Checking reliability of quotations in historical texts-A digital humanities approach-

Cristina Vertan[1] and Alptug Güney[1] and Walther v. Hahn[1]

[1] University of Hamburg, Vogt-KöllnStrasse 30, 22527 Hamburg, Germany
cristina.vertan@uni-hamburg.de , alptug.gueney@uni-hamburg.de,vhahn@informatik.uni-hamburg.de,

Abstract. The analysis of reliability of quotations in historical texts is particular difficult, first because the quotation style differs from the modern one and is usually done in a more hidden way, secondly because quoted works have themselves reliability issues. In this paper we will describe a mixed hermeneutic-computational approach, trying to support the humanist researcher with plausible evidence about the reported historical facts. Starting from an initial hermeneutic investigation we built a fuzzy ontology and combine the queries on the ontology with analysis of linguistic vagueness and uncertainty. We describe the hermeneutic and computational methods and present add-on value for the humanist researcher.

Keywords: fuzzy ontology, vagueness, uncertainty, reliability of historical sources

1 Introduction

The quotation style in historical texts differs a lot from the one used nowadays both natural sciences and humanities works. Until 19th century texts relating historical facts were quoting indirectly, using expressions like „it is said ", „in the chronicle X it is told "etc. Usually neither there is no bibliographical list with the consulted sources nor it is clear if the author really saw the sources or the quotations are more a result of oral transmission.

Thus it is usually very difficult to check the reliability of presented facts and their historical support. One of the first historians, who started to change this system is the Moldavian prince Dimitrie Cantemir, author of the first exhaustive history of the ottoman empire. Written most probably in Latin, at the beginning of the 18th century, at the demand of the Berlin Royal Academy of Sciences, the work gets translated few years afterwards into English. The English translation respects only partially the text and introduces itself many errors, also concerning the quotations. This translation it is used however as base for the translations into other languages: German [3], French, Romanian, Turkish, and became the most used book about the ottoman empire until Middle of 19th century. One can consider that a great part of the image the ottoman Empire had in the western European countries at the begin of the 19th century is based on this book. As consequence it is one of the most prominent works for any scholar in turcology.

Already researchers in the 1920, e.g. Babinger [1], tried to analyse the quotations used by Dimitrie Cantemir. The lack of access to materials in the Turkish libraries, the difficult access to documents spread all over the world lead at that time to the conclusion that most part of Cantemir assertions do not have a solid scientific, historiographical base. Babinger and other historians claimed that many of Cantemir's sources are invented and persons he claimed to have discussed with never exist.

In light of new gained access to (on-line) materials as well as a real progress in ottoman historiography since the 1920ies, as well as the discovery of original Latin versions (originals or copies of originals) of Cantemir's books, his work and style of quotation has to be re-analyzed.

In the project HerCoRe – „Hermeneutic and Computer-based analysis of Reliability, consistency and vagueness in historical texts"(https://www.inf.uni-hamburg.de/inst/dmp/hercore/projects.html) an interdisciplinary team of scholars in ottoman studies, linguists and computer scientist investigate how accurate Dimitrie Cantemir quoted the ottoman sources, if he could have met mentioned oral sources and to which extension the translation process has changed the original text.

A central brick in this investigation is the creation and deployment of a large knowledge base (a fuzzy ontology) about the ottoman Empire and the neighboring countries. In this work we will discuss:

- The role the ontology plays for the inference processes and thus for querying the corpus
- The influence the work at the ontology has on the hermeneutic analysis
- The experience of the mixed team in the process of developing the ontology

2 Hermeneutic investigation

The challenge of this investigation was, on one hand to provide enough information for the computer modelling, on the other hand not to replace it. Thus we decided to use mainly just the German translation.

It consists of two volumes with four books and a set of so called annotations done by the author; the first volume includes three books (pp. 1-408) which are dealing with the growth of the Ottoman Empire. The second volume includes the fourth book (pp. 409-770) which is handling the decay of the Empire. In a first iteration we explored introduction of Cantemir and then one chapter from each book. The decision was taken according to information given by the researcher in ottoman studies and relies on different degrees to which ottoman history is overall documented. After manual extraction and analysis of quotations, and corresponding reported historical texts, we decided to extend the investigation to additional chapters for the parts where Cantemir's narrative deviates from or contradicts strongly the quoted sources. This process is in-line with the hermeneutic cycle.

.

This manual investigation we concluded that Cantemir gives references to multiple Turkish and European historians and a few times, he mentions only the title of a book without mentioning the author. Consequently, a list of these names and titles that he has used as source was produced and the works of these historians for their content and the accuracy of Cantemir’s references were checked. In this regard, we tried also to find out the methodology of Cantemir, after the reconstruction of these sources.

According to that Cantemir used the Turkish and Greek manuscripts and also European works on Ottoman history. He deployed generally wide-ranged and well-reputed Ottoman sources which are regarded also by Ottoman scholars as standard-work of the historiography.

The German translator of the book used also some additional sources and made some remarks regarding the information provided by Cantemir.

Following sources were selected as relevant to be compared with Cantemir’s assertions:

- Ottoman Sources:
    - Âşıkpaşazâde (1400-1484), Tevârîh-i Âl-i Osman [Mid. C13th-1472]
    - Neşrî (?-1520?), Kitab-ı Cihan-Nüma [Mid. C13th-1481]
    - Hoca Sadettin (1536/7-1599), Tâcü't-Tevârih [Mid. C13th-1520]
    - Peçevî (1574-1649?), Tarih-i Peçevi [1520-1648]
    - Hezarfen Hüseyin (?-1691), Tenkih üt-Tevarih [Mid. C13th-1672/73]
    - Byzantine/Greek Sources:
    - Nikephoros Gregoras (1295?-1359/61), Byzantine History [1204-1359]
    - Laonikus Chalkondylas (1423?-1490), Proofs of Histories [1298-1463]
    - Georgius Phranza (1400?-1477?), Chronicle [1258-1476]
    -
- Latin Sources:
    - Philipp Lonicerus (1532-1599), Chronikorum Turcicorum [Mid. C13th-1529]
    - Johannes Gaudier (Hans Caudir von Spiegel) (XVI. Jh.-1579)/
    - Johannes Leunclavius, Annales Svltanorvm Othmanidarvm, A Tvrcis Sva Lingva Scripti (Translation of: Muhyiddin Cemali (?-1550),
    - Tevarih-i Al-i Osman [Mid. C13th-1549], German Translation: Johannes Leunclavius (1541?-1594), Neue Chronika türkischer Nation (1590)

Additionally, research within the Orthodox Patriarchate in Istanbul lead to a number of other 55 documents which may be consulted by Cantemir (i.e. were available at the time he lived in Istanbul).

Further hermeneutic investigation, e.g. research in law documents, official edicts, etc. let us construct a network of about 50 people which built Cantemir’s network, and who could have been oral sources for his book.

One of the most important observations was that markers for linguistic vagueness cannot be used exclusively for attestation of reliability of historical facts. There are paragraphs were Cantemir uses expressions with a strong semantics of “being sure” “true” and the reported facts are invalidated by other works. On the other hand, there are paragraphs were he remains vague, uses expressions with a semantics like “being unsure”, “improbable” but in contradiction the reported events are attested but several other written testimonies. This is an important finding showing that the linguistic annotation and analysis has to be accompanied by an annotation and analysis of the knowledge expressed in the text

3 Computer-bases approach

Computational methods are employed with a twofold goal:

1. As proof for the hypotheses from the hermeneutic study
2. In order to analyse historical claims done by Cantemir and compare them with historical evidence.

The backbone of the computer-based approach consists of:

- A rich knowledge base aiming at model in the ottoman world, in light of current state-of-research
- The mark-up of linguistic vagueness indicators

3.1 The construction of the knowledge base

The HerCoRe knowledge base is a Fuzzy OWL Ontology[1], trying to model he ottoman empire world in its administrative, social, geographical and religious facets.

Following aspects need a particular attention:

The modelling of geographical respectively political entities. Political entities (e.g. countries) tend often to share name with some geographical entities. Political entities keep name but change often borders. Thus we considered as fix, unambiguous individuals the geographical elements which are visible nowadays. Historically attested geographical zones which do not exist are modelled as fuzzy concepts.

Political entities are defined as a sum of several historical contexts. We introduce additionally the concept of „historical zone“ in order to model concepts as „Europe“ which from the point of view of the Ottoman Empire e.g. began at the border with Hungary, or „Balkan“ which for the Ottoman Empire was represented by the Wallachia and Moldavia principalities.

The ontology is still under development, and contains for the moment:350 Classes, 130 Object properties, and 2000 Individuals

For dealing with multilinguality we attach to each individual the name used in German translation in the official ottoman documents, as well as in the Romanian sources.

Time Intervals and geographical positions include fuzzy concepts which allow us to model uncertain dates and coordinate

3.2 Annotation of linguistic vagueness

A preliminary linguistic analysis of the texts showed that the author uses a large number of vague expressions and doubts often about the correctness of events. Especially he seems aware that legends can be trust only to limited extent. Given the extensive

[1] https://www.w3.org/TR/owl2-syntax/

use of vague expressions we decided to annotate them in text and embed them in the analysis.
Concrete we recorded so called language dependent vagueness lists following Pinkal's [5] classification of vagueness.

1. comparatives, inexact adjectives e.g. *"mehr/more" "größer/bigger", "älter/older"*
2. non-intersectives e.g. „vermeintlich/supposed", „so-genannt/so-called"
3. Hedges e.g. „ziemlich/quite", „einigermaßen/approximately „etwa/about"
4. inexact measures *„4 Tagereisen/4 days trip", 10 Fuß /10 feet"*
5. modals (attitudes) e.g. „vielleicht/maybe", „hoffentlich/hopefully"; subjonctives verbs
6. lexical quotation markers :"es wurde gesagt /it is said"
7. vague quantifiers e.g. „viele", „meistens /mostly"
8. complex quantifiers e.g. [2]*"etwa die Hälfte von den 20-30 tausend Soldaten / about a half from the 20-30 thounsand soldiers"*
9. numbers
10. range expressions e.g. *"Anfang des 18. Jhds./begin of 18th century"*
11. unclear place *„Syrfia", „Moramor"*
12. unclear person e.g. *„der ehemaligen Herzog / the former duke"*
13. unclear time e.g. *„in alten Zeiten /in old times"*
14. Domain specific *e.g. „Wesir/vizier" vs. „Wesire/viziers"*

The list of vagueness indicators were created manually and then enriched semi-automatic with elements of synsets extracted from the language specific Wordnet. Semi-automatic refers to the following procedure: for each term in the vagueness list, its synset is extracted automatic from the Wordnet. A human evaluates then, if the synset elements were part of the vocabulary of the 18th century, and in positive case, if the semantics was the same.
The annotation of these vagueness parkers is preceded by an automatic morphological annotation, assigning to each token a part-of-speech and morphological values, including e.g comparative degrees for adjectives. In order to ensure a uniform annotation across languages we decided for the CoNLL-U (Universal Dependencies) – Format.

3.3 System architecture

The system architecture [4] is presented in Figure 1
Following user scenario is the basis for the system architecture:

- The user formulates a SPRAQL[3] Query
- The Query is mapped against the knowledge base of the system
- The
- result is a list of possible interpretations (RDFTriples Subject-Predicate-Object associated with the Source from where they origi-

[2] https://universaldependencies.org/format.html
[3] https://www.w3.org/TR/sparql11-query/

nate). Each triplestore has associated a so called vagueness –score composed of following bricks: lexical vagueness, factual uncertainty, named entity uncertainty and hedges

It is important to mention that the system is not choosing one solution among these answers but leaves this for the hermeneutic interpretation

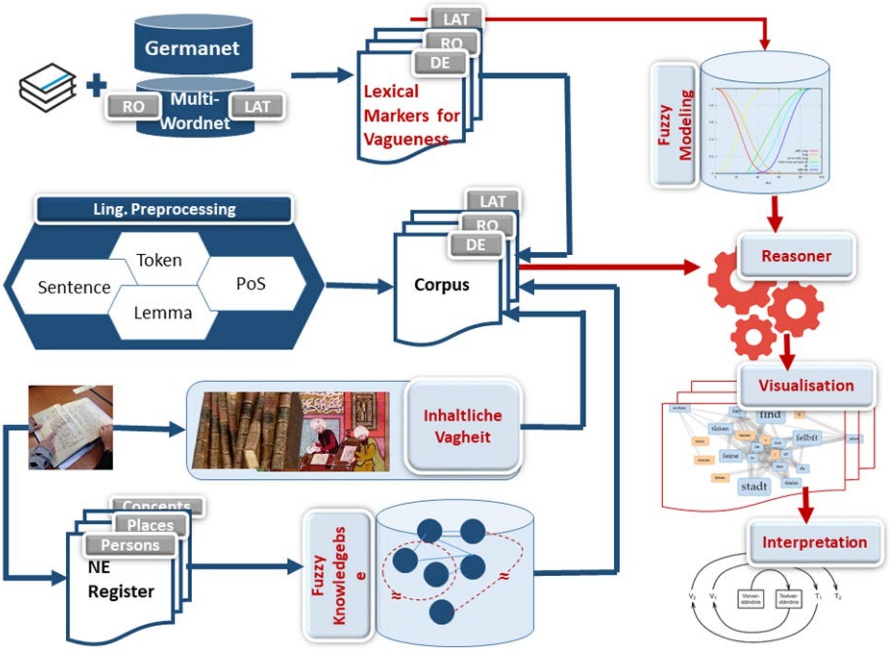

Fig. 1. HerCoRe System Architecture

4 Challenges of the interdisciplinary work and first results

The ontology is built with the help of Protégé-Tool. The manipulation of the Protégé tool is not trivial for a humanities scholar and needs some weeks of training. On the other hand, a formalization done exclusively by the computer scientist will lead to wrong defined concepts. In a first phase the computer scientist and the ottoman research scholar worked together: the research scholar explained a certain domain (e.g. military organization) and the computer scientist translated this into OWL statements.

In a second phase the concept and the object relations and data properties were created by the computer scientist whilst the humanities scholar created the corresponding individuals.

An interesting aspect was that there was a smooth transition to the next step in which the research scholar started to create himself concepts and object relations.

A more complicated issue is represented by the fuzzy concepts and relations. This modelling implies dep knowledge of mathematical foundation. Although it exists a plug-in for Protégé supporting fuzzy representations [2], the GUI of this plug-in is not working for the newest Protégé version. This means that the user must encode directly in OWL these concepts, object relations and data types. In this case the research scholar is writing in form of a comment what has to be modelled as „fuzzy" and the computer scientist is transiting in OWL.

Although there is a Web version of Protégé, its manipulation is much more difficult than the desktop version. Additionally, some features and plugins are not supported. We choose the desktop version and work iteratively: after a development phase, the ontology is checked by the computer scientist, cleaned from redundancies and given back to the humanist scholar for development.

In this phase we could already detect, with the help of the system, different parts of Cantemir's work which are extremely accurate and some parts were the historical accuracy is missing. We could also observe that there is not a 100% correlation between the usage of vagueness expressions and the accuracy of related facts: some of them are announced as less probable, but are attested also by all other chronicles. On the other hand, sometimes the author seems to be very sure and the reported facts are erroneous.

Extremely interesting is the observation done by the research scholar who declares that already the work at the ontology itself brings complete new insights on the texts. The scholar is more attentive to each detail; additionally, the introduction e.g. of a new Individual in the ontology together with all object relations leady to a broader research, and often conducts to detection of new facts, insights, errors.

For example, for recording geographical places, the researcher had to look after details in ottoman archives. This had at consequence, that one obtains not only the required information but additionally gains knowledge about the relation ottomans had with their geographical neighborhood. Also the importance of small geographical details, features, to which it was not paid attention until now become evident.

In order to model historical figures (persons) in the ontology one has to specify a number of family relationships (filiation, marriages, etc.), public and administrative functions, relations with art etc. The researcher is forced to conduct a more detailed research, which otherwise would have been neglected. The same holds for the modeling of (vague) concepts and the relations between them. An important feature of the ontology is the recording of sources responsible for a certain concept or individual description. This forces the researcher also to a more careful processing of the information and selection of features and values for each concept.

Furthermore, we expect that the ontology will be used for various research scenarios related to the history of the ottoman empire.

5 Acknowledgments

This work was performed with the project HerCore (Hermeneutic and Computer based Analysis of Reliability, Consistency and Vagueness in historical texts) funded by the Volkswagen Foundation within the Framework "Mixed Methods in The Humanities" 2017-2021.

References

1. Babinger, Franz, Die Geschichtsschreiber der Osmanen und ihre Werke. Leipzig, (1927)
2. Bobillo, Fernando and Delgado, Miguel and Gomez-Romero, Juan, "Reasoning in Fuzzy OWL 2 with DeLorean, in Uncertainity Reasoning for the Semantic Web II, Bobillo, F.,Costa, P.C.G.,d'Amato, C.,Fanizzi, N.,Laskey, K.B.,Laskey, K.J.,Lukasiewicz, Th.,Nickles, M.,Pool, M. (Eds.), Lecture Notes in Artificial Intelligence, Springer Verlag, (2013)
3. Cantemir, Dimitrie, Geschichte des osmanischen Reichs nach seinem Anwachse und Abnehmen, Herold, Hamburg, (1745)
4. Güney, A. and Vertan, C. and von Hahn, W. "Combining hermeneutic and computer based methods for investigating reliability of historical texts", Proceedings of the "Twin Talks" workshop collocated with DHN 2019, Steven Krauver and Darja Fiser (Eds), University of Copenhagen 2019, https://cst.dk/DHN2019Pro/TwinTalksWorkshopProceedings.pdf, pp. 25-38, (2019)
5. Pinkal, Manfred, 1985: Logik und Lexikon: Die Semantik des Unbestimmten, (1985)