Making historical texts accessible to everybody

Cristina Vertan
University of Hamburg
Vogt-Kölln Strasse 30
22529 Hamburg
cristina.vertan@uni-hamburg.de

Walther v. Hahn
University of Hamburg
Vogt-Kölln Strasse 30
22529 Hamburg
vhahn@informatik.uni-hamburg.de

Abstract

In this paper we discuss the degree of readability of historical texts for a broad public. We argue that text simplification methods can improve significantly this aspect and bring an added value to historical texts. We present a specific example, a genuine multilingual historical texts, which should be available at least to researchers from different fields and propose a mechanism for simplifying the text.

1 Introduction

During the last decade there was a massive digitization campaign, which lead to a large number of electronicly available collections of historical documents. Most of these collections offer the possibility to navigate through the documents and display not only the associated metadata but also content. Thus researchers and students in various fields, which are related to one document's topic, may have access to it.

This, however, is not a barrier-free access as many historical languages either differ significantly from their modern correspondent or they are not at all in use any longer.

Thus only scholars can understand such texts with deep knowledge in the respective language(s). We use the plural form „languages" as most historical texts are multilingual, being composed from a mixture of paragraphs in one main text language and one or e more secondary languages which were either linguae francae at the time when the document has been written (e.g. Latin, Ancient Greek, Arabic) or reflect cultural or geographical particularities of the topic being described (e.g. a Latin Document written about the organisation of the Turkish empire).

Text simplification is a technique used up to now for making modern texts accessible to groups with special requirements (persons with disabilities, language learners). Simplification means a broad range of techniques from lexical replacements of less used terms by more frequent ones, through syntactic adaptation (substitution of relative clauses, elimination of long distance dependencies) up to reshaping on the discourse level (e.g. eliminating anaphora) (Saggion et. al. 2013), (Dornescu et. al. 2013). However, it is assumed that the users know orthography and morphology of the language.

In this paper we argue that text simplification is also an adequate method for making historical texts understandable for a broader public and we describe first approaches with a genuine multilingual historical text. The paper is organised as follows: in section 2 we introduce simplification requirements for historical texts and exemplify them by means of a particular scenario, which we will describe. Section 3 is dedicated to our approach towards text simplification. Finally in section 4 we present our first conclusions and further work to be done.

2 The Need of Text Simplification for historical Texts

As we mentioned in section 1, historical texts must suffer a certain transformation in order to be understood by non-trained readers. These transformations are language dependent and should satisfy two criteria:

- They should try to bring the text as close as possible to the modern language form (if available)
- They should preserve the cultural and geographical setting of the time when they were written.

In the following paragraph we will consider texts (originals or historical translations) for which a modern variant of the language is still in use.

As an example we will discuss the works of Dimitrie Cantemir, political figure, philosopher, historian, musician, geographer, who lived at the end of XVIIth. century and prepared two important works for the history of Eastern Europe for the Royal Academy of Sciences in Berlin. The first one, „Decriptio Moldaviae“ (The Description of Moldavia) is - as the title suggests - a detailed presentation of his (Cantemir's) native country Moldavia (spreading today from the eastern part of Romania to the current Republic of Moldavia). Cantemir describes the history of the country, as well as its geography, the language and the traditions of people living there. It includes also the first detailed map of the region. The work was written in Latin and translated into German and French at the beginning of the XVIIIth century, later into Romanian. The second work is „The History of Growth and Decay of the Ottoman Empire“. This was again written in Latin and translated more or less immediately into German, French, English and Russian. It remained a reference work for studies of the Ottoman Empire until the middle of XIXth century.

Both works are thus relevant for historians but also for ethnographers, linguists, as well as for people interested in the history of these three territories.

The fact that they were translated seems to make their reception easier. However we will show through several examples that this is a false assumption. The following examples illustrate also the need of text simplification at four linguistic levels: orthography morphology, syntax and semantics.

The following examples are extracted from the German translation from 1771 of "Descriptio Moldaviae" (Cantemir 1771). They are, however illustrative for a wide range of historical texts in other language combinations. A more detailed description can be found in XXXXXXXX

2.1 2.1. Orthographic level

In this text we encounter passages in German, Romanian, Latin and Ancient Greek. German Text is written in black-letter typeface. Latin and part of Romanian words (see below) are written with roman typeface. Greek paragraphs are easyly to detect and to isolate due to their specific alphabet.

Two approaches of the writer were identified when dealing with local (Romanian) names

- Named Entities (geographical names, person names) as well as names for specific roles in the army or society are written with black-letter typeface. They are adapted to the German pronunciation, like in the following examples:

  e.g. The river Prut became Pruth; The ruler Dragoș became Dragosch, the role of being a „pivnicier“ (person responsible for keeping wine and goods in the basement of a castle) became „pivnichar“

- Lexical items illustrating the language, remained in Latin font and were not adapted phonetically. However, as at Cantemir's time Romanian language was written in Cyrillic alphabet, the Latin-alphabet transcription is deviant from the current Romanian orthography

  e.g. "*muiere*" (colloquial term for woman) appears in text as "*mujere*".

2.2 Morphological and syntactic level

Old morphological forms deviant from those used in current German are present throughout the book

e.g. *„zweyten“* or *„Theil“* instead of *"zweiten"* and *"Teil"*

For any modern reader unknown named entities appear: e.g. „*in dem bergigten Theile von Moramor, (*)*“. Even in the text the *„Moramor“* region is not clearly identified and the text passage contains two footnotes “(*)”, one from Cantemir himself and one from the German translator from 1771, both commenting the word *Moramor*.

2.3 Semantic level

There are either words which still exist in the modern vocabulary but mostly used with a different meaning. An example is the word *„flüchtigen“* used in the XVIIIth century exclusively with the meaning of „*running away from somebody*“ whereas nowadays it is predominantly used with the meaning of „*volatile substance*“. The main challenge here is that both meanings were and are still valid through the whole period from XVIIIth century until now, just the usage frequency of one or the other meaning changed.

Time references are often relative. In an expression like *„von dem heutigen Ungarn“* (engl. „*from Hungary nowadays*”) one should understand and interpret the temporal expression „nowadays“ as referring to the time when the text was written (even not: published). This also implies that the corresponding political or geographical unit, in this case *„Hungary“* may have changed since that time.

2.4 Knowledge level

Additionally at knowledge level one can observe a different conceptual representation. We present here just one relevant example: It refers to geographical units / population groups, which changed their denomination or may refer to different entities depending of the historical/geographical context. In the sentence

> „*Die auf der andern Seite angränzende Polen und Russen nennen die Moldauer Wolochen, d. i.* ***Wälsch****e oder Italiäner, die Walachen aber, die auf dem Gebirge wohnen, heissen sie die Berg-Walachen, oder die Leute jenseits des Gebirges*

we find the term „***Wälsche“***. In Central Germany up to the last century *„Wälsche“* was the name for French, in Southern Germany for Italians and still today in Eastern Austria it is the name for Slovenians. Thus the term depends on the historical and geographical context and is not fixed to a specific population. However, readers may be confused without this background knowledge.

From the examples above it is clear that a non-trained reader (i.e. a reader being not familiar with Early New High German, Romanian and old terms in Romanian, Romanian geography and history) will have difficulties in reading and interpreting the text. We should mention here that there does not exist any modern German Edition of the text.

3 Text simplification for historical texts

We argue that a process of text simplification should take place at all above-mentioned levels. Some parts (the orthographical and morphological/syntactical level can be done semi-automatically through a rule-based process. Prerequisite is that the text is digitized and the information concerning the typeface and the font is preserved.

- STEP 1: Within the black-letter typeface paragraphs
  - Match each word against a German language model,
  - If a word is not matched but there are candidates from which it differs by 1 or 2 words try to apply normalization rules like (ey → ei, ev → eu),
  - In contrary match the word against a Romanian language model and try the same with a set of Romanian normalization rules. Words which could not be matched should be rendered to the user and proposed for manual correction.
- STEP 2: Within the Latin-typeface phrases
  - Match words against a Latin language model and a Romanian one

- Word which could be found in both should be rendered for manual annotation and language disambiguation,
- For words not found in the Latin model but with some close variants in the Romanian language model try to apply a Romanian normalization rule.

The output of Steps 1 and 2 will be a normalized text in with language is identified and marked for all paragraphs.

The paragraphs marked by "Romanian" have to be manually translated, i.e. explained to the reader in German or English.

Additional annotation is necessary to enable text processing for text simplification at upper levels. We propose an annotation scheme, aiming not only at marking words which could not be corrected throughout the normalization process, but enhancing also the meaning of the word (Vertan and v. Hahn 2014)

The main unit of the annotation is called „phrase". By phrase we mean a word or a multi-word expression. For each phrase the semantic frame includes information about the named entity (if any) as well as the obsolete meaning and the modern meaning of the word.

In figure 1 we present the structure of the annotation, while in figure 2 we present an example of an annotation as well as the possible linkage to a domain ontology

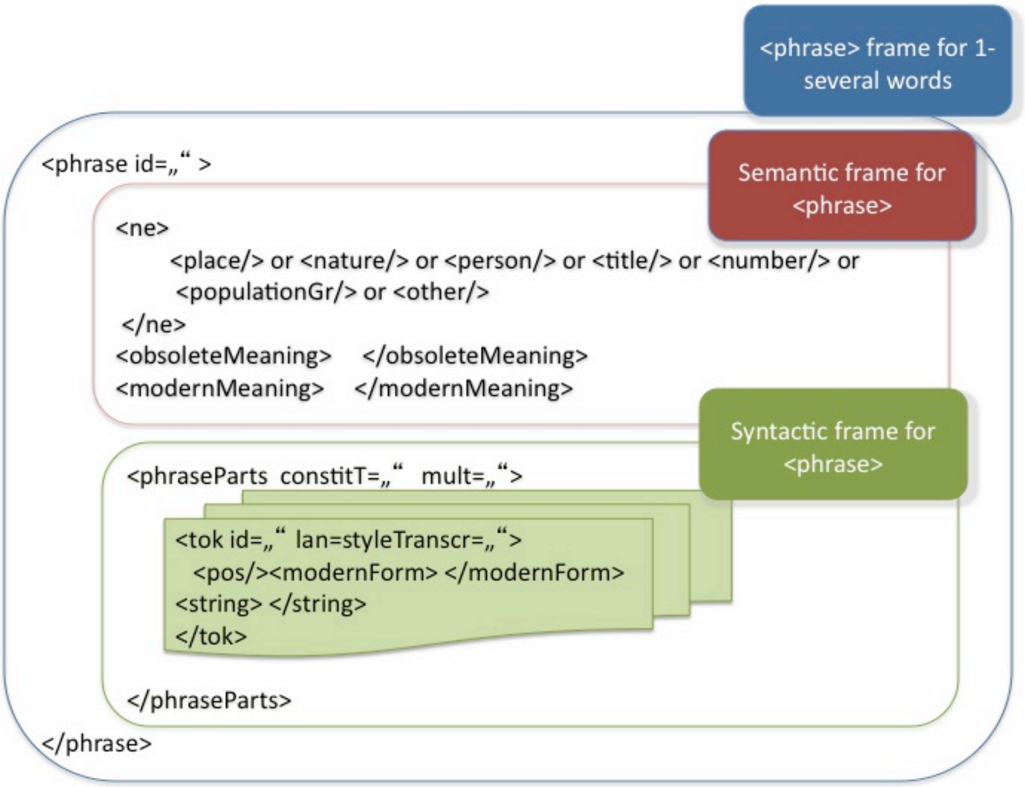
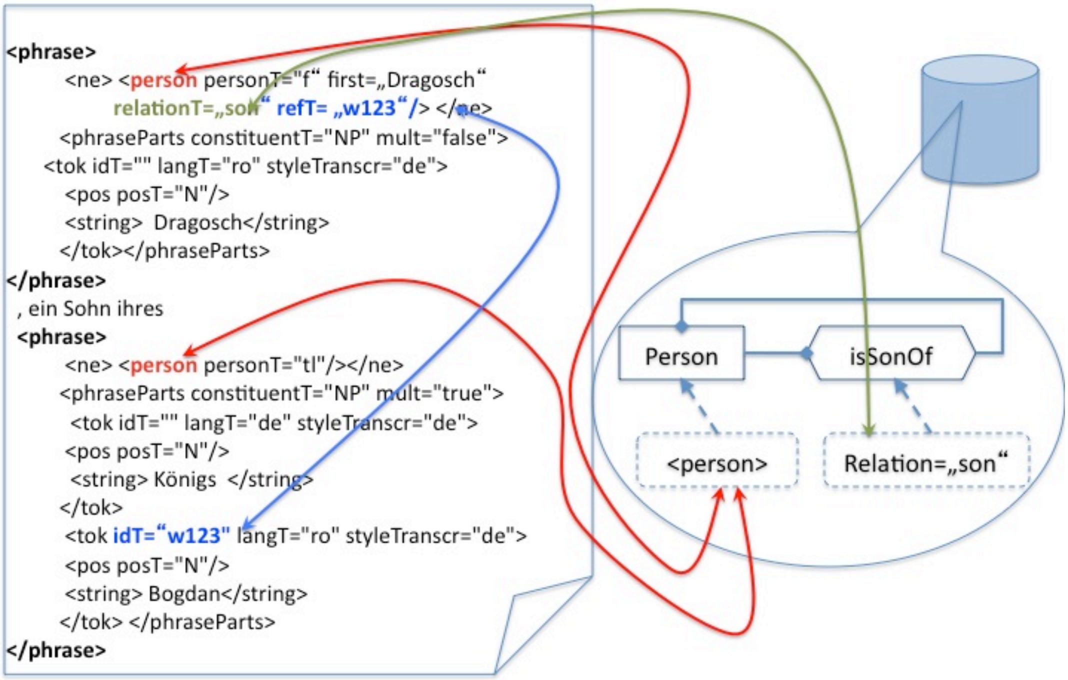

<phrase> frame for 1-several words
<phrase id=„" >
Semantic frame for <phrase>
<ne>
<place/> or <nature/> or <person/> or <title/> or <number/> or <populationGr/> or <other/>
</ne>
<obsoleteMeaning> </obsoleteMeaning>
<modernMeaning> </modernMeaning>
Syntactic frame for <phrase>
<phraseParts constitT=„" mult=„">
<tok id=„" lan=styleTranscr=„">
<pos/><modernForm> </modernForm>
<string> </string>
</tok>
</phraseParts>
</phrase>

Figure 1 Structure of the annotation scheme

Following this annotation step, a replacement of each annotated phrase with its modern form or in case of Romanian or Latin words with its translation will be obtained. Our first attempts in applying a rule-based constraint dependency parser (Beuck et. al. 2011) on such text were successful but this needs deeper investigation. The dependency parser can be used for identifying relative clauses and proposes candidates for further simplification.



```
<phrase>
    <ne> <person personT="f“ first=„Dragosch“
        relationT=„son“ refT= „w123“/> </ne>
   <phraseParts constituentT="NP" mult="false">
  <tok idT="" langT="ro" styleTranscr="de">
    <pos posT="N"/>
    <string>  Dragosch</string>
   </tok></phraseParts>
</phrase>
 , ein Sohn ihres
 <phrase>
    <ne> <person personT="tl"/></ne>
   <phraseParts constituentT="NP" mult="true">
    <tok idT="" langT="de" styleTranscr="de">
    <pos posT="N"/>
    <string> Königs  </string>
   </tok>
    <tok idT=“w123" langT="ro" styleTranscr="de">
    <pos posT="N"/>
    <string> Bogdan</string>
   </tok> </phraseParts>
</phrase>
```

Figure 2 Example of annotation scheme and ontology linking

4 Conclusions and further work

In this paper we showed that text simplification is a useful technique for making historical texts understandable for modern readers. We identified particularities of historical texts, which need special attention and pre-processing. In the second step we are able to apply state-of-the-art methods for text simplification. We propose an algorithm dealing with multilingual entries for text normalization.

Currently we are annotating manually the words rendered by the normalization process. Further work is planned for the application of the WCDG parser on the normalized text and selection of relative clauses, which can be either deleted or transformed into a main clause, in order to make sentences shorter and clearer. We intend also to exploit the existence of a comparable corpus containing translations of the same text in five languages (Vertan 2014).

References

Niels Beuck, Arne Köhn, and Wolfgang Menzel. Incremental parsing and the evaluation of partial dependency analyses In DepLing 2011, Proceedings of the 1st International Conference on Dependency Linguistics, 2011.

Cantemir Dimitire Beschreibung der Moldau. *Faksimildruck der Original Ausgabe von 1771, Maciuca C. (Ed).* , Bukarest, Kriterion Verlag, 1973

Iustin Dornescu, Richard Evans and Constantin Orasan, A tagging Approach to Identify Complex Constitutents for Text Simplification, in Proceedings of Recent Advances in Natural Language Processing (RANLP 2013), Hisar, Bulgaria, pp.221-229

Horacio Saggion, Elena Gomez-Martinez, Alberto Anula, Lorena Bourg, Esteban Etayo, Text Simplification in Simplext: Making texts more Accessible, last retrieved at www.simplext.es

Cristina Vertan and Walther v. Hahn, Discovering and Explaining Knowledge in historical Documents, Proceedings of the Workshop on “Language Technology for Historical Languages and Newspaper Archives”, Kristin Bjnadottir, Stewen Krauwer, Cristina Vertan and Martin Wyne Eds., Workshop associated with LREC 2014, Rejkyavik Mai 2014, pages 76-79

Cristina Vertan, Less explored multilingual issues in the automatic processing of historical texts – a case study, Proceedings of the Digital Humanites Conference 2014, Lausanne, pages 406-407