Historical Text Corpora for the Humanities and Social Sciences. Digitization, Annotation, Quality Assurance and Analysis

Workshop at the Joint "Culture & Technology" and CLARIN-D European Summer School „Digital Humanities & Language Resources “ Leipzig, July 22nd – 26th, 2014

Susanne Haaf, Christian Thomas
Deutsches Textarchiv, BBAW – Berlin
haaf@bbaw.de, thomas@bbaw.de
Introduction:
Building up a text corpus from scratch –
What is there to do?
Preliminaries for Corpora of Historical Texts

• Purpose of the corpus
  – Which research questions should my corpus be able to answer?
  – Which methods do I want to apply for corpus analysis?

• Text selection
  – Topic specific or reference corpus?
  – If reference corpus: What should it represent?
  – If reference corpus: What texts are representative for my field of interest?

• Sample size
  – Entire documents or sections?
  – If sections: Which size would be straightforward?
Preliminaries for Corpora of Historical Texts

• Methods for Quality Assurance:
  – Which text recognition method would suit my material? (OCR vs. single or double keying)
  – How can I improve the recognition quality? (automatic methods vs. manual proofreading)

• Annotation
  – Which structures?
  – Which linguistic information?
  – Suitable tagsets which also ensure homogeneity
  – Necessary annotation depth
  – Automatic vs. manual annotation
Preliminaries for Corpora of Historical Texts

• Metadata
  – Which metadata belong to my corpus, the single texts, their sources?
  – How do I get and record them?

• Intended life-cycle of corpus
  – Which formats would be sustainable?
  – Which standards should I use?
  – How do I achieve long-term availability and archiving?
  – How do I achieve interoperability?
Preliminaries for Corpora of Historical Texts

• Documentation
  – Which parts of my workflow should be documented?
  – How should they be documented? (for understanding vs. applying the workflow)

• Infrastructure
  – How to collect data?
  – How to store it?

• Time management
  – Where do I have to compromise? (nice to have vs. project capacities)
Ok, so I decided on all the preliminaries!

What next?
Task:

This historical text should be made part of my text corpus.

→ What is there to do?
Getting images for the texts in my selection

Alonso de Casarubios: Compendium privilegiorum fratrum Minorum. 1532, Original: Bayerische Staatsbibliothek; http://books.google.de/books?id=R6g5AAAACAAJ&hl=de&pg=PT129#v=onepage&q&f=false

Nota bene: DFG-Guidlines for Digitization

DFG-Praxisregeln
„Digitalisierung“

http://www.dfg.de/formulare/12_151/12_151_de.pdf
Which way of text recognition is suitable for the material?
Optical Character Recognition (OCR)?

Fontane 1894

Desse Schein ich nicht mal sah, nur die zusätzliche Veranlassung gab. Meine Eltern befanden sich an jenem Tage auf einem kleinen Diner, ganz am entgegengesetzten Ende der Stadt. Als die Tischgesellschaft von der Nachricht, dass alle Schauen in Feuer stünden, überrascht wurde, stand es für meine Mutter, die eine sehr nervöse Frau war, sofort fest, dass ihre Kinder mit verbrennen müssten oder mindes... (Text unterbrochen)

Corvinus 1715

Betten fließen, Heißer aus Bördent wend... (Text unterbrochen)


Sonst ist betrachtungswürdig / dass diese beyde Hochfürstlichen Damen auch Oij selbst

Double Keying: common method for large text collections

- 2 persons are typing the same text
- Both versions are compared automatically
- Differences between both versions are judged by an expert
Text recognition for the DTA core corpus

- Main part (> 1100 texts) recognized via double keying; very high accuracy rate (> 99.98%) cf. Haaf/Wiegand/Geyken 2013 (99.9909%)

- Smaller part (~200 texts) recognized via OCR; a lot of effort necessary for post-correction
Transcription Guidelines
Task:

This historical text should be made part of my text corpus

→ What is there to do?
Enter Theseus, Hippolita, with others.

Theseus.

Ow faire Hippolita, our nuptiall houre
Drawes on apace: foure happy daies bring in
Another Moon: but oh, me thinkes, how slow
This old Moon wanes; She lingers my desires
Like to a Step-dame, or a Dowager,
Long withering out a yong mans reuenuew.

Foure daies wil quickly steep thefeleues in nights
Foure nights wil quickly dreame away the time:
And then the Moone, like to a siluer bow,
Now bent in heauen, Shal behold the night
Of our solemnities.

The. Go Philostrate,
Stirre vp the Athenian youth to merriments,
Awake the pert and nimble spirit of mirth,
Turne melancholy forth to Funerals:
The pale companion is not for our pompe,
Hippolita, I woo'd thee with my sword,
And wonne thy love, doing thee injuries:
But I will wed thee in another key,
With pompe, with triumph, and with revelling.

Enter Egeus and his daughter Hermia, Lysander,
and Demetrius.

Ege. Happy be Theseus, our renowned Duke.
The. Thanks good Egeus, what's the news with thee?
Ege. Full of vexation, come I, with complaint
Against my childe, my daughter Hermia.

Stand forth Demetrius.

With cunning haft thou filch'd my daughters heart,
Turn'd her obedience (which is due to me)
To stubborne harshnesse. And my gracious Duke,
Be it so she will not heere before your Grace,
Consent to marrie with Demetrius,
I beg the ancient privilege of Athens;
As she is mine, I may dispose of her;
Which shall be either to this Gentleman,
Or to her death, according to our Law,
Immediately provised in that case.

The. What say you Hermia? be advis'd faire Maide.
To you your Father should be as a God;
One that compos'd your beauties; yea and one
To whom you are but as a forme in waxe
By him imprinted: and within his power,
To leave the figure, or disfigure it:

Demetrius is a worthy Gentleman.

Her. So is Lysander.

The. In himselfe he is.
But in this kinde, wanting your fathers voyce.
The other must be held the worthier.

Her. I would my father look'd but with my eyes.
The. Rather your eies must with his judgment looke.

Her. I do entreat your Grace to pardon me.
I know not by what power I am made bold,
Nor how it may concerne my modestie
In such a presence heere to pleade my thoughts:
But I beseech your Grace, that I may know
The worst that may befall me in this case,
If I refuse to wed Demetrius.
The. Either to dye the death, or to abide
Transcription Guidelines

• Normalizations?
  – Punctuation
  – Orthography

• Handling of printing/writing errors

• Handling of historical characters, e.g.:
  – Long s (ſ) vs. round s (s):
  – Sharp s (ß):
  – Capital letter I vs. J:
  – Writing of Umlaut:

→ DTA Transcription Guidelines: http://www.deutschestextarchiv.de/doku/richtlinien