Combining hermeneutic and computer based methods for investigating reliability of historical texts

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

[1] University of Hamburg, Hamburg Germany
{cristina.vertan,alptug.gueney}@uni-hamburg.de,
vhahn@informatik.uni-hamburg.de

Abstract. Within the framework of the project HerCoRe[1] we are analyzing two historical works from 18th century "History of Rise and Decay of the Otoman Empire" and "Description of Moldavia" (both written by Dimitrie Cantemir) and investigate them with regard to the historiography of its time. We evaluate the usage of sources by the author and also the reliability of his references. We also seek to shed more light on the motivation behind the writing-process of these works by taking into account the political and cultural dynamics of the time and the position of Cantemir within the Ottoman elite. To determine missing or incorrectly translated parts of the work, the German and English translations are also compared with a copy of the Latin manuscripts. This comparative approach serves also to discuss the causes of the (un)conscious mistakes and omissions in the translations. We are performing this study by means of hermeneutic and IT approaches.

Keywords: historical documents, uncertainty and vagueness annotation, hermeneutics.

1. Rationale of the research

Dimitrie Cantemir (1673-1623) was prince of Moldavia (a historical area including regions from current eastern Romania, Republic of Moldavia and some parts from Ukraine), He was a 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, met important persons around the sultan and learned a lot about the 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, on

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

their request, he produced the two books which are the target of this proposal:

- *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, as well as the political and administration system. Local denominations and toponyms, as well as names are written in Romanian with Latin script as his intention was to demonstrate the Latin origin of his folk. The transcriptions are not standardized and one retrieves for the same toponyms, several name variations. Quotations as known today were very rare, there is no bibliography. According to [3], 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 into German, and only later - at the middle of the XIXth century - into 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.
- *Historia incrementorum atque decrementorum Aulae Othomanicae*, the history of the Ottoman Empire. In contrast to the previous work about Moldavia, here Cantemir indicates very carefully the sources of information. [3] supposes the existence of previous works, known in the western countries, behind this decision. This work was written also on the request of the Academy in Berlin. Cantemir follows the same principle: text in Latin, while the toponyms 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 original manuscript (or a copy of it) reaches the western world after Cantemir's death, carried by his son to London. Here, a first translation into 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 at that time were perceived 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 is based on the English one and inherits the same alterations, and presumably adds new ones. The Romanian translations, in contrast, use the Latin versions. The last translation [2] is being used in this research.

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 the latin copies) and their translations is practically impossible ~~to be done~~ only with traditional hermeneutic methods. One needs expertise at the same time in Latin, German, English, Romanian, Turkish, to enumerate just the main languages used in the two books, which additionally sum up to a quantity of about 1000 pages. Both German editions are printed in "Fracture" ("black letter") script, which nowadays is very difficult to be read.

Already in the 1920s it was demonstrated (by using only a selections of texts), that the translations are not respecting the original all the time. E.g. information sources indicated by Cantemir were omitted, because they seemed too unreliable to the translator.

In the XXth century researchers claimed that some of the sources, persons and facts quoted by Cantemir were not existing at all (e.g. [1]).

But given the:

- geographic distribution of material (originals in libraries in USA and Russia; translations and copies all 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 Quantity of data which has to be processed in parallel,

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

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

- compare the Latin copies and the English and German translations,

- identify translation mistakes or gaps (made by purpose or not),
- search after the quoted works and identifiy related Ottoman sources,
- analyse Cantemir‘s writing and discourse style,
- assess the importance of the work in the Ottoman studies and compare them with other works contemporaneous to Cantemir or follow-up research about the Ottomans,
- develop electronic resources which may be of use for follow-up work about the Ottoman empire and the history of Balkans.

2. Hermeneutic investigation

The hermeneutic investigation concentrates on the identification of sources quoted directly or indirectly by Dimitrie Cantemir, as well as the mentioned places, persons, events and dates.

The two works are very different with respect to the quotation style. While in the “Description of Moldavia” the quotation sources are almost missing, in the “History of rise and decay of Ottoman Empire” the author refers explicitely to different sources. However, there is no quotation style like in modern scientific works. Most references are real quotations or the author indicates the source of quotation through syntactic phrases followed by a reformulation of the semantic substance of a text section. Especially these cases are subject to the hermeneutic investigation.

By now we identified the main works quoted by Cantemir. These works are available only in paper form and are written in Ottoman Turkish (with Arabic alphabet) thus only a manual comparison can be performed.

This systematic comparison led to a very unexpected result: we observed that linguistic expressions of certitude (e.g. *“for sure”*, *“without any doubt”*) are not an unambiguous indicator of the reliability of the quotation. E.g.: Cantemir is sure that all investigated sources mention 4 sons of Sultan Bayazid. However, all reliable sources of the time mention that the sultan had five sons.

We do not know why these inconsistencies occur. One possible motivation is the context in which he wrote the two books: in exile in St. Petersburg, probably with few notes at hand, that he made in Istanbul. Whilst we cannot find the reason of the inconsistency, the hermeneutic analysis showed that a pure automatic annotation (searching for quotation marks) will not help in the case mentioned above, as the semantics of the quotation mark does not match the degree of reliability of the quoted information. This is something which cannot even be inferred by automatic methods; at least

not at this stage, where documents in Ottoman Turkish are rarely digitised and no linguistic tools are available for this historical language variant.

A second part of the hermeneutic investigation concerns the collection of persons, places, and domain specific concepts which are mentioned by the author. An automatic identification is practically impossible, as names are not standardised (e.g. for the city of Iasi in Moldavia we identified at least 12 writing variants).

The third part of the hermeneutic investigation is concerned with the identification of missing paragraphs in the German and English translation. First result: all paragraphs written in the Latin original with Arab characters were systematically omitted. This leads to misunderstandings in the two translations.

3. Computer-based approach

Digital methods can facilitate analysis on the reliability of translations but also of the historical facts claimed by the author [8]. In order to be effective, these methods must consider an intrinsic feature of all natural languages: the ability of producing and understanding vague utterances. The project HerCoRe aims at modelling and annotating five levels of vague assertions

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, contemporary knowledge),
4. the factual uncertainty (range expressions, time expressions, geo relations), and
5. historical change (named entities, abbreviations, meaning changes)

We develop an annotation formalism which allows for:

- the mark-up of different types of vagueness and its source; the implementation of a set of inference rules for the combination of such vague features to calculate an overall result of their reliability;
- the definition of a similarity measurement of the inferred results obtained for the same queries on different translations. The system architecture is presented in figure 1. It relies on annotation on 4 levels (linguistics, lexical markers for vagueness/uncertainty, ontological and factual/quotation markers).

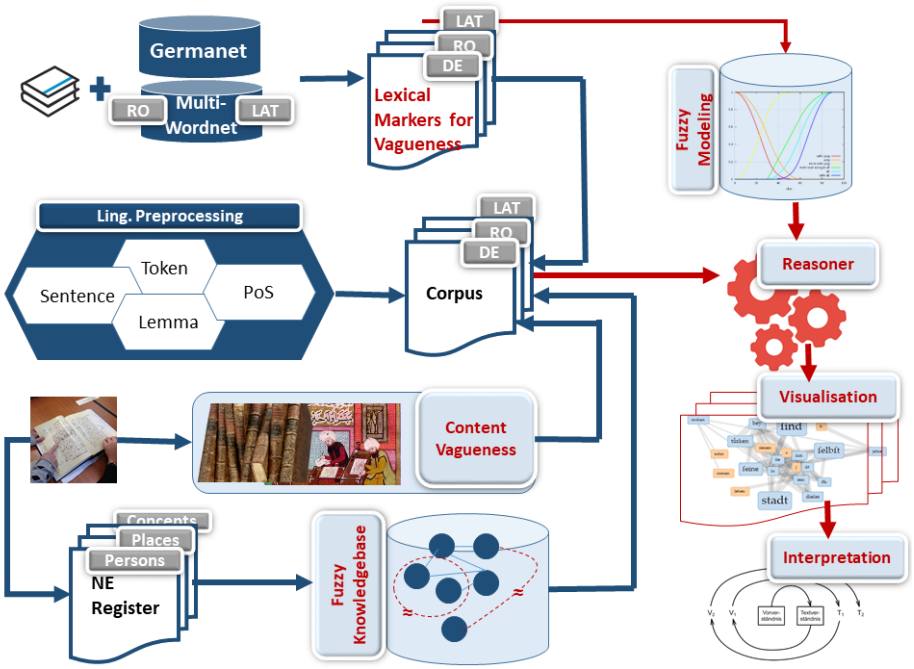

Figure 1. HerCoRe System Architecture

For the detection of linguistic vagueness we follow a multilingual approach. We collected the above listed indicators in the three languages involved in the project (Latin, German and Romanian). Based on [5] [6] we distinguish between:

- Vague quantifiers, e.g.: some, most of, a few, about, etc.
- Modal adverbs, e.g.: probably, possibly, etc.
- Verbs e.g.: to believe, think, prefer, assume 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 past perfect/pluperfect for Latin, all of them indicating a “counterfactive” or non-reality (doubt, hear-say, possibility, etc.)

The initial collections of linguistic indicators are enriched through synsets in the corresponding Wordnets.

The knowledge base backbone is ensured by a fuzzy ontology modelled in OWL2. 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 fuzzy data properties (e.g. time, placement).

4. A case study

In "The History of the Growth and Decay of the Ottoman Empire (1734)" Cantemir tells the story of a battle between the *Moldavian Prince Ștefan* and the *Ottoman Sultan Bayezid I.* Cantemir does not give the exact date of the battle, but one could think that this can be inferred from other details of the text. *Ș*tefan attacked the Ottoman camp in *Rasboeni*. According to the account of Cantemir, in this first confrontation, Ștefan lost the battle and retreated to his castle in *Neamț*. Then, after the inspiring speech of his mother in *Neamț*, he drew his soldiers together and stroke back the Ottoman army twice in succession. After the last defeat in Vaslui, Sultan Bayezid fled back to Edirne [7]. In the same paragraphs, Cantemir gives in a footnote some information - among many other detailed and important details - about the Moldavian Prince Ștefan, who fought two times against *Bayezid I.*:

"*He overthrew the renown'd Matthias King of Hungary, and wrested from him Transilvanian Alps […] His son Bogdan made Moldavia tributary to the Turks.*"

This account of Cantemirs includes several problems, which can mislead the reader and even a historian who does not have detailed knowledge about the Ottoman and Romanian (Wallachian and Moldavian) history.

Known and proven historical facts:

- There were two sultans with the name Bazeyid in the history of Ottoman Empire Bayezid I and Bayezid II but only the first one had the additional name "Yildirim" i.e. the Thunderbold. Cantemir mentions exactly this appellative and not the numbering (I or II) so we can exclude any typo or damaged spot in the manuscript. We checked this information in all translations and the Latin facsimile.
- The reign time of Bayezid I is known for sure (according to diplomatic text sources): 1389 – 1402.
- The frontiers of the empire at that time leaned already towards the Danube River and the Ottoman Empire was yet neighbor to Wallachia and Moldavia, thus a military confrontation with both principalities is historically possible.

- During this time Wallachia was ruled by several princes: Mircea I (1386-1395), Vlad I (1394-1397) and then again by Mircea I (1397-1418).
- The Ottoman chronicles report on a battle in 1391 between the Wallachian Prince *Mircea* and *Bayezid I* in a place called Arkaş (in Romanian *Rovine*). According to the Ottoman historians, Bayezid won the battle and Mircea recognized the Ottoman sovereignity [4].
- At the time of Bayezid Wallachia was ruled by Mircea I (1386-1395), Vlad I (1394-1397) and then again by Mircea I (1397-1418). In Moldavia there was just one Ruler called Ștefan (Ștefan I 1394-1399)
- There was a Moldavian ruler Ștefan III (1457-1512) who defeated the Ottomans in 1475 after a loss in Rasboieni, and who defeated also the Hungarian King Matthias (known as Matthias Corvinus [1443-1490]). Moreover, this ruler had a son called Bogdan who made Moldavia tributary to the Ottomans. These facts are confirmed by Ottoman chronicles [4].

At a closer look there is a strong mismatch between Cantemir and all other established chronicles, but a historian would not know here how to interpret the text:

- Is it referring to a battle against Moldavia or Wallachia?
- Which Ruler opposed Bayezid Yildirim?
- Where took the battle place?

An historian using only traditional methods (source inspection, reflection, re-evaluation of text) will face here a bunch of unsure and contradictory information, very difficult to resolve. An historian with less background knowledge about the Romanian history will be tempted to interpret wrongly the text section, either choosing the wrong rulers or the wrong place.

The HerCoRe System aims at helping historians in their interpretation, and suggests different reading paths. From the ontology and additional annotations the following inferences are possible:

- `Class Ruler and wasRulerof value 'OttomanEmpire' and has Name Bayezid and has Additional Name Yildirim -> Bayezid Yildirim 1389-1402`
- `Class Ruler and hasName 'Bayezid' and 'has additional-Name Yildirim' and hasBattlesIn some (Class Principality and belogsTo some (Historical Region and (liesIn value NorthDanube) -> Moldavia and Wallachia`
- `Class Ruler and was RullerOf exactly Moldavia and had BattleWith value 'Bayezid Yildirim' ->` this will show all rulers which fulfil the criteria.

Continuing this queries to the ontology, or invoking a complex inference chain the system will propose following solutions:

- The paragraph is about a battle in Wallachia in a place called Rovine and against a ruler *Mircea I*. -> contradicts Cantemir text: (it is not a Moldavian king but a Wallachian). The user may infer also that the place called '*Rasboe*' by Cantemir might be the place known as '*Rovine*'
- The paragraph is about a battle in Moldavia against the Ruler Ștefan I, and Cantemir mistakes the information about Ștefan I for Ștefan III
- The paragraph is about the battle in *Răsboieni* against the Moldavian prince *Ștefan III* and the mismatch here is about the Sultan (*Bayezid I* or *Bayezid II*). Thus, the battle place here called by Cantemir '*Rasboe*" would be '*Răsboieni*'.

All this information will have attached a score indicating a degree of truth. The latter one e.g. will have a lower score when introducing an additional scoring parameter: the metadata. The metadata will say, that the mentioned paragraph is within a chapter about the sultan Bayezid I, so it is less probable that Cantemir mistakes the name of the Sultan.

HerCoRe System does not aim at proposing a final solution. This decision is left entirely to the user/researcher, who is the hermeneutic subject and also can store the inference paths presented above as motivation for his choice.

5. Conclusions and further work

In this article we intend to show how hermeneutic and IT methods can be combined in order to investigate the reliability of historical texts (original and their translations). We show that a deep analysis can be performed only by the combination of the two approaches. Current research focus on the semi-automatic annotation and development of the ontology. Further work concerns the implementation of the reasoner and the visualisation of results.

6. Remarks on cooperation

In our project we need the cooperation between computer scientists, linguists (Latin, German and Romanian) as well as researchers in turcology. We cooperate also very close with the editor and translator into Romanian of the two works. The cooperation already revealed to date interesting aspects:

- The simple availability of raw texts in digital form does not help. E.g. the German translation of the History of Ottoman Empire is available from

the German Text Archive. However, the text was digitized for visualisation purposes. The usage of the underlying text versions leads to an unsorted mixture of paragraphs written by the Cantemir, his side notes and the comments of the translator. These parts, marked correspondingly in the TEI version are melted in the .TXT version. We had to invest additional work on separating these distinct parts.

- Mark-up in the editions have to be considered, as they can enrich the text. However, first one has to know the semantics of the mark-up.
- The ontological formalisation of the notions mentioned in the text was a great help for the humanist researchers leading to a better reflexion of the used notions.
- Many of the computational linguistics approaches had to be revised given the particularities of the historical text.

References

1. Babinger, Franz, 1927, Die Geschichtsschreiber der Osmanen und ihre Werke. Leipzig
2. 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)
3. Lemny, Stefan, 2010, Cantemirestii -Aventura europeana a unei familii princiare din secolul al XVIII-lea, Polirom Publishing House.
4. Parmaksızoğlu, İsmet (Ed.), Hoca Sadeddin, Tâc'üt-tevârih, Bd. III, Ankara, 1979, 153-158; Abdülkadir Özcan, „Boğdan", TDV İslam Ansiklopedisi, Bd. VI, 269.
5. Pinkal, Manfred, Semantische Vagheit: Phänomene und Theorien. In: Linguistische Berichte 70. 1980. 1-26. und 72. 1981. 1-26.
6. Pinkal, Manfred, 1985 Logik und Lexikon: Die Semantik des Unbestimmten.
7. Unat, Faik Reşit (Ed.), Mehmed Neşrî, Kitâb-ı Cihan-Nümâ, Bd. I, Ankara, 1949, 327; Parmaksızoğlu, İsmet (Ed.), Hoca Sadeddin, Tâc'üt-tevârih, Bd. I, Ankara, 1979, 200-201.
8. Vertan, Cristina and v. Hahn, Walther, 2014, Discovering and Explaining Knowledge in Historical Documents, In: Kristin Bjnadottir, Stewen Krauwer, Cristina Vertan and Martin Wyne (Eds.), Proceedings of the Workshop on "Language Technology for Historical Languages and Newspaper Archives" associated with LREC 2014, Reykjavik Mai 2014, p. 76-80.