Towards the Integration of Language Tools within Historical Digital Libraries

Cristina Vertan
University oh Hamburg, Research Group "Computerphilology"
Von-Melle Park 6, 20146 Hamburg, Germany
E-mail: cristina.vertan@uni-hamburg.de

Abstract

During the last years the campaign of mass digitization made available catalogues and valuable rare manuscripts and old printed books vie the Internet. The Manuscriptorium digital library (Uhli and Knoll 2009) ingested hundreds of volumes and it is expected that the volume will grow up in the next years. Other European initiatives like Europeana (Europeana 2009) and Monasterium (Monasterium 2009) have also as central activities the online presentation of cultural heritage.
With the growing of the available on-line volumes, a special attention was paid to the management and retrieval of documents within digital libraries. Enabling semantic technologies and intelligent linking and search are a big step forward, but they still do not succeed in making the content of old rare books intelligible to the broad public or specialists in other domains or languages.
In this paper we will argue that multilingual language technologies have the potential to fill this gap. We overview the existent language resources for historical documents, and present an architecture which aims at presenting such texts to the normal user, without altering the character of the texts

1. Types of Documents within a Digital Library for Old Books

A special particularity of such digital libraries is the fact that the great part of existing objects are high resolution images. In contrast with digital libraries containing modern texts, where OCR technologies are enough powerful to provide text versions of scanned pages, for old printed books and even more for old manuscripts is OCR technique far from being used on large scale with high precision and recall. Especially for manuscripts one faces the problem that one manuscript was written by several hand and there is not enough training material for automatic recognition. Additional problems are made by illuminations and annotations made by different readers of the manuscript on the border.

Recent advances in palaeography (Stansbury 2009) made however possible the identification of different hands within a manuscript by means of statistical methods. Without having the OCR technology, old books remain mostly in electronic form as images. Only few of them are transcribed.

Another particularity of old books and manuscripts is the existence of so called "manuscript descriptions". These are meta-data provided by researches (codicologists). These manuscript descriptions have the big advantage that they are written in modern languages and present not only physical characteristics of the book (bindings, covers transmission) but also report extensively about the Content. The main disadvantage of such manuscript descriptions is that their language is understandable only for the specialists: they are full of abbreviations, not syntactical sentences, and contain a big number of specialised terms, as one can see in Example 1.

The language of manuscript descriptions can be considered however a specialised case of "controlled language", which means:

1. it can be expanded to natural language following a set of rules and
2. it can be translated following methods of controlled language.

```
<msItem>
<locus>(ff. 1v, 3 - 8v, 2rv, 9 &#8212; 333)</locus>
Mete. I &#8212; II, inc. mutile …] I 3, 339b13.
Lücke: II 9, 369b11 κωπηλασίας[…]370a32 (=
Blattausfall nach f. 332).
</msItem>
<msItem>
<locus>(ff. 1rv, 3 &#8212; 8v, 2rv, 9 &#8212; 334v,
mit dem Text von Mete. alternierend)</locus>
Alexander von Aphrodisias, In Mete. comm. (CAG
III 2), inc. mutile …] S. 8, 10, des. mutile S. 132, 17
φερομένη ς [… Lücke: S. 128, 36 […] εναι δέ τινας
S. 131, 22 (= Blattausfall nach f. 332). Auslassung
(f. 1): S. 8, 12.
</msItem>
```

Example 1. TEI-annotated manuscript description – content part

Additional materials accompanying a digital manuscript are : metadata describing its structure and sometimes lists ob bibliographical materials and /or lists of watermarks associated with the manuscript.

During the last years one could observe a shift in the presentation of such manuscripts: from animated Presentation by means of proprietary multimedia tools, to separate storing and description of each manuscript page. This corresponds to the shift of the interest from

simple digital preservation towards digital working space, in which one can access and interrelate different parts of manuscripts. Additionally through projects like TextGrid (TextGrid 2009), tools for annotations inside images became possible. This allows linking on manuscript descriptions references to certain parts of the manuscript with the image respectively targeted part of the associated image.

2. Challenges of Language Technology for Old Texts

When trying to apply language technology for old texts, one should get rid of following false assumptions, which hold for modern documents:

- documents have one author,
- documents are written in one language,
- have a well defined orthography,
- have definite grammatical rules,
- belong to a homogenous linguistic and historical layer
- are readable by every speaker of the language,

are part of a vast homogenous corpus.

According to the levels of processing in modern texts we identify the following major problems with historical texts:

a) Morphological and syntactical problems can be resumed under the following points:

- Sentences were longer in medieval times,
- Medieval German e.g. tends to build long sequences of adjectives and nouns, esp. in legal documents,
- syntactic means changed: e.g. word order and composition,
- inflexion system changed (Example: shift from strong to weak verb paradigm, loss of imperfect in favour of perfect),
- Syntactic change is a continuous flow, because there is no sequence of static rules,
- Similarly, textual rules changed:

b) semantical and conceptual problems occur because:

- obsolete words occur in texts,
- lexical semantics of known words changes over time,
- idioms change
- false friends over periods (Germ. übel, wohl),
- even technical terms change (Germ. Verleger),
- references might be wrong and might look like a translation error,
- medieval texts often convey information about heterogenous fields (The Physiologus about Jesus Christ, stones and animals).

All these kind of exceptions make rule-based approaches extremely difficult and less robust. Statistical methods suffer also from the frequent lack of homogenous large corpora. There is also a stylistic component which should not be neglected namely the adaptation of texts to the modern languages risks to modify the original character of the text.

In the following section we will present a flexible architecture which combines modules from controlled language translation, MAT and language generation with the aim of explaining old texts to broad public.

3. System Architecture

Given the fact that a translated transcription (word-to-word) does not meet the needs of most users and a replica of old texts in other old languages is historically impossible, we propose an tool integrating elements of machine aided translation, natural language generation starting from templates, and controlled language translation, which allows for

–a modern narrow paraphrase in a second language, and
–a broad documentation of the source text, target text, and translation.
- an augmented translation of the attached manuscript description (translation of the controlled language and generation from templates of non-specialised texts).
We present the system architecture in Figure 1

We are implementing for the moment this architecture for the Teuchos-System (Teuchos 2010). The Teuchos Center for Manuscript and Text Research was set-up in 2007 by the Institute for Greek and Latin Philology of the University of Hamburg in cooperation with the Aristoteles –Archive at the Free University Berlin. Teuchos is a long-term infrastructure project, which is financed in its starting phase (until mid-2010) by the German Research Foundation according to the funding scheme "Thematic Information networks" in the framework of the Scientific Library Services and Information Systems programme.

In its final form Teuchos is to provide a web-based knowledge portal suited for manuscript and textual studies, offering tools for capturing, exchange and collaborative editing of primary philological data. The data shall be made accessible to the scholarly community as primary or raw data in order to be reusable as source material for various individual or collaborative research projects.

There are several groups of digital objects to be stored in the repository:
We store tracings of watermarks from dated paper

manuscripts as digital images on the one hand, and descriptive data on these watermarks and their motif groups in an XML format on the other. Images are associated with Dublin Core -like information about the data and linked to the descriptive metadata.

The textual transmission group is divided into two subgroups that are themselves subdivided: material related to individual manuscripts and material related to a particular work, e.g. a particular source text by a particular author.

The manuscript group encompasses digital page images of manuscripts (or parts of manuscripts) that are aggregated on a per manuscript basis scholarly manuscript descriptions that may reference page images if available for the one manuscript described, and transcription data, which may range from a first set of basic structural data to full transcriptions, and usually links to pages of exactly one manuscript (exceptions are e.g. texts spanning more than one manuscript volume and re- or miss-bound manuscripts).

The group of works encompasses a wide range of materials referring to a source text with its entire set of manuscripts rather than to one particular witness, and ranges from full critical editions (with several intermediate stages) and translations to various kinds of commentaries (and other explanatory or descriptive materials).

The biographical dictionary group hosts information on the biography of historical persons relevant to our other research materials.

A special group is dedicated to research papers that may reference material from the other groups, without themselves falling into any of the other categories.
Finally, bibliographic data (also including online resources) pertinent to specific research areas is collected and stored using a separate group of objects.
Whilist bibliographic –data, watermarks and research papers may be of interest only for researchers in classical philology, works on Aristoteles, could be relevant also for a broader public like philosophers, researchers on Middle –Age, mathematicians. Therefore this is a relevant use-case for the system we intend to implement.

We constructed a corpus of about 200 manuscript descriptions in German. They are annotated conform TEI, not only at the macro-level but also contain phrase elements, marking the folios, the segments in other languages than German the person and the topographical names. We are now working on extracting a database of examples covering frequent expressions occurring within these descriptions. Consequently the translation of the controlled language will rely on an Example-based paradigm. First tests will be done on English , followed by other languages used within the community.

The Lexicon material is ensured through a web interface to Perseus Digital library (Perseus 2010).

4. Conclusions

In this paper we present a flexible architecture aiming to integrate different language technology tools in order to make understandable materials within historical digital libraries to a broader public. We explain the challenges of the approach and describe the proposed system as well as the current implementation of the use-case.

References

Uhlii, Z. and Knoll, A. (2009), “Manuscriptorium Digital Library and the ENRICH Project –Means for dealing with digital palaeography and codicology”, in “Paleography and Codicology in Digital Age, Malte Rehbein,Patrick Sahle,Torsten Schaßan (Eds.), Schriften des Instituts für Dokumentologie und Editorik, Müchen, 2009, pag. 67-96

Europeana (2009) www.europeana.eu/

(Monasterium (2009) http://www.monasterium.net/

(Stansbury M.(2009), “Teh computer and the classification of script”, in “Paleography and Codicology in Digital Age, Malte Rehbein,Patrick Sahle,Torsten Schaßan (Eds.), Schriften des Instituts für Dokumentologie und Editorik, Müchen, 2009-11-08, pag. 237-250

(TextGrid (2009) http://www.textgrid.de/

(Teuchos (2010) http://beta.teuchos.uni-hamburg.de

(Perseus (2010) http:// www.perseus.tufts.edu

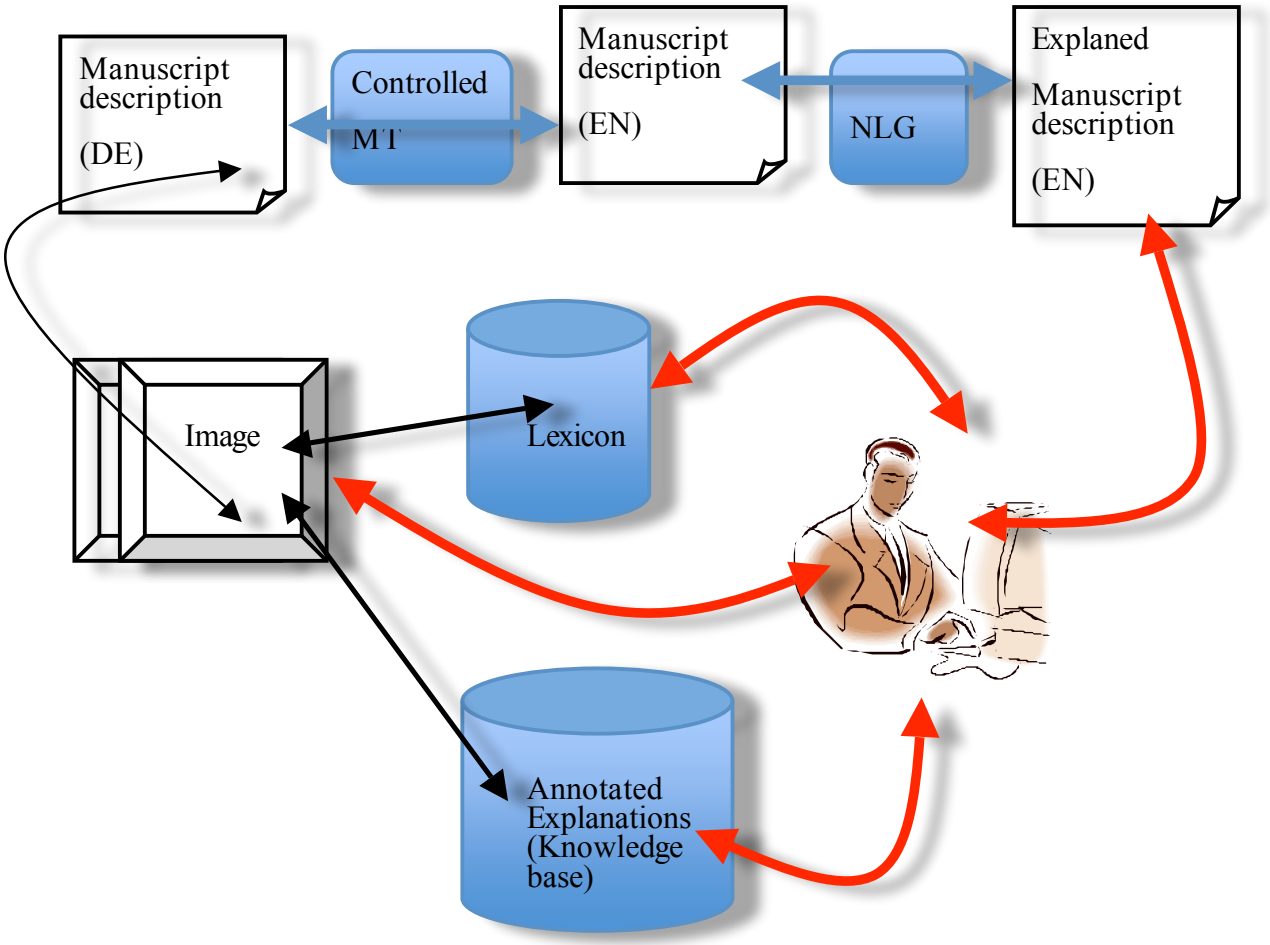

Figure 1. system Architecture