Quality Assurance of Large TEI Corpora

Frank Wiegand Alexander Geyken (Deutsches Textarchiv — DTA)

Deutsche Forschungsgemeinschaft DFG
berlin-brandenburgische Akademie der Wissenschaften

About the Project

The DFG-funded *Deutsches Textarchiv* started in 2007 and is located at the Berlin-Brandenburgische Akademie der Wissenschaften (BBAW). Its goal is to digitize a large cross-section of German texts from 1650 to 1900, annotated in TEI format. The DTA presents almost exclusively the first editions of the respective works.

- 700+ books (one more every day)
- ~260.000 pages (400M characters)
- double keying, OCR
- texts from external submitters (DTAE)
- a corpus balanced with respect to text genres
- all encoded in one XML/TEI dialect (DTA "base format")

Problem Statement

Data entry companies using a double keying process report a 99.95 % accuracy of the transcriptions. However, mistakes in the transcription process are not the only source of errors.

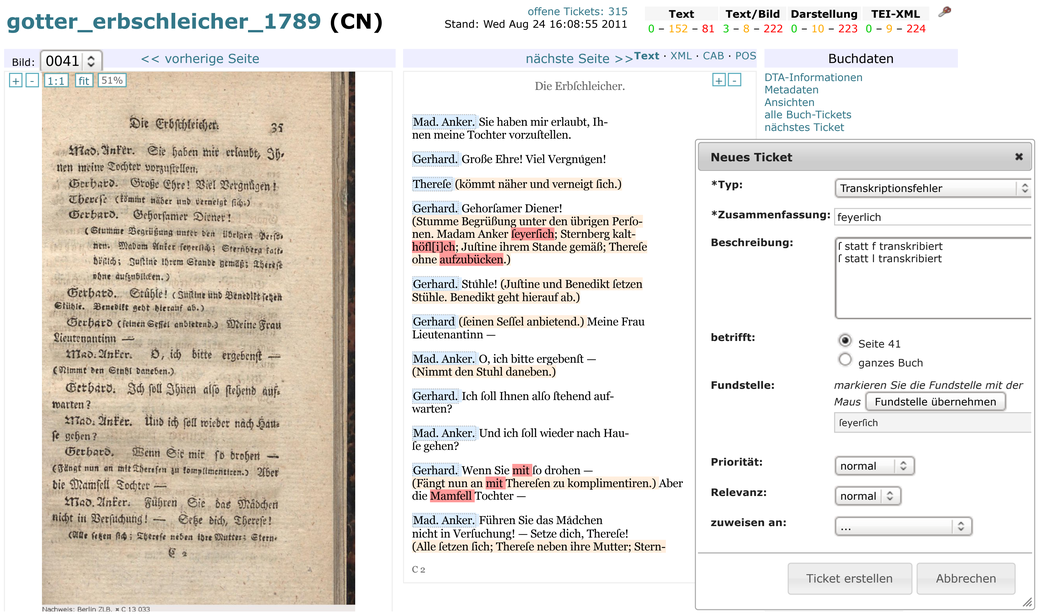
Quality assurance (QA) has to take into account other levels of error prone representations and tasks, namely metadata, XML annotation, HTML presentation (and other media), and the adequacy of workflow. DTAQ is a QA system dealing with all potential errors: They need to be reported, stored and fixed.

We `use Modern::Perl; # and more`

The backend of DTAQ is built upon many open source packages. Using *Perl* as a glue, the system runs on *Catalyst*, connects to a *PostgreSQL* database via *DBIx::Class* and builds its web pages with *Template Toolkit*. The frontend makes heavy use of *jQuery* and *Highcharts JS* to create a very interactive and responsive user interface.

Our XML/TEI files are automatically split up into individual pages and stored inside a *git* repository. The development of DTAQ itself also occurs within a distributed *git* repository.

A Web Based Frontend — DTAQ



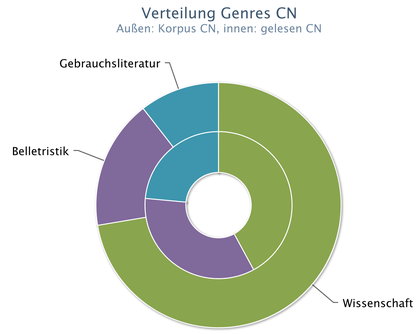
DTAQ provides a parallel view between the original image and the transcribed text (alternatively XML/TEI, CAB, or Part of Speech)

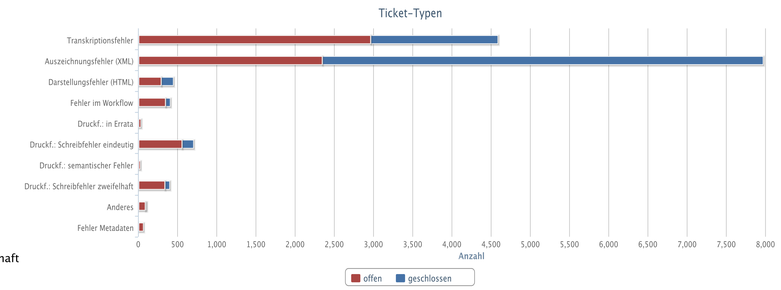
Quality Assurance in a Collaborative System

DTAQ is a browser-based tool to find, categorize, and correct different kinds of errors. Using a simple authentication system combined with a fine granularity access control system, new users can easily be added to our QA system. The GUI of our tool is highly customizable, so we can offer different views of our source images, transcriptions, and presentations.

Our linguistic tools (*CAB*) are integrated into this environment, not only to check their performance for errors, but also to provide other views of our texts. To avoid unnecessary repetitions in proofreading, users can mark pages as proofread. Using this technique, we are also able to provide several quality stages of our pages or books.

Statistics and Analyses





All tickets and proofread pages are stored within a database, thus DTAQ provides in-depth analysis and visualisation about the accuracy of the DTA corpus (cf. Haaf et al. TEI MM Würzburg 2011).

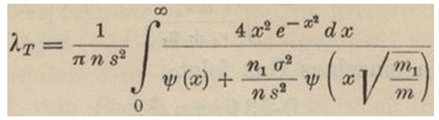
Integrated Ticketing System

Spotting an error leads to creating a ticket in our QA system, which may be classified, commented and assigned by the user like in a software bugtracking system. To keep track of thousands of reports, administrators can create importance levels, blockers, and milestone lists. Due to the fact that the physical position of each character is encoded in the XML/TEI, each erroneous spot can be highlighted on the digital image. Since 2011/06 we got:

- ~15.000 tickets created (7.700 solved)
- ~8.000 pages proofread

Integrated Formula Editor

As of October 2011, we have ~20.000 formulae annotated within our XML/TEI, using the `<formula/>` element as a placeholder to mark their appearance. Since most of them are not transcribed yet, DTAQ provides an integrated formula editor to help users to create TEX transcriptions.

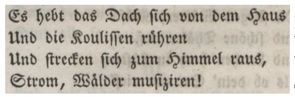


$$\lambda_T = \frac{1}{\pi n s^2} \int_0^\infty \frac{4 x^2 e^{-x^2} dx}{\psi(x) + \frac{n_1 \sigma^2}{n s^2} \psi\left(x \sqrt{\frac{m_1}{m}}\right)}$$

$$\lambda_T = \frac{1}{\pi n s^2} \int_0^\infty \frac{4x^2 e^{-x^2} dx}{\psi(x) + \frac{n_1 \sigma^2}{n s^2} \psi\left(x \sqrt{\frac{m_1}{m}}\right)}$$

Example from: *Boltzmann, Ludwig: Vorlesungen über Gastheorie. Bd. 1. Leipzig: Barth, 1896, page 73.*

Integrated CAB View



Es hebt das Dach sich von dem Haus
Und die Kulissen rühren
Und strecken sich zum Himmel raus,
Strom, Wälder musizieren!

For each text, DTAQ provides a transformation to a normalised modern spelling form using the *Cascaded Analysis Broker* – Bryan Jurish, JLCL 2010, 25(1).

Links & Contact

http://deutschestextarchiv.de/
http://deutschestextarchiv.de/dtaq

e-mail: dta@bbaw.de
phone: +49 (0)30 2037 0523