Modelling linguistic vagueness and uncertainty in historical texts

Cristina Vertan, Walther v. Hahn
University of Hamburg Department of Computer Science, Research Group"Computerphilologie", Vogt-Kölln Strasse 30, 22527 Hamburg, Germany
cristina.vertan@uni-hamburg.de,
vhahn@informatik.uni-hamburg.de

Abstract

Many applications in Digital Humanities (DH) rely on annotations of the raw material. These annotations (inferred automatically or done manually) assume that labelled facts are either true or false, thus all inferences started on such annotations us boolean logic. This contradicts hermeneutic principles used by humanites in which most part of the knowledge has a degree of truth which varies depending on the experience and the world knowledge of the interpreter. In this paper we will show how uncertainty and vagueness, two main features of any historical text can be encoded in annotations and thus be considered by DH applications.

1 Introductions

Most Digital Humanities projects tend to collect data as facts in a (relational) data base. According to Wilhelm Dilthey humanities make use of a hermeneutic paradigm for establishing hypotheses. Accordingly, social data often consist either of texts mirroring attitudes, allegations, beliefs, etc., or are reactions of test subjects to verbal stimuli. Such material cannot reasonably be treated as facts like numbers or positive propositions. On the other hand, analysing only formal features in the material does rarely contribute to the hermeneutic aims of the investigation intended by a humanist researcher. In this paper we show by means of a particular historical corpus how vagueness and uncertain language features can be kept in annotations and used in reasoning engines.

1.1 Case study: The corpus of historical texts wriiten by Dimitrie Cantemir

Dimitrie Cantemir (1673-1623) was prince of Moldavia (historical region including regions from current eastern part of Romania, Republic of Moldavia and some parts from Ukraine), man of letters-philosopher, historian, musicologist, linguist, ethnographer and geographer. He received education in classical studies (Greek and Latin in his country of origin), then he lived for several years in Istanbul where he learned Turkish, and familiarized himself with the cultural traditions of the ottomans, meet important persons around the sultan and learned a lot about history of the Empire. After a very short period of being prince of Moldavia he was forced to immigrate to Russia, where he became an important person at the court of Tsar Peter the Great. During this period, his works gained attention in the Western countries. He became member of the Royal Academy in Berlin and, at their request, he produced the two books which are the subject of this project:

Descriptio antiqui et hodierni status Moldaviae, written in Latin, a history of his country in which he describes not only pure historical facts but also traditions, the language, the political and administration system. Local denominations and troponins, as well as names are written in Romanian with Latin script as his intention is to demonstrate the Latin origin of his folk. The transcriptions are not standardized and one retrieves for the same troponin several name variations. Quotations as known today are very rare, there is no bibliography. According to (Lemny 2010), as there was practically no consistent previous work about the region, Cantemir himself was not particularly careful with indicating sources of knowledge. The work is accompanied by a map, the first detailed cartography of the region. The names on the map are in Romanian language. The Latin original was translated for the first time in German, and only later at the middle of the XIXth century in Romanian. The Latin manuscript seemed to be lost for a long time, so that the first Romanian translation was following the German one. The German translation is containing editorial notes of the translator

Proceedings of the Workshop on Language Technology for Digital Historical Archives in conjuction with RANLP-2019, pages 34–38,
Varna, Bulgaria, Sep 5, 2019.
http://doi.org/10.26615/978-954-452-059-5_007

(Cantemir 1771).The first parallel Latin-Romanian Edition considering all available manuscripts was publisher recently (Costa 2015).

Historia incrementorum atque decrementorum Aulae Othomanicae, the history of the Ottoman Empire. In contrast with the previous work about Moldavia, here Cantemir indicates very carefully the sources of information. (Lemny 2010) supposes the existence of previous works, known in the western countries, behind this decision. This work was written also at the request of the Academy in Berlin. Cantemir follows the same principle: text in Latin, while the troponins and local denominations are written this time in Ottoman Turkish. Although there were already some previous works about the Ottoman Empire, the novelty of his approach is the quotation of Turkish sources. The reliability of these sources is untrusted sometimes by Cantemir himself. The manuscript reaches the western world after Cantemir's death, carried by his son to London. Here, a first translation in English is produced: The history of Raise and Decay of the Ottoman Empire. The translator reinterprets the texts, probably also being confused by the presence of Turkish information sources, which were perceived in that time as completely unreliable. The Latin original remains lost for centuries and is rediscovered only at the end of the XXth century in the USA. Thus, the German translation (Cantemir 1745) is based on the English one and inherits the same alterations, and presumably adds new ones. The Romanian translations use in contrast the Latin original. The last translation (Costa 2015) will be used in this proposal.
Until now there is no systematic study on the reliability of the text sources in Cantemir's works, nor the degree of alterations produced by the translations of the two works.
Given the fact that both works became standard reference for western authors until the middle of XIXth century, it is expected that their reception influenced also following historical material. There is no reprint / new edition of his works in German or English. There are however, several reprints of the Romanian versions. Recent Romanian translations of Decriptio Moldaviae are done after the original Latin manuscript.
A lot of works were dedicated to the personality of Dimitrie Cantemir and its perception in different parts of Europe. A study of the reliability and consistency of the historical facts as they are described in originals and their translations is practically impossible to be done only with traditional hermeneutic methods. One needs expertise in the same time in Latin, German, English, Romanian, Turkish, just to enumerate the main languages used in the two books, which additionally sum up to a volume of about 1000 pages. Both German editions are printed in "Fraktur" script, which is nowadays very difficult to be read. A recent digitalization done by the BBAW for the History of the Ottoman Empire, makes the text more accessible. The digital version is freely available in TEi-P5 format. However, the TEI-P5 concentrates only on a diplomatic transcription and a flat linguistic annotation (lemma and part of speech) and does not touch any aspects of vagueness or reliability of sources.
Cantemir's texts are a real challenge with respect to multilinguality: in Descriptio Moldaviae, the original version in Latin there are paragraphs classical Greek, Romanian and isolated in Turk ish. The Romanian Names are written with Latin characters, unusual for that period (Romanian was written until the middle of XIXth century with Cyrillic script). Thus, the transcriptions in Latin script is random because Cantemir uses sometimes the rules used at the Moldavian court, and some other times, the Polish system to translate Cyrillic (Nicolae 2004). The German translation imports original Romanian names for troponins, persons or professions, and tries to adapt it to the German Phonetics which increases once more the variants for one single name.
Given the:

- Geographic distribution of material (originals in libraries in USA and Russia; translations and copies across Europe; most part of the quoted sources in Turkey),
- The multilingual character of the materials to be investigated (Latin, German, Romanian, English, Turkish at least) and
- The volume of data which has to be processed in parallel

no study about the reliability and consistency of the original and the translations could be performed until now.

In the HerCoRe project we propose the mix hermeneutic and IT-methods in order to:

- compare the copy of the original (Latin) and the English and German translations

- identify translations mistakes or gaps (done by purpose or not);
- search after the quoted works and identification of related ottoman sources;
- analyse Cantemir‘s writing and discourse style;
- asses the importance of the work in the ottoman studies and compare them with other works contemporary with Cantemir or follow-up research about the ottomans;
- develop electronic resourced which may be of use for follow-up works about the ottoman empire and the history of Balkans.

For these purposes we combine methods from natural language processing, ontology reasoning and fuzzy modelling which we describe in the following sections

2 Annotation of Vagueness

For the particular corpus presented in section 1.1 we decided to represent vagueness and other types of uncertainty at least five levels (Vertan et al 2017)

1. the text uncertainty (uncertain readings, losses, translations, multilinguality, etc.),
2. the linguistic vagueness (metonymies, vague adjectives, comparatives, non-intersectives, hedges, homonyms,),
3. the author reliability (genres, time style, general recognition),
4. the factual uncertainty (range expressions, time expressions, geo relations), and
5. historical change (named entities, abbreviations, meaning changes).

In a first phase we collect for each of the processed languages (German, Romanian and Latin) explicit lexical vagueness markers like words or expressions such as:

- Vague quanitfiers, e.g.: some, most of, a few, about, etc.
- Modal adverbs, e.g.: probably, possibly, etc.
- Verbs e.g.: to believe, think, prefer, etc.
- Lexical quotation markers, e.g. introduced by quotation marks or verbs with explicit meaning (say, write, mention)
- Inexact measures and cardinals
- Complex quantifiers
- Non-intersective adjectives
- Implicit syntactic clues: mainly verb moods such as conditional-optative for Romanian, conjunctive mood or imperfect/pluperfect for Latin, all of them indicating a non-reality (doubt, hear-say, possibility, etc.)

To annotate vague expressions like the ones above, the first step is to (semi-automatically) identify them. Identifying the three distinct categories of expressions that induce vagueness (explicit-lexical, implicit-syntactic and pragmatic) requires different strategies.

To automatically identify (mark up in text) the explicit lexical-semantic clues, our strategy is the following: one manually create a list of words and expressions that are possible indicators of vagueness for the three languages (Latin, Romanian and German), from selected parts of texts. After the pre-processing step (chunking, lemmatizing, PoS tagging, NP-chunking), based on the previously created list, one automatically finds and marks all the (inflected forms of) explicit vagueness terms. Finally, one manually checks the marking for a short part of text for evaluation, followed by feedback and slight improvement.

The automatic identification of syntactic clues is a much more difficult/complex task. There is an inherent ambiguity in the text between vagueness and plain quotation (often intentionally created by the author) that is difficult to decide upon even for a human annotator, and thus impossible for the machine. A possible strategy to be investigated is: to use machine learning techniques (may be the power of deep learning) on a training set of positive examples obtained from explicit clues and negative examples of certain text.

Uncertainty is especially given by named entities like persons and places, especially when they differ in transliteration, spelling within the text or across similar historical sources. Thus the annotation of named entities is of central role.

However, the unclear person, time, place identification is even more difficult to automatize or at least assist by computer techniques, being more

of a matter of hermeneutical research for humanists and historians.

The annotated entities are modelled as individuals within a knowledge-base which we will describe in the following section.

3 Fuzzy ontological knowledge base

The knowledge base of the system is ensured by a manual developed ontology written in OWL 2 (Bobillo et al 2010)), modelling the administrative, religious, military and conceptual world of the Ottoman Empire and the Moldavia and Wallachia Principalities. The ontology is built according to the current generally accepted sate-of the art concerning the history of this territories. In this section we will detail on three main features of the ontology

The modelling of time: As for may events and biographical data only uncertain dates are available we decided to represent a year not as a string , reflected then in other concepts as a Datatype Property but as an object (thus a concept into the ontology). Each concrete year is thus an instantiation of this concept. For a "*Year*"-concept we specify the

- *exactValue* a string
- *aroundValue, beforeValue, afterValue,* defined as fuzzyDatatypes
- *shortBefore, shortAfter* values defined as a combination between a modifier and a fuzzy datatype

The modelling of geographical regions: there are a number of geographical places for which the concrete placement is still not clear. For this we define a concept *GeographicalVagueZone* having as properties fuzzy datatype *neighborhoodOf*

The modelling of historical political entities: We distinguish between fixed concepts and relations (like geographical elements: river, mountain, island) and notions for which several "contexts can be defined. E.g. a geographical notion like "Danube" is within one historical context a border of the administrative notion "Ottoman empire", and in another one the border to the so called administrative notion "Roman empire". The historical contexts are specified by further objects containing fuzzy data properties (e.g. time, placement).

3.1 Conclusions and further work

Annotation and interpretation of vagueness is a central issue in digital processing of historical texts. However, this issue was completely neglected until now, and has as consequence often distorted interpretation of digitized historical texts. In this article we presented the current state of the art on vagueness annotation and introduce the first approached for considering vague expressions as part of the annotation process. We describe also the introduction of fuzzy properties into the ontological knowledge base as main backbone for interpretation of vague and uncertain facts Further work concerns the completion of the ontology, the linkage between the ontology and the corpus and the adaptation of a fuzzy reasoner (as in Bobillo et al 2013) dealing with the different types of annotations

Acknowledgements

Research described in this article is supported by HerCoRe project, funded by Volkswagen Foundation (Project no. 91970)

References

Bobillo, Fernando and Straccia, Umberto, 2010, "*Fuzzy Ontology Representation using OWL 2*", http://arxiv.org/pdf/1009.3391.pdf (last retrieved 28.08.2019)

Fernando Bobillo and Delgado, Miguel and Gomez-Romero, Juan, 2013, *Reasoning in Fuzzy OWL 2 with DeLorean, in Uncertainity Reasoning for the Semantic Web II,* in Bobillo, F.,Costa, P.C.G.,dAmato , C.,Fanizzi, N.,Laskey, K.B.,Laskey, K.J.,Lukasiewicz, Th.,Nickles, M.,Pool, M. (Eds.), Lecture Notes in Artificial Intelligence, Springer Verlag

Ioana Costa 2015, *Dimitrie Cantemir, Istoria măririi şi decăderii Curţii othmane,* 2 volume, editarea textului latinesc şi aparatul critic Octavian Gordon, Florentina Nicolae, Monica Vasileanu, traducere din limba latină Ioana Costa, cuvânt înainte Eugen Simion, studiu introductiv Ştefan Lemny, Bucureşti, Academia Română-Fundaţia Naţională pentru Ştiinţă şi Artă, 2015. ISBN 978-606-555-135-0 (978-606-555-136-7, 978-606-555-137-4)

Wilhelm Dilthey,1922, *Einleitung in die Geisteswissenschaft.* .

Cantemir, Dimitrie, 1771, *Beschreibung der Moldau, Faksimiledruck der Originalausgabe von 1771*, Frankfurt und Leipzig

Cantemir, Dimitrie, 1745 *Geschichte des osmanischen Reichs nach seinem Anwachse und Abnehmen*, 1745, Herold, Hamburg

Stefan Lemny, Stefan, 2010, Cantemirestii -A*ventura europeana a unei familii princiare din secolul al XVIII-lea*, Polirom Publishing House.

Nicolae, Florentina, 2004, *Toponime si Hidronime in literature Cantemiriana*, Annals of Philology XV, pag., 143-152

Cristina Vertan and Anca Dinu and Walther v. Hahn, 2017, On the annotation of vague expressions: a case study on Romanian historical texts, Proceedings of the RANLP 2017 Workshop on Language Technology for Digital Humanites in Central and South Eastern Europe