Measuring the Correctness of Double-Keying: Error Classification and Quality Control in a Large Corpus of TEI-Annotated Historical Text

Philology in the Digital Age

2011 Annual Conference and Members' Meeting of the TEI Consortium in Würzburg, Germany

Susanne Haaf, Frank Wiegand, Alexander Geyken

Deutsches Textarchiv, Berlin-Brandenburgische Akademie der Wissenschaften
<http://www.deutschestextarchiv.de/>, dta@bbaw.de

(*Paper to be published in the Journal of the Text Encoding Initiative, Issue 4, 2013*)

Abstract

Among mass digitization methods, double-keying is considered to be the one with the lowest error rate. This method requires two independent transcriptions of a text by two different operators. It is particularly well suited to historical texts, which often exhibit deficiencies like poor master copies or other difficulties such as spelling variation or complex text structures.

Providers of data entry services using the double-keying method generally advertise very high accuracy rates (around 99.95% to 99.98%). These advertised percentages are generally estimated on the basis of small samples, and little if anything is said about either the actual amount of text or the text genres which have been proofread, about error types, proofreaders, etc. In order to obtain significant data on this problem it is necessary to analyze a large amount of text representing a balanced sample of different text types, to distinguish the structural XML/TEI level from the typographical level, and to differentiate between various types of errors which may originate from different sources and may not be equally severe.

This paper presents an extensive and complex approach to the analysis and correction of double-keying errors which has been applied by the DFG-funded project "Deutsches Textarchiv" (German Text Archive, hereafter DTA) in order to evaluate and preferably to increase the transcription and annotation accuracy of double-keyed DTA texts. Statistical analyses of the results gained from proofreading a large quantity of text are presented, which verify the common accuracy rates for the double-keying method.

Introduction

Among mass digitization methods, double-keying is considered to be the one with the lowest error rate. The double-keying method requires two independent transcriptions of a text by two different operators. The two resulting versions are compared in order to detect transcription errors. Since two human operators are unlikely to make the same mistakes, the double-keying method yields very high accuracy rates. It is particularly well suited to historical texts, which often exhibit deficiencies like poor master copies or other difficulties such as spelling variation or complex text structures. Therefore, for the digitization of large amounts of historical text, it is common to employ the method of double-keying rather than applying (semi-)automatic methods (see DFG 2009, 10–11). In addition to data entry, the double-keying process may also include an enrichment of the text with structural annotations. In such cases, the comparison of the double-keyed output texts may also reveal ambiguities in the underlying tag set.

Providers of data entry services using the double-keying method generally advertise accuracy rates around 99.95% to 99.98%. The "DFG Practical Guidelines on Digitization" (DFG 2009, 11) even foresee accuracy rates of 99.997%, resulting in "virtually error-free texts." While we do not dispute that the accuracy rate of double-keying texts is very high, the advertised percentages are generally estimated on the basis of small samples, and little if anything is said about either the actual amount of text or the text genres which have been proofread, about error types, proofreaders, or other factors that might affect accuracy.[1] Studies on accuracy rates and quality control have been undertaken for Optical Character Recognition systems (see, for instance, Furrer et al. 2011; Holley 2009a and 2009b; Tanner et al. 2009), whereas—to our knowledge—the double-keying accuracy has not yet been subjected to further research. But in order to develop quality control methods for the digitization of large amounts of (historical) text via double-keying, knowledge about typical error categories and their corresponding error rates leading to precise evaluations of text transcription and annotation approaches is crucial.

In order to obtain significant data on this problem it is necessary to analyze a large amount of text representing a balanced sample of different text types, to distinguish the structural XML/TEI level from the typographical level, and to differentiate between various types of errors which may originate from different sources and may not be equally severe.

This paper presents an extensive and complex approach to the analysis and correction of double-keying errors which has been applied by the DFG-funded project "Deutsches Textarchiv" (German Text Archive, hereafter DTA).[2] Our aim was to evaluate and preferably to increase the transcription and annotation accuracy of double-keyed DTA texts. Statistical analyses of the results gained from proofreading a large quantity of text are presented, which verify the common accuracy rates for the double-keying method.

Large-Scale Double-Keying for the DTA

Since 2007, the DTA has been compiling a steadily growing, balanced corpus of German historical texts of different genres ranging from the late 18th to the end of the 19th century.[3] About 260,000 full-text digitized pages corresponding to more than 400 million characters have already been digitized by various methods, among which transcription by non-native speakers via double-keying is the most important one (175,151 pages and about 298 million characters).

As part of the double-keying process, structural annotations were added to the transcribed texts by the typists using a simplified pseudo-XML tag set. These XML tags were then converted (semi-)automatically into the DTA "base format", a subset of the TEI P5 annotation standard. The DTA "base format" consists of about 80 elements within `<text>` and a stable set of attribute-value pairs. It has been applied consistently over the entire DTA corpus in order to preserve intertextual coherence on a structural level.[4] Furthermore, use of the DTA "base format" may serve to raise the degree of interoperability among TEI-annotated historical texts.[5]

[1] For example, the accuracy level of the digitized dictionary of J. H. Campe has been ascertained to achieve up to 99.996%; cf. <http://www.textgrid.de/fileadmin/berichte-1/report-4-1.pdf>, p. 16.

[2] Deutsches Textarchiv (DTA): <http://www.deutschestextarchiv.de/>.

[3] For further information about the DTA project, see Geyken et al. (2011).

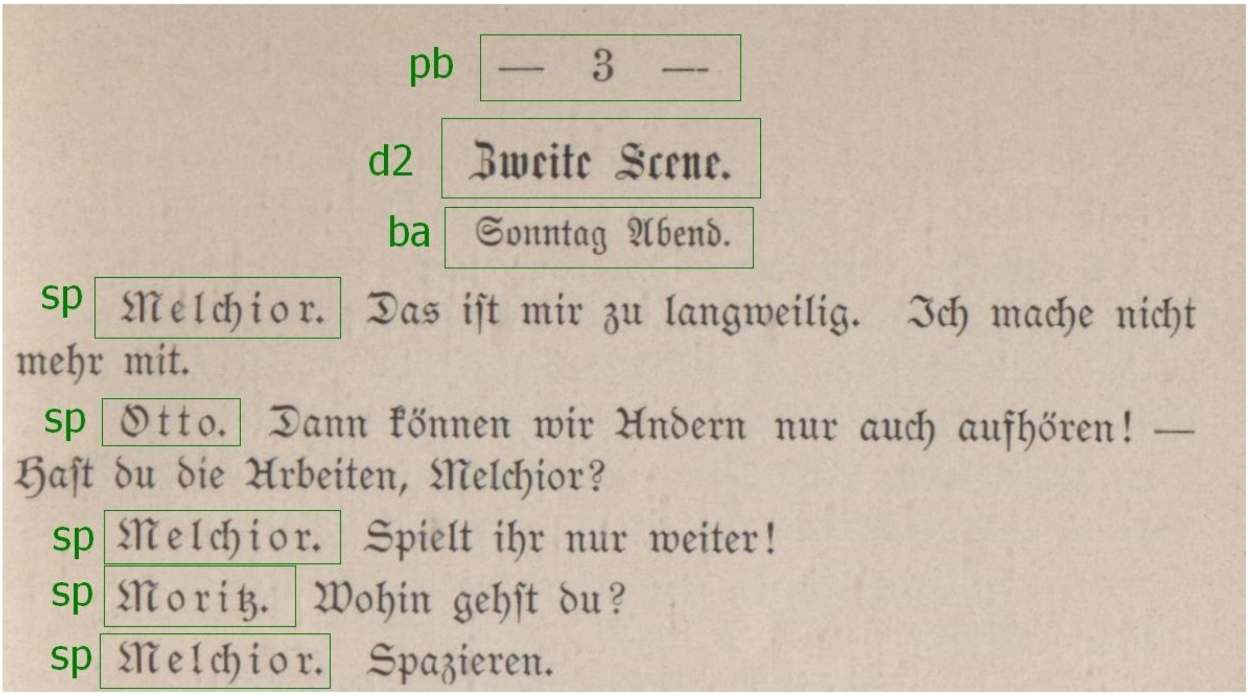
[4] DTA "base format": <http://www.deutschestextarchiv.de/doku/basisformat>.

[5] The necessity of using a basic TEI subset for text structuring to ensure the interoperability of TEI texts has been pointed out by Unsworth (2011), among others.

Methods of Quality Control at the DTA

In order to ensure both transcription and annotation accuracy, various quality control methods are applied before and after text recognition.[6]

Generally, data typists are provided with guidelines specifying the basic encoding format as well as transcription rules (hereafter referred to collectively as DTA guidelines).[7] In addition, the DTA provides special information about each book to the typists to support the transcription and annotation task, e.g. information about special cases and exceptions of the source text as well as structuring examples. The latter are added as labels to the images provided of each book (see figure 1), using the DTA ZOT ("zoning tool"), which was developed specifically for this task.[8]



— 3 —

Zweite Scene.

Sonntag Abend.

Melchior. Das iſt mir zu langweilig. Ich mache nicht mehr mit.

Otto. Dann können wir Andern nur auch aufhören! — Haſt du die Arbeiten, Melchior?

Melchior. Spielt ihr nur weiter!

Moritz. Wohin gehſt du?

Melchior. Spazieren.

pb
d2
ba
sp
sp
sp
sp
sp

Figure 1. Structuring examples for a DTA image (Frank Wedekind, *Frühlings Erwachen* [Zürich: Groß, 1891], 3)

Even though these precautions improve the textual and structural adequacy of text transcriptions considerably, errors cannot be avoided completely in advance. Not only are transcription problems still likely to occur, but tagging errors may occur as well because of underspecifications in the tag set. Automatic methods may help to systematically identify typical transcription and printing errors,[9] but proofreading remains indispensable for detecting individual exception cases. Therefore, the DTA employs word-by-word proofreading by native speakers, especially philologists, of historical texts with high degrees of both structural complexity and spelling variation.

[6] For further information on the quality assurance methods applied in the DTA project, see Geyken et al. (2012).

[7] DTA guidelines: <http://www.deutschestextarchiv.de/doku/richtlinien>.

[8] DTA ZOT: <http://www.deutschestextarchiv.de/doku/software>.

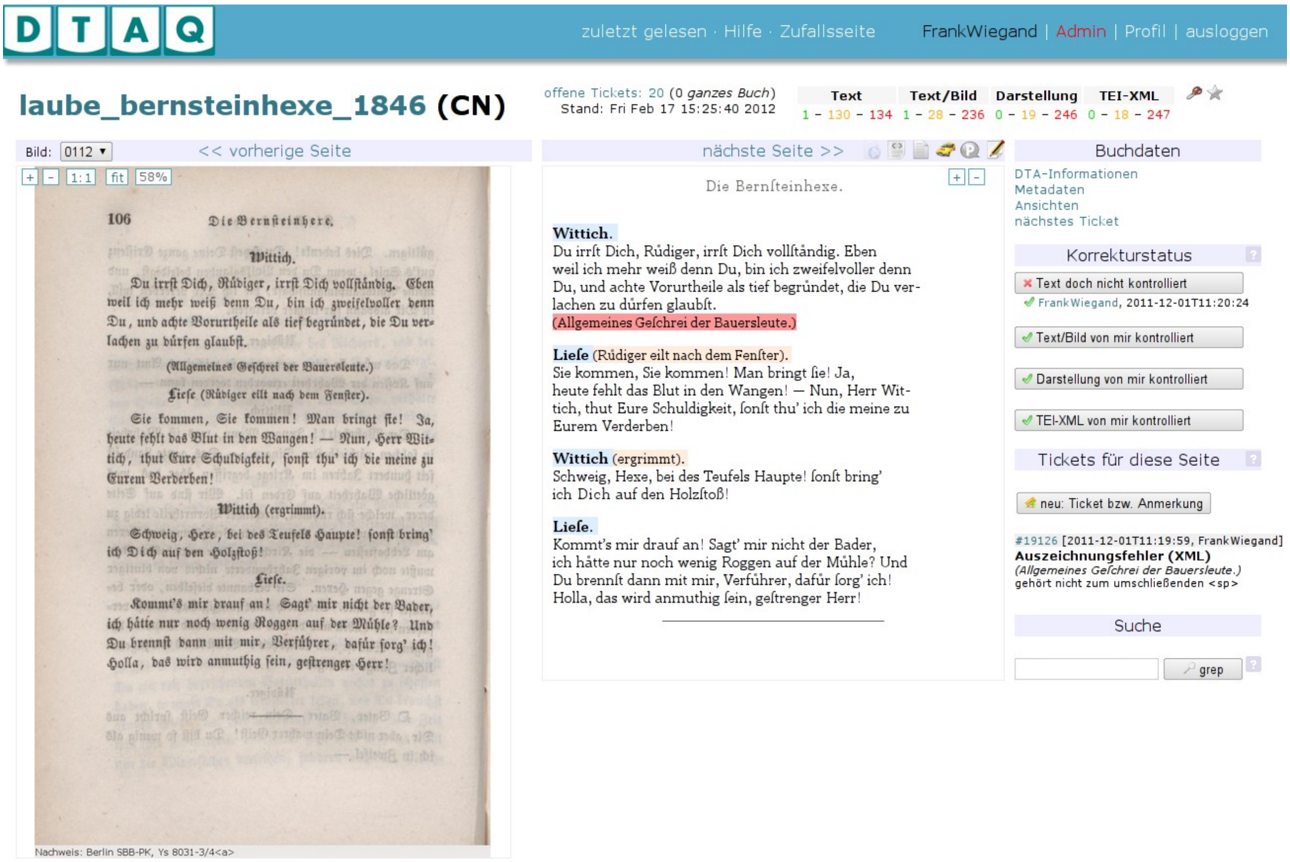
[9] For example, for orthographic standardization purposes some DTA texts are compared based on string matching with normalized transcriptions of the particular works. As a side effect of this procedure it is possible to extract non-matching strings that contain transcription or printing errors. See Jurish et al. (forthcoming).

DTAQ

To support the proofreading task, the quality assurance platform DTAQ[10] has been developed, a web-based collaboration tool which allows distributed proofreaders to review DTA texts page by page with reference to the source images. After data entry and conversion of the text into XML/TEI, each book is integrated in DTAQ.

Along with their metadata, books are presented in various ways via a customizable web front end. To provide different views of the source images together with their corresponding TEI transcriptions, each book is split up into single page files, using the `<pb>` element as a separator (the separation process is reversible, so that these single files can be merged together to get the original TEI document back).

The DTAQ GUI is highly customizable. Users can set up the workspace according to their screen resolution, and there are various ways to present special Unicode characters, even if there is no font available on the client side to display them in a satisfactory way.[11] Furthermore, there are options to adjust the presentation of facsimiles according to the user's particular preferences when proofreading: facsimiles can be zoomed into, and they may be moved individually within their frame to get a better look at a particular text passage. The transcription itself is offered either as raw XML/TEI, as rendered HTML using customized XSLT stylesheets (see figure 2), or as plain text which is searchable via a GUI wrapper using the `egrep(1)` command.


DTAQ
zuletzt gelesen · Hilfe · Zufallsseite
FrankWiegand | Admin | Profil | ausloggen
laube_bernsteinhexe_1846 (CN)
offene Tickets: 20 (0 ganzes Buch)
Stand: Fri Feb 17 15:25:40 2012
Text 1 - 130 - 134
Text/Bild 1 - 28 - 236
Darstellung 0 - 19 - 246
TEI-XML 0 - 18 - 247
Bild: 0112
<< vorherige Seite
nächste Seite >>
+ - 1:1 fit 58%
106 Die Bernſteinhexe.
Nachweis: Berlin SBB-PK, Ys 8031-3/4<a>
Buchdaten
DTA-Informationen
Metadaten
Ansichten
nächstes Ticket
Korrekturstatus
Text doch nicht kontrolliert
Text/Bild von mir kontrolliert
Darstellung von mir kontrolliert
TEI-XML von mir kontrolliert
Tickets für diese Seite
neu: Ticket bzw. Anmerkung
Suche
grep

Die Bernſteinhexe.

Wittich.
Du irrſt Dich, Rüdiger, irrſt Dich vollſtändig. Eben
weil ich mehr weiß denn Du, bin ich zweifelvoller denn
Du, und achte Vorurtheile als tief begründet, die Du ver-
lachen zu dürfen glaubſt.
(Allgemeines Geſchrei der Bauersleute.)

Lieſe (Rüdiger eilt nach dem Fenſter).
Sie kommen, Sie kommen! Man bringt ſie! Ja,
heute fehlt das Blut in den Wangen! — Nun, Herr Wit-
tich, thut Eure Schuldigkeit, ſonſt thu' ich die meine zu
Eurem Verderben!

Wittich (ergrimmt).
Schweig, Hexe, bei des Teufels Haupte! ſonſt bring'
ich Dich auf den Holzſtoß!

Lieſe.
Kommt's mir drauf an! Sagt' mir nicht der Bader,
ich hätte nur noch wenig Roggen auf der Mühle? Und
Du brennſt dann mit mir, Verführer, dafür ſorg' ich!
Holla, das wird anmuthig ſein, geſtrenger Herr!

FrankWiegand, 2011-12-01T11:20:24

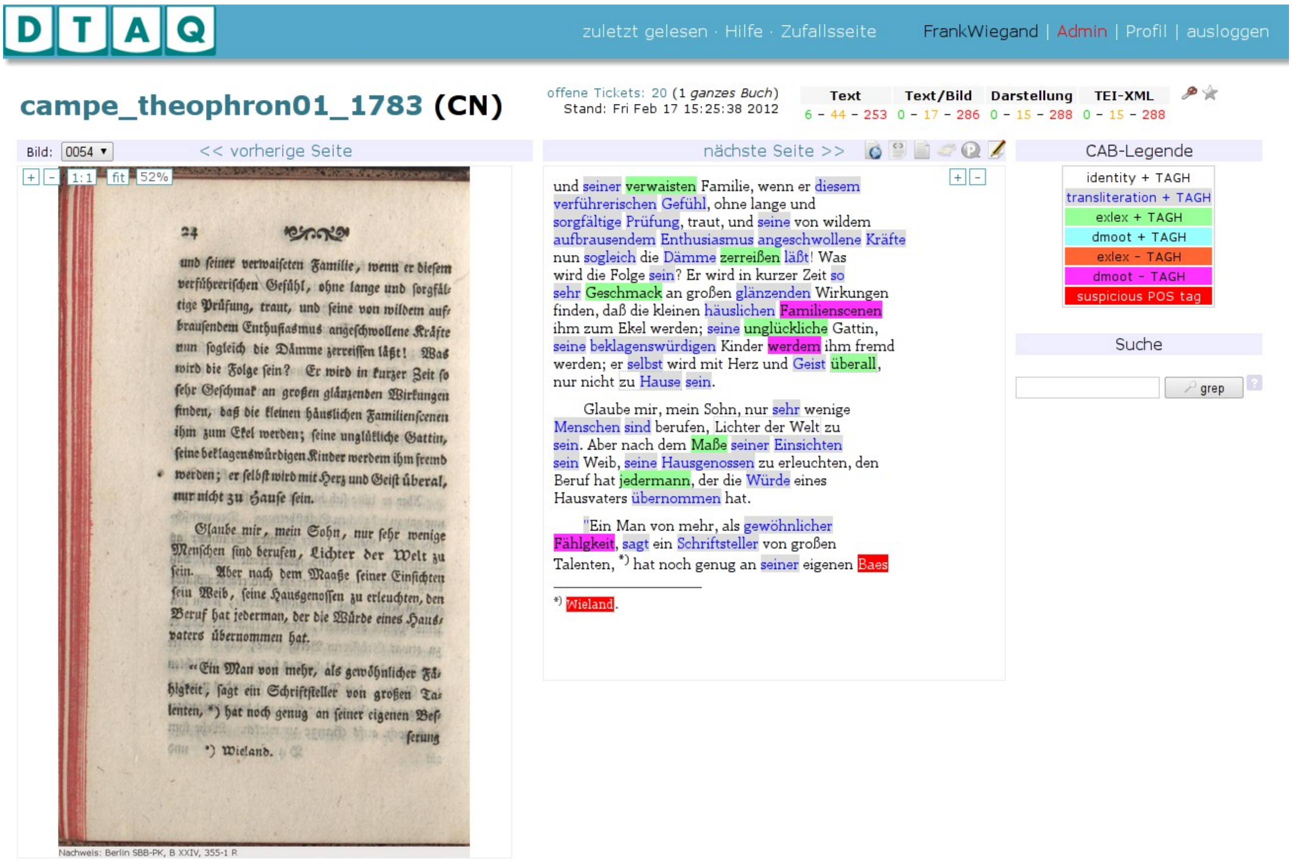
#19126 [2011-12-01T11:19:59, FrankWiegand]
Auszeichnungsfehler (XML)
(Allgemeines Geſchrei der Bauersleute.)
gehört nicht zum umschließenden <sp>

Figure 2. DTAQ, parallel view (image and rendered HTML)

[10] DTAQ: <http://www.deutschestextarchiv.de/dtaq>.

[11] For example, diacritical characters like the "combining Latin small letter e" (U+0364), or the "Latin small letter r rotunda" (U+A75B) are very frequent in the DTA corpus, but because of the lack of proper Unicode support even in modern browsers they are sometimes rendered into squares or result in scrambled lines, so there is a need to circumvent these inconveniences.

A fourth mode of text presentation is provided in DTAQ, showing the results of the linguistic analyses associating each token with a normalized modern orthographic word form using CAB[12] (see figure 3).



24

und ſeiner verwaiſeten Familie, wenn er dieſem verführeriſchen Gefühl, ohne lange und ſorgfältige Prüfung, traut, und ſeine von wildem aufbrauſendem Enthuſiasmus angeſchwollene Kräfte nun ſogleich die Dämme zerreiſſen läßt! Was wird die Folge ſein? Er wird in kurzer Zeit ſo ſehr Geſchmak an großen glänzenden Wirkungen finden, daß die kleinen häuslichen Familienſcenen ihm zum Ekel werden; ſeine unglükliche Gattin, ſeine beklagenswürdigen Kinder werdem ihm fremd werden; er ſelbſt wird mit Herz und Geiſt überal, nur nicht zu Hauſe ſein.

Glaube mir, mein Sohn, nur ſehr wenige Menſchen ſind berufen, Lichter der Welt zu ſein. Aber nach dem Maaße ſeiner Einſichten ſein Weib, ſeine Hausgenoſſen zu erleuchten, den Beruf hat jederman, der die Würde eines Hausvaters übernommen hat.

„Ein Man von mehr, als gewöhnlicher Fähigkeit, ſagt ein Schriftſteller von großen Talenten, *) hat noch genug an ſeiner eigenen Beſ-
ſerung

*) Wieland.

und seiner verwaisten Familie, wenn er diesem verführerischen Gefühl, ohne lange und sorgfältige Prüfung, traut, und seine von wildem aufbrausendem Enthusiasmus angeschwollene Kräfte nun sogleich die Dämme zerreißen läßt! Was wird die Folge sein? Er wird in kurzer Zeit so sehr Geschmack an großen glänzenden Wirkungen finden, daß die kleinen häuslichen Familienscenen ihm zum Ekel werden; seine unglückliche Gattin, seine beklagenswürdigen Kinder werdem ihm fremd werden; er selbst wird mit Herz und Geist überall, nur nicht zu Hause sein.

Glaube mir, mein Sohn, nur sehr wenige Menschen sind berufen, Lichter der Welt zu sein. Aber nach dem Maße seiner Einsichten sein Weib, seine Hausgenossen zu erleuchten, den Beruf hat jedermann, der die Würde eines Hausvaters übernommen hat.

"Ein Man von mehr, als gewöhnlicher Fählgkeit, sagt ein Schriftsteller von großen Talenten, *) hat noch genug an seiner eigenen Baes

*) Wieland.

Figure 3. DTAQ, CAB view (modernized spelling)

Finally, there is the option to obtain the part-of-speech analysis corresponding to each single token of a page.

This way, users of DTAQ can analyze the digital image together with a chosen representation of the corresponding transcription page by page, in order to compare them and flag erroneous pieces. If errors are found, proofreaders can report them as “tickets” (as in a software bug-tracking system), providing information about the type and location of the error, as well as the correct form. The information comprised by each ticket may be modified or supplemented by any user. Tickets—as well as information about all changes made to them—are stored in the database back end; various RSS feeds as well as severity and priority levels are provided, too.

The back end of DTAQ is built upon many open source packages. Using Perl[13] as a glue language, the system runs on Catalyst[14] and connects to a PostgreSQL[15] database via the DBIx::Class[16] ORM, and the web pages are built with Template Toolkit.[17] The front end makes heavy use of jQuery[18] and Highcharts JS[19] to create a very interactive and responsive user interface.

[12] Cascaded Analysis Broker; see Jurish (2012).
[13] The Perl Programming Language: <http://www.perl.org/>.
[14] Catalyst—The Elegant MVC Web Application Framework: <http://www.catalystframework.org/>.
[15] PostgreSQL: <http://www.postgresql.org/>.
[16] DBIx::Class: <https://metacpan.org/module/DBIx::Class>.
[17] Template Toolkit: <http://template-toolkit.org/>.
[18] jQuery: <http://jquery.com/>.
[19] Highcharts JS: <http://www.highcharts.com/>.

Using only Open Source technologies and a robust web application framework along with modern Javascript libraries, no complicated setup is required, so that collaborators have easy access to DTAQ and benefit from the various possibilities that modern web sites can provide.

Quality assurance via proofreading using DTAQ began in April 2011. During 10 months of work with DTAQ (through January 2012), more than 15,000 different pages were proofread.

Error Types

During the first examinations of our texts with regard to the overall text quality, several typical error sources resulting in characteristic error categories were identified. These categories include violations of the DTA guidelines, printing errors (i.e. textual anomalies of the source text), and errors arising during the transformation of our simplified pseudo-XML format to TEI P5. The first error category can be further subdivided into transcription errors, annotation (structuring) errors with regard to the DTA "base format," and HTML rendering problems (which may in turn necessitate changes to the tag set). This analysis therefore results in five different error types:

1. transcription error
2. annotation error (XML)
3. representation error (HTML)
4. workflow error
5. errata (in corrigenda, certain, uncertain, semantic)

These error types were used to categorize errors during the text proofreading process of our case study, which is presented in the next section.

Measuring the Correctness of Double-Keying: A Case Study

Between August 2011 and January 2012, we conducted a case study at the DTA, during which a considerable amount of text drawn from the DTA corpus was proofread. The goal of this proofreading case study was twofold. First, we intended to measure the error rates of double-keying on the basis of a larger text sample by means of a careful proofreading process using the DTAQ platform. Second, we tried to take advantage of the proofreading results to improve our quality assurance methods in the DTA workflow and thus our ability not only to correct, but also to prevent, errors.

The study was carried out in two phases lasting three months each. For each phase, a sample was extracted from the DTA corpus and provided to the proofreaders. Twenty-two persons took part in the proofreading process, which consisted of checking the transcribed texts for errors by comparing them to the corresponding images using DTAQ. Errors were reported as tickets and classified according to the error types listed above.

Texts were chosen by genre and typeface on the one hand (science/Fraktur, science/Antiqua, fiction/Fraktur, functional literature/Fraktur) and by vicennium (twenty-year period) on the other hand (1780–1799, 1800–1819, 1820–1839, 1840–1859, 1860–1879, 1880–1899), which led to 24 different categories. Thus, we aimed at getting a wide cross-section of our corpus based on the assumption that varying degrees of difficulty are encountered in text transcriptions and annotations depending on the above-mentioned properties time of publication, genre, and typeface (see figure 4).

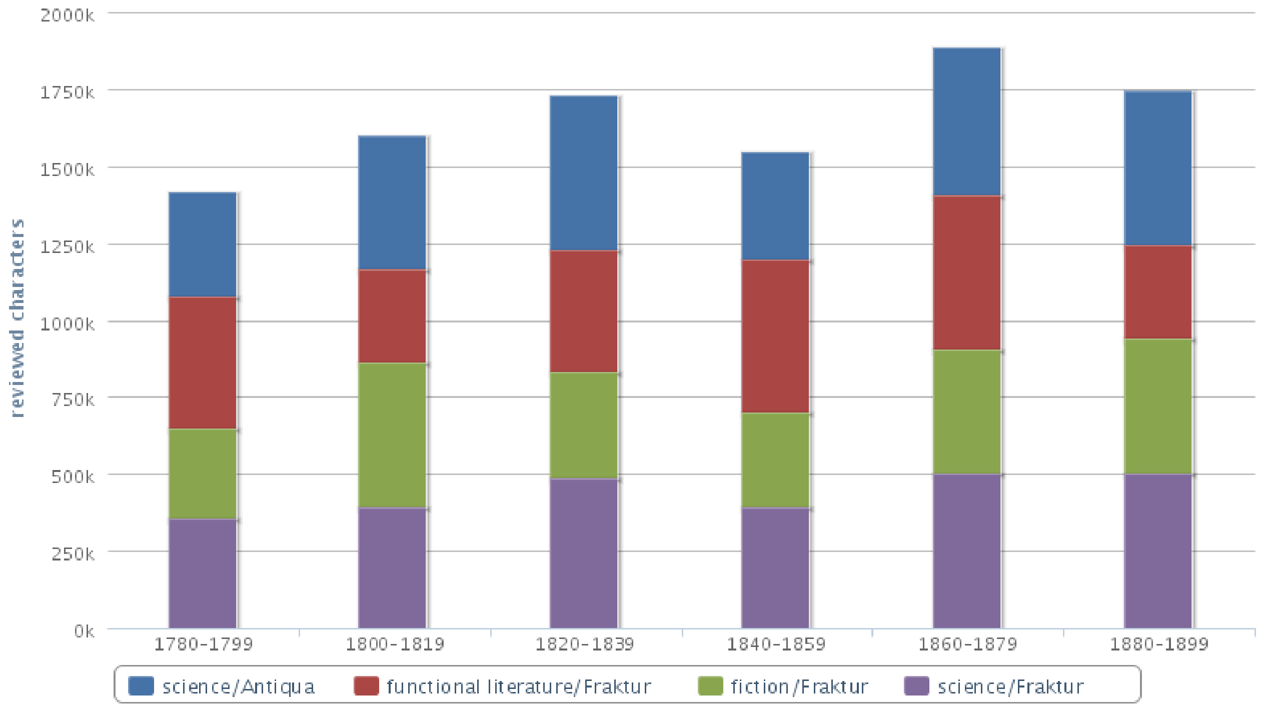


Figure 4. Reviewed characters by category

In addition, the selection of text samples was based on as many works as possible, since transcription and annotation quality may depend on factors such as the condition of the master copy, the structural complexity of the source text, and the presence of foreign language material.[20] More precisely, the following constraints were imposed on each category: A minimum length of 27,000 characters and a maximum length of 100,000 characters[21] for each source text was required. Within one source text, only consecutive pages were considered, generally taken from within the `<body>` element of the respective TEI document to avoid an abundance of title pages, tables of contents, indices, and the like. Nevertheless, a limited number of examples for the aforementioned structures was taken into account as well. Furthermore, in order to minimize category bias due to sample selection, each category had to be represented by at least 300,000 but not more than 500,000 characters, drawn from at least three different sample texts.

The resulting text sample consisted of more than 9.9 million characters on 7,208 pages taken from 170 books. Table 1 shows the distribution of books and characters over categories.

[20] See the Apex Covantage Price Matrix: <http://accesstei.apexcovantage.com/Home/PriceMatrix>. Prices for data entry are set according to the presence of frequent error sources, e.g. difficult typefaces, broken characters, or physical damage to the source document.

[21] Samples consisted of entire pages; therefore slight variation around the 100,000-character limit is possible.

Table 1. Characters (books) per category

	science/ Fraktur	science/ Antiqua	fiction/ Fraktur	funct. literature/ Fraktur	total
1780–1799	357,609 (9)	341,464 (5)	293,757 (7)	426,130 (9)	1,418,960 (30)
1800–1819	396,308 (8)	434,112 (8)	465,459 (6)	305,677 (4)	1,601,556 (26)
1820–1839	485,709 (9)	501,483 (10)	344,919 (9)	399,993 (6)	1,732,104 (34)
1840–1859	393,898 (7)	350,118 (4)	310,060 (5)	494,828 (6)	1,548,904 (22)
1860–1879	501,619 (9)	481,593 (7)	405,302 (5)	500,461 (6)	1,888,975 (27)
1880–1899	502,353 (10)	501,856 (12)	438,729 (6)	302,283 (3)	1,745,221 (31)
total	2,637,496 (52)	2,610,626 (46)	2,258,226 (38)	2,429,372 (34)	9,935,720 (170)

Results

The first major result of our case study was that the proofreading conducted on the sample corpus (7,208 pages) yielded only 2,758 tickets. In other words, only one error was reported for every three full pages proofread. This outcome already supports claims of high accuracy rates for the double-keying method. Further analysis of the results of our study—namely the proportions of annotation and transcription errors, as well as of errors which go beyond the level of double-keying errors in a strict sense—is presented in the remainder of this section.

Annotation Errors

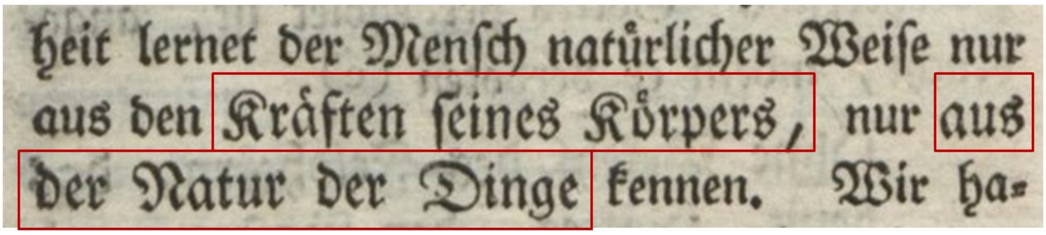
Annotation errors are not only a problem for the semantic structuring of texts; they almost always result in presentation errors as well, so they may have a considerable impact on the outcome of the text digitization process at different levels. As mentioned above, the annotation of the transcribed text is part of the double-keying process, so that annotation errors may be taken into account for the calculation of double-keying accuracy. However, certain parts of the text annotation are carried out during other stages of the digitization process as well.

Our study yielded several typical sources for annotation errors, such as uncertainties regarding the correct application of the DTA guidelines during image preparation in advance as well as during text transcription, or even during proofreading. Furthermore, some annotation problems were the result of automatic processes which had been applied for the conversion of the pseudo-XML tagged texts to TEI P5.

Nevertheless, based on the data from our proofreading task, we were able to identify annotation error types which usually emerge during the text recognition process.

1. Concerning typographical particularities, the recognition of a change between different Fraktur typefaces seems to cause most difficulties.

Example 1. Misinterpretation of typographical conditions (1)



heit lernet der Menſch natürlicher Weiſe nur
aus den Kräften ſeines Körpers, nur aus
der Natur der Dinge kennen. Wir ha-

Wrong:

```
<p>[...]heit lernet der Menſch natürlicher Weiſe nur <lb/> aus den
Kra&#x0364;ften ſeines Ko&#x0364;rpers, nur <hi
rendition="#fr">aus<lb/>der Natur der Dinge</hi> kennen. Wir ha-[...]</p>
```

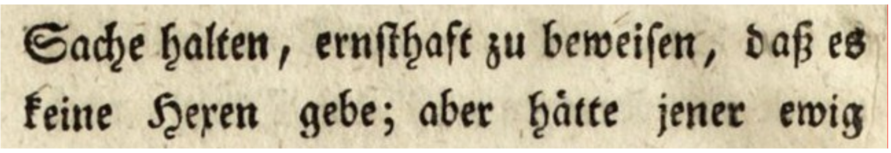
Right:

```
<p>[...]heit lernet der Menſch natürlicher Weiſe nur <lb/> aus den <hi
rendition="#fr">Kra&#x0364;ften ſeines Ko&#x0364;rpers,</hi> nur <hi
rendition="#fr">aus<lb/>der Natur der Dinge</hi> kennen. Wir ha-[...]</p>
```

[excerpt from http://www.deutschestextarchiv.de/sailer_selbstmord_1785/36]

Nevertheless, misinterpretations of the typographical conditions of the text source did occur on other levels as well. For example, in some cases increased letter-spacing was erroneously applied to text parts, where no change of letter-spacing had been intended by the printer.

Example 2. Misinterpretation of typographical conditions (2)



Sache halten, ernſthaft zu beweiſen, daß es
keine Hexen gebe; aber hätte jener ewig

Wrong:

```
Sache halten, ernſthaft zu beweiſen, <hi rendition="#g">daß</hi> es<lb/>
keine Hexen gebe; aber ha&#x0364;tte jener ewig
```

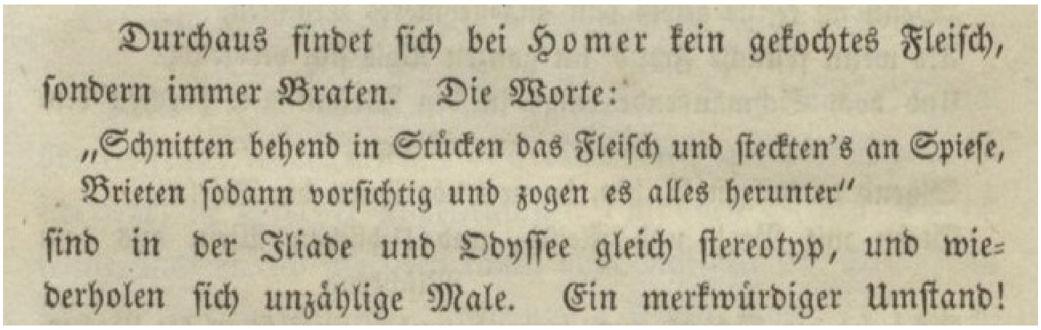
Right:

```
Sache halten, ernſthaft zu beweiſen, daß es<lb/>
keine Hexen gebe; aber ha&#x0364;tte jener ewig
```

[excerpt from http://www.deutschestextarchiv.de/dohm_juden02_1783/16]

2. Because of the image-oriented (and therefore page-oriented) preparation of the source texts, difficulties are encountered regarding structuring across page breaks. Typists sometimes decided wrongly whether a certain structure (e.g. a paragraph) did or did not finish at the end of a page.
3. Sometimes, a single paragraph containing a block insertion was erroneously transcribed as several separate paragraphs.

Example 3. Paragraph containing a block insertion



Durchaus findet ſich bei Homer kein gekochtes Fleiſch, ſondern immer Braten. Die Worte:

„Schnitten behend in Stücken das Fleiſch und ſteckten's an Spieſe,
Brieten ſodann vorſichtig und zogen es alles herunter“

ſind in der Iliade und Odyſſee gleich ſtereotyp, und wiederholen ſich unzählige Male. Ein merkwürdiger Umſtand!

Wrong:

```
<p>Durchaus findet ſich bei <hi rendition="#g">Homer</hi> kein gekochtes
Fleiſch,<lb/>ſondern immer Braten. Die Worte:<lb/></p>
<p>„Schnitten behend in Stu&#x0364;cken das Fleiſch und ſteckten’s an
Spieſe,<lb/>Brieten ſodann vorſichtig und zogen es alles herunter“<lb/>
ſind in der Iliade und Odyſſee gleich ſtereotyp, und wie-<lb/>derholen
ſich unza&#x0364;hlige Male. Ein merkwu&#x0364;rdiger Umſtand!<lb/></p>
```

Right:

```
<p>Durchaus findet ſich bei <hi rendition="#g">Homer</hi> kein gekochtes
Fleiſch,<lb/>ſondern immer Braten. Die Worte:<lb/>
<cit><quote>„Schnitten behend in Stu&#x0364;cken das Fleiſch und
ſteckten’s an Spieſe,<lb/>Brieten ſodann vorſichtig und zogen es alles
herunter“<lb/></quote></cit>
ſind in der Iliade und Odyſſee gleich ſtereotyp, und wie-<lb/>derholen
ſich unza&#x0364;hlige Male. Ein merkwu&#x0364;rdiger Umſtand!<lb/></p>
```

[excerpt from http://www.deutschestextarchiv.de/anthus_esskunst_1838/39]

Transcription Errors

Regarding the double-keying accuracy in the DTA corpus, transcription errors are of special interest, since they are really only a matter of correct text recognition by the typists. In addition to pure characters, we considered recognition errors involving line breaks as transcription errors as well, because they have a direct effect on the number of characters in the raw text transcription. In contrast, for example, missing ornamental elements were regarded as a matter of annotation, even though they entail a loss of source-text material, since they are indicated only through XML elements and therefore don’t have any influence on the number of characters in the raw text.

Following these rules, the 7,208 pages of our text sample contained 830 transcription errors.

Characteristics of Transcription Errors

The analysis of the transcription errors found in our proofreading sample brought to light some interesting facts. Obviously, some letters caused more recognition difficulties than others, leading to pairs of letters typically and frequently causing transcription errors (see table 2).

Table 2. Pairs of letters frequently causing transcription errors

source text	**transcription**	**example source**	**example transcription**
and vice versa			
f	ſ (long s)	tiefſter	tieffter
occurs in Fraktur and Antiqua		Schriften	Schriſten
u	n	Abſcheu	Abſchen
occurs mostly in Fraktur		Schuhſchnallen	Schuhſchuallen
b	h	erhellt	erbellt
occurs in Fraktur and oblique Antiqua		geben	gehen
r	t	neigten	neigren
occurs in Fraktur		innerhalb	innethalb
V	B	Vetter	Better
occurs in Fraktur		Brief	Vrief
l	i	eilig	ellig
occurs in Fraktur and Antiqua		taumelt	taumeit

Other, less frequent transcription problems involved the letters ſ/s, c/e, r/x, t/i, o/e, C/T, J/I (capital i), N/R, e/c, m/n, a/u, and k/t.

Often (in 124 cases), punctuation marks (such as . , : ! ? / ( ) [ ] { } ‚‘ „“) were affected, or spaces and line breaks were wrongly omitted or inserted. Such errors would not affect corpus searches for word forms, but may influence the results of the tokenization and consequently of many linguistic analyses based on it.

Occasionally, transcription errors resulted in “false friends,” i.e. valid word-forms which were however lexically distinct from the correct transcription. In such cases the incorrect words would be found via corpus search, whereas the correct forms would not. Examples for such false friends were:

- wrong: Annaten (annates)—right: Annalen (annals)
  [http://www.deutschestextarchiv.de/klueber_voelkerrecht02_1821/36]
- Auſſatz (leprosy)—Aufſatz (essay)
  [http://www.deutschestextarchiv.de/wedekind_erwachen_1891/22]
- Laute (lute)—Laufe (course)
  [http://www.deutschestextarchiv.de/beck_eisen03_1897/20]
- Halde (dump)—Haide (heath)
  [http://www.deutschestextarchiv.de/platen_oedipus_1829/10]
- Chat (chat)—That (act)
  [http://www.deutschestextarchiv.de/sanders_woerterbuchschreiber_1889/37]

Questions about false friends arising from transcription errors, the parts of speech affected, and word accuracy in general are of interest in other contexts (see Nartker et al. 2003; Tanner 2009). For us, however, character accuracy remains the foremost concern, since many DTA-corpus usage scenarios require correct transcriptions of the source text on the character level.

Transcription Accuracy

After evaluating all transcription error tickets, each error was assigned the appropriate Levenshtein distance *d* (where insertion, deletion, and substitution are each assigned a cost of 1; see Levenshtein 1965). In most cases, one error per ticket was found. Nevertheless, there were cases in which *d* turned out to be greater than 1 (see table 3).

To calculate the length of the transcribed text in total, each Unicode character that occurred (excluding combining diacritical marks, but including line breaks) was counted as 1 character. After fixing each error, the accuracy (correctness) *c* of the transcribed text *T* with respect to the original source *S* was calculated as:

$$c=\frac{max(|S|,|T|)-d}{max(|S|,|T|)}$$

Table 3. Examples for the calculation of transcription correctness

source *S*	transcription *T*	|*S*|	|*T*|	*d*	*c*
als	als	3	3	0	1
Luft	Luſt	4	4	1	0.75
neceſſariis	neceſſarns	11	10	2	0.82
ſchwangern	ſchroangern	10	11	2	0.82
nun	mm	3	2	3	0

We calculated transcription correctness for each category as well as in total. Not surprisingly, transcription accuracy tended to decrease with increasing age of the texts (see table 4).

Table 4. Transcription accuracy by vicennium

vicennium	transcription accuracy
1780–1799	99.9719%
1800–1819	99.9858%
1820–1839	99.9943%
1840–1859	99.9976%
1860–1879	99.9975%
1880–1899	99.9966%

Reasons for the increasingly poor transcription accuracy of the older texts may include the comparatively poorer condition of older prints, as well as the fact that older Fraktur typefaces usually are more ornate than newer ones.

Concerning text genre, it turned out that contrary to our hypothesis, fictional texts were not easier to transcribe than non-fictional scientific texts (see table 5).

Table 5. Transcription accuracy by genre

genre	transcription accuracy
fiction	99.9886%
functional literature	99.9916%
science	99.9916%

Interestingly, in contrast to OCR output (see Furrer et al. 2011, [1]) the typeface did not have any notable effect on the accuracy rate (see table 6).

Table 6. Transcription accuracy by typeface

typeface	transcription accuracy
Antiqua	99.9918%
Fraktur	99.9906%

Table 7 shows the accuracy rates for each of our defined categories.

Table 7. Transcription accuracy by category

	science/ Fraktur	science/ Antiqua	fiction/ Fraktur	funct. literature/ Fraktur	total
1780–1799	99.9771%	99.9751%	99.9493%	99.9805%	99.9719%
1800–1819	99.9816%	99.9841%	99.9951%	99.9794%	99.9858%
1820–1839	99.9961%	99.9948%	99.9904%	99.9952%	99.9943%
1840–1859	99.9987%	99.9986%	99.9955%	99.9974%	99.9976%
1860–1879	99.9964%	99.9985%	99.9990%	99.9962%	99.9975%
1880–1899	99.9956%	99.9988%	99.9934%	99.9993%	99.9966%
total	99.9915%	99.9918%	99.9886%	99.9916%	99.9909%

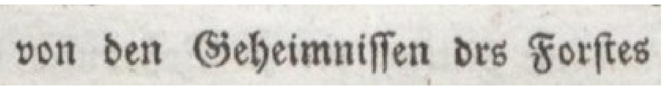
The overall accuracy rate for our text sample was 99.9909%, that is, 11 transcription errors in every 100,000 characters. Hence, it exceeded the accuracy rates of 99.95%–99.98% advertised by double-keying companies (see the introduction above). In addition, the guaranteed accuracy rates were obtained for 23 of 24 categories.

Errors Beyond Double-Keying Correctness

Knowledge about **workflow errors** and **presentation errors** is important in order to improve our text quality and quality assurance methods, as well as the presentation of our texts. Nevertheless, these kinds of errors generally do not arise from double-keying. Therefore, they were reported during the proofreading process but were not taken into account for measuring the double-keying correctness of our texts.

Printing errors (errata) are properly a matter of text revision after data entry, since non-native speakers would not be able to annotate printing errors and offer the correct form. However, they sometimes are accompanied by transcription errors, for example in cases where they were inadvertently corrected by the typists.

Example 4. Transcription problems with printing errors (1)



von den Geheimniſſen drs Forſtes

Wrong transcription:

```
von den Geheimniffen des Forftes
```

Right transcription:

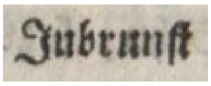
```
von den Geheimniffen drs Forftes
```

[excerpt from http://www.deutschestextarchiv.de/fouque_undine_1811/19]

Such corrections of printing errors or illegible letters may also lead to misinterpretations, thus introducing new errors.

Finally, printing errors may result in mistranscriptions of surrounding characters because of instructions in the DTA guidelines.

Example 5. Transcription problems with printing errors (2)



Jubrunſt

[excerpt from http://www.deutschestextarchiv.de/campe_theophron01_1783/36]

According to the DTA guidelines, the Fraktur capital letter “J” should be transcribed either “I” (before consonant) or “J” (before vowel), since there is usually only one letter for these two phonemes in Fraktur typefaces. Here, it was accordingly transcribed “J” instead of “I”, because of the printing error “u” for “n”. Such problems transcribing “I” vs. “J”, which were brought to light through our study, led us to change the above-mentioned transcription rule: The Fraktur letter “J” will henceforth be transcribed as “J”, irrespective of its context.

All such cases were classified as transcription errors. In contrast, printing errors of the text source that were silently reproduced by the typists were not taken into account in measuring the transcription accuracy, since as far as the equivalence of the transcription with the source texts is concerned, the affected words were transcribed correctly.

Prospects

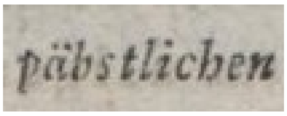
The DTA corpus is currently being extended in two ways. First, in the course of the second phase of the DTA project, earlier texts dating from 1600 to 1780 are being digitized. Second, in the course of the subproject DTAE (DTA Extensions),[22] our corpus is being extended by texts from all time periods represented in the DTA that have themselves been digitized by other projects. This way, we provide a basis for comparative corpus research by offering a balanced historical core corpus as well as specialized sub-corpora focusing on particular discourses. In addition, DTAE represents a platform for publication of digitized historical texts which would otherwise not be available to a broader scientific public.

The older our texts get, the more likely we will have to deal with transcription and annotation difficulties. For example, text structuring layouts in the master copies will become more and more complex, texts will increasingly contain special characters or foreign language material, and German text material generally will appear in Fraktur typeface with a large variety of decorated letters.

Therefore, the knowledge about potential error sources and frequencies of error types that we gained through our study will enable us to periodically adjust our quality control methods. For example, in DTAQ, all ticket information (i.e. information about the errors that were found through proofreading) is stored in a database and therefore available for further analysis. Based on these data, potential error sources in particular texts (e.g. changes between typefaces within one text, similar appearance of different letters) can be anticipated and avoided by providing typists with representative examples. In addition, lists of frequent errors can be communicated to the typists in order to attract special attention to similar cases. Subsequent to text recognition, new texts can be searched for erroneous strings that occurred repeatedly in the proofread texts and that are uncommon in the German language (see example 6).

[22] DTAE: <http://www.deutschestextarchiv.de/dtae>.

Example 6. Uncommon string “cb” due to transcription error



päbstlichen

Wrong transcription:

```
päbstlicben
```

Right transcription:

```
päbstlichen
```

[excerpt from http://www.deutschestextarchiv.de/moser_politische01_1796/102]

Therefore, based on the results of our study, we compiled a large list of certainly erroneous words or strings, which may be applied regularly to the DTA corpus in order to detect errors in an efficient way. However, this error detection method may only be applied semi-automatically, since each potential error needs to be reviewed and classified as transcription or printing error manually in order to be treated properly.

Conclusion

The project Deutsches Textarchiv aims to create a text corpus of German printed works dating from 1600 to 1900 that is suitable for linguistic research on the development of the (historical) New High German language. In this context, in order to avoid misinterpretations (e.g. concerning questions of spelling variation, lexical variation, or specifics of the print itself), it is necessary to ensure that the transcribed texts are highly accurate representations of the text sources. Since the extent of the text corpus may have an impact on the results of linguistic analyses as well, it is necessary to apply transcription and annotation methods which allow for the digitization of large amounts of text in a justifiable amount of time, while leading to highly accurate results as well.

The study presented in this paper aimed to measure the correctness of the double-keying transcription method, which is estimated to be highly accurate and therefore generally applied very frequently by text digitization projects. It has been the most important digitization method for the DTA corpus texts as well. In order to measure the accuracy of double-keyed texts, a subcorpus of 7,208 pages chosen from the DTA corpus with respect to different criteria was proofread by 22 different persons using the quality assurance platform DTAQ. The results of this extensive proofreading task confirmed that the transcription accuracy of double-keying is very high, showing that this method can be considered suitable for the digitization of historical printed texts. However, the quality assurance methods applied by the DTA in advance are likely to have a positive effect on the transcription quality. As a result of our study, we were able to gain insights on where transcription problems are still likely to occur and how our quality assurance methods may be improved in order to avoid errors in advance or efficiently eliminate them subsequent to text recognition.

Acknowledgements

We want to thank our proofreaders for their support in establishing a representative sample of proofreading pages.

References

DFG (Deutsche Forschungsgemeinschaft). 2009. *Scientific Library Services and Information Systems (LIS): DFG Practical Guidelines on Digitisation*. Developed by the Subcommittee on Cultural Heritage. http://www.dfg.de/download/pdf/foerderung/programme/lis/praxisregeln_digitalisierung_en.pdf.

Furrer, Lenz, and Martin Volk. 2011. "Reducing OCR errors in Gothic script documents." *Proceedings of Workshop on Language Technologies for Digital Humanities and Cultural Heritage* (Hissar, Bulgaria, September 16, 2011, associated with RANLP 2011), 97–103. http://aclweb.org/anthology-new/W/W11/W11-4115.pdf.

Geyken, Alexander, Susanne Haaf, Bryan Jurish, Matthias Schulz, Jakob Steinmann, Christian Thomas, and Frank Wiegand. 2011. "Das Deutsche Textarchiv: Vom historischen Korpus zum aktiven Archiv." *Digitale Wissenschaft. Stand und Entwicklung digital vernetzter Forschung in Deutschland, 20./21. September 2010, Köln. Beiträge der Tagung, 2., ergänzte Fassung*, edited by Silke Schomburg, Claus Leggewie, Henning Lobin, and Cornelius Puschmann, 157–161. Köln: Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen (hbz). http://www.hbz-nrw.de/dokumentencenter/veroeffentlichungen/Tagung_Digitale_Wissenschaft.pdf.

Geyken, Alexander, Susanne Haaf, Bryan Jurish, Matthias Schulz, Christian Thomas, and Frank Wiegand. 2012. "TEI und Textkorpora: Fehlerklassifikation und Qualitätskontrolle vor, während und nach der Texterfassung im Deutschen Textarchiv." *Jahrbuch für Computerphilologie online.* http://www.computerphilologie.de/jg09/geykenetal.pdf.

Holley, Rose. 2009a. "How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs." *D-Lib Magazine* 15 (3/4). doi:10.1045/march2009-holley.

———. 2009b. "Many Hands Make Light Work: Public Collaborative OCR Text Correction in Australian Historic Newspapers." N.p.: National Library of Australia. http://www.nla.gov.au/ndp/project_details/documents/ANDP_ManyHands.pdf.

Jurish, Bryan. 2012. "Finite-state Canonicalization Techniques for Historical German." Ph.D. diss., Universität Potsdam. urn:nbn:de:kobv:517-opus-55789; http://opus.kobv.de/ubp/volltexte/2012/5578.

Jurish, Bryan, Marko Drotschmann, and Henriette Ast. Forthcoming. "Constructing a Canonicalized Corpus of Historical German by Text Alignment." In *New Methods in Historical Corpus Linguistics, volume 3 of Corpus Linguistics and Interdisciplinary Perspectives on Language (CLIP)*, edited by Paul Bennett, Martin Durrell, Silke Scheible, and Richard J. Whitt. Tübingen: Narr.

Levenshtein, Vladimir Iosifovich. 1966. "Binary Codes Capable of Correcting Deletions, Insertions and Reversals." *Soviet Physics Doklady* 10(8): 707–10.

Nartker, Thomas A., Kazem Taghva, Ron Young, Julie Borsack, and Allen Condit. 2003. "OCR correction based on document level knowledge." In *Proceedings of the IS&T/SPIE 2003 International Symposium on Electronic Imaging Science and Technology* 5010 (Document Recognition and Retrieval X, Santa Clara, CA, January 20, 2003), 103–10. doi:10.1117/12.479681.

Tanner, Simon, Trevor Muñoz, and Pich Hemy Ros. 2009. "Measuring Mass Text Digitization Quality and Usefulness: Lessons Learned from Assessing the OCR Accuracy of the British Library's 19th Century Online Newspaper Archive." *D-Lib Magazine* 15(7/8). doi:10.1045/july2009-munoz.

Unsworth, John. 2011. “Computational Work with Very Large Text Collections: Interoperability, Sustainability, and the TEI.” *Journal of the Text Encoding Initiative* 1. http://jtei.revues.org/215.

Keywords

1. double-keying
2. quality control
3. error classification
4. digitization
5. tools
6. transcription accuracy
7. proofreading

Biographical statements

Susanne Haaf

e-mail: haaf@bbaw.de

Susanne Haaf works as research assistant/coordinator for the German Text Archive at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW).

Frank Wiegand

e-mail: wiegand@bbaw.de

Frank Wiegand works as research assistant/software developer for the German Text Archive at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW).

Alexander Geyken

e-mail: geyken@bbaw.de

Alexander Geyken works at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW) and is head of the project groups of the Digital Dictionary of German language (DWDS, a long-term BBAW-project), and the German Text Archive.