Encoding, Processing and Interpreting Vagueness and Uncertainty in Historical Texts – A Pilot Study Based on Multilingual 18th Century Texts of Dimitrie Cantemir (HerCoRe)

Cristina Vertan

Abstract *Qualitative and Quantitative Analysis of historical documents by means of computational methods can lead to a better understanding of the particular texts but also of the époque to which they relate. This can be however achieved only if 1.) the digital methods encode, embed and allow visualisation of the various manifestations of vagueness and ambiguity present at all levels in text (metadata, edition, discourse, lexical etc.), 2.) the knowledge-base is built in close cooperation with specialists in the respective time period and 3.) the final interpretation is left to the user. In this contribution we will present an innovative approach for dealing with the vagueness of natural language and show how Mixed Methods can help in detecting reliability and consistency of historical documents from the 18th century. We use a multilingual corpus consisting mainly of texts in Latin, German and Romanian and apply methods from fuzzy logic, graph theory, knowledge acquisition and natural language processing.*

Introduction

Over the last decades, digital processing of historical texts (in fields such as Digital History, Historical Linguistics or Digital Culture Heritage) has focused primarily on digitization and adaptation of computer linguistic tools for (nearly) extinct languages (e.g. Classical Greek, Latin, or Coptic), and on old language variants (e.g. Middle High German or Old French). Massive digitization campaigns not only allowed preservation of historical documents but also increased significantly the distribution of these texts to a broad spectrum of researchers. However, only in recent years has a movement occurred from 'digital reading' (search through digital catalogues and display of texts with progressive zoom facilities) to 'digital analysis' of the texts. However, it goes rarely beyond keyword search and quantitative (i.e. statistical) measurements and lacks interpretation/contextualization/specification of

individual words and terms. This may lead to inaccurate/wrong interpretations because, for example, words have different meanings according to the context and simple statistics about co-occurrences of two terms do not automatically imply causality between them ('Romans were eating a lot of bread' and 'Romans died younger' are two true assertions but they do not automatically imply that 'Romans died younger because they were eating a lot of bread'. For the latter utterance one may need more detailed analysis beyond statistical measurements).

Approximate dates, unclear places and less documented existence of persons are just some examples of types of vagueness and uncertainty encountered in such documents. Ignoring this characteristic of historical documents not only may lead to an incomplete digital recording, but also has important consequences on the interpretation process whether it is done by man or machine. The project HerCoRe (Hermeneutic and Computer-based Analysis of Reliability Consistency and Vagueness in Historical Texts) has investigated, in a concrete multilingual corpus, how vagueness and uncertainty can be recorded digitally and used for hermeneutic interpretation. Most of the current approaches tend to reduce expressions like 'towards the beginning of the 18th century' to crisp values like 1700, or in best case an interval 1700–1720. An event happening in 1721 is thus no longer considered early of the 18th century. Vague expressions such as 'it is probable that the event occurred' are reduced to 'the event occurred'. Thus important information is omitted and the analysis lacks accuracy whether it is performed by computers or humans (hermeneutic).

Methods from semantic web and natural language processing have the potential to enrich these digital historical recordings with intelligent functionalities, so that new insights in the materials are gained or hitherto hidden connections between events or persons are revealed and can be visualized.

The basis of all these new functionalities is in many cases a deep annotation of the contents. Methods of machine learning by using large training resources are difficult to be applied because these training resources are missing. Additionally the embedding of vagueness and uncertainty in the analysis (throughout a genuine method in hermeneutic approach) is often neglected Following a brief analysis of the state of the art in section 2, we proceed to explore the reliability and consistency of two works of Dimitrie Cantemir, a humanist of the early 18th century and expert (among others) in the culture and history of the Ottoman Empire, by using advanced computer science tools such as ontologies and inferences. We show how uncertain and vague information, so often neglected by state-of-the-art Digital Humanities approaches, can be modelled and included in the computer-aided analysis of texts.

Overview of State-of-the-Art Annotation of Vagueness and Uncertainty in Digital Humanities (DH)

Language is one of the most powerful tools people use to express themselves and their cultural as well as social environment. Words have meanings, follow certain syntactic rules and many are just labels for concepts. In real world, it is quite common that information cannot be classified as 100% true or false. However, all these features we associate with a word are part of our background knowledge. Computers represent words as a sequence of '1' and '0'. All background information (semantic, syntactic, or conceptual) has to be supplied through annotations.

Many annotation systems implemented up to now tend to ignore an intrinsic character of natural language which is of great importance for the analysis of historical documents: the vagueness of almost every utterance.[1] Two parameters, vagueness and uncertainty, are mostly neglected in annotations of historical documents. Studies on natural language have shown that even in technical domains (e.g. manuals for equipment) language expressions are vague. Texts dealing with past events and written at a time when access to information was restricted are full of utterances expressing doubt or remaining vague about events or places. Additionally, variations in spelling, use of oral tradition instead of documented sources lead to uncertainty especially about named places (e.g. 'not far away from the Danube'), persons ('known as the son of…') and dates (e.g. 'after the last call of the muezzin').

TEI (Text Encoding Initiative)[2] is the state-of-the-art standard for annotation of historical documents. It offers three possibilities for recording (i.e. annotating) vagueness.

1. Using the <note> element the user can write unstructured text, mentioning the degree and scope of the identified vague aspect.
2. The <certainty> element: it offers the possibility of structuring the information about vagueness in the following ways. The <certainty> element can refer to the name of the annotation tag considered uncertain (e.g. a person or a place name), the position in text where the annotation tag starts, or a value of an attribute contained in the annotation tag. Through the attribute @degree it is possible to refine the level of certainty. The <certainty> element can refer to one or more annotation elements through XPath expressions.
3. The <precision> element, which can be applied for any numerical value (a date, or a measure), indicates the numerical accuracy associated with some aspects

1 Walther von Hahn, "Vagheit bei der Verwendung von Fachsprachen", in: Lothar Hoffmann, Hartwig Kalverkämper and Herbert Ernst Wiegand (eds.): *Fachsprachen. Ein Internationales Handbuch zur Fachsprachenforschung und Terminologiewissenschaft* (Berlin 1998), pp. 383 – 390.

2 www.tei.org.

of text markup. If a standard value is precise and known, the user can express it with the element <precision> and the attribute @stdDev, which represents its standard deviation.

Additional TEI offers the possibility of indicating the responsibility for the whole content or partial annotators. In this way the user can specify if the vagueness is due to the author, the quoted source or the editor. Such detailed information is usually done manually.[3]

However, if we want to annotate vague expressions or other levels of uncertainty, we can face several drawbacks of the TEI approach:

1. Overlapping annotations concerning vagueness are possible only as stand-off annotation. However, stand-off annotation in TEI is extremely complicated.
2. There are different levels of vagueness introduced by the author by the referred source, dating etc. Not all of these sources of vagueness can be specified with the <response> tag which can be attached just to individuals.
3. The <precision> element can be specified just for numerical values. An expression 'some kilometres south from the city' introduces a non-numerical vague coordinate. When we speak about historical documents, geo-location of the place is not always possible.
4. There is no reasoner (computer programme for realizing logical connections and inferences between assertions) which can be applied to the TEI annotation.

The HerCoRe project has developed an annotation-tool which is able to handle drawbacks 1 and 2.

HerCoRe Case Study for Mixing Methods from Humanities and Computer Sciences

As mentioned above, digital processing of historical documents is particularly challenging with regard to vagueness and uncertainty. Approximate dates, unclear places and less documented existence of persons are just some examples of the types of vagueness and uncertainty encountered in such documents. Ignoring these characteristics of historical documents may lead not only to an incomplete digital recording but also to important consequences for the interpretation process whether it is done by man or machine.

3 More details and examples of using these TEI-elements can be found on https://web.uvic.ca/lancenrd/martin/guidelines/tei_CE.html.

The project is investigating Latin manuscripts and their translations in German and Romanian of two most relevant works on the Balkan history, written in the middle of 18th century by Dimitrie Cantemir. Section *Rationale* introduces the author, his relevance for Ottoman and Balkan studies and describes the still unsolved research issues. In section *Corpus*, we provide details about the corpus, the transmission and the selection of texts for the HerCoRe project. Finally, in section *Mixed-Method Investigation Approach* we present the mixed-method approach including a pure humanistic (hermeneutic) approach as well as modern technologies from computer science.

Rationale

For historians, research on the Ottoman Empire is challenging given the time interval (more than 700 years) and the number of languages and cultures involved. Many documents in Turkey or the Balkans have become accessible physically only in recent years; still they cannot be fully analysed due to language or script barriers. Only within the Balkans (as a former part of the Ottoman Empire) do we encounter four alphabets (Latin, Cyrillic, Arabic and Greek) and several national languages in addition to the ones considered official at the time of Ottoman Empire (Greek and Slavonic in the churches, Ottoman Turkish and sometimes also Latin). For many centuries, information about the Ottoman Empire was sparse and limited to a few diplomats and travellers who reported it in a fragmentary way. In this context the works of Dimitrie Cantemir represent an essential milestone in the reception of the Ottoman society in Western Europe. Born in Romania in 1673 as the son of a Moldavian ruler (Wojevode), Cantemir was sent to Istanbul as guarantee for the loyalty of his father. He continued his education (which he begun in Moldavia with Greek monks) in Istanbul and became one of the last universal scientists of the Age of Enlightenment. D. Cantemir published works in history, geography, philosophy and music, being the first one to record Ottoman music on paper. Of particular relevance were two of his works, written at the request of the Royal Academy of Sciences in Berlin, of which he became a member: 'Descriptio Moldaviae' (The History of Moldavia, DM) and 'The Rise and Decay of the Ottoman Empire' (HO). These works were written in Russia, at the court of the Tsar Peter the Great, where he received asylum after a short period of ruling over Moldavia. This detail is important as it may have implications about the two mentioned works: it is not clear if historical mistakes in his writings were due to lack of sources or were done deliberately in order to comply with the ideology and the religion of his new protector, the tsar.

The two works mentioned above reached London after his death, being brought by his son as a diplomat. They were translated by Georges Tindal[4] immediately. The

4 https://en.wikipedia.org/wiki/George_Tindall.

originals were lost shortly afterwards. Tindal's translations were used for French, German and, later in the 19th century, Romanian versions. All these translations represented for more than one century the most comprehensive information about the Ottoman Empire. Even later historians at the end of 19th and beginning of 20th century, who claim to quote Cantemir, quote in reality one of these translations, especially the English and the German one. In the 1920s, several researchers raised doubts about the accuracy of Cantemir claims. Again, they compared translations with historical facts described, meanwhile, by other researchers. Only in the second half of the 20th century were the Latin manuscripts of Cantemir's works discovered (three manuscripts for Descriptio Moldaviae and one for History of the Ottoman Empire). Even the shallow comparison between these manuscripts and the translations revealed that Tindal had made serious deviations from the original, for example, paragraphs written in Arabic were omitted and some facts or descriptions of persons or traditions were changed. We will describe in detail these differences in section *Corpus*.

For researchers in Ottoman studies, Cantemir's writings are of great importance, yet a thorough comparison of existing manuscripts and translations is still lacking. This aspect could not have been realized until today as the number of scripts and languages involved as well as the size of the material are difficult to be handled by humans. Additionally, recent digitalization campaigns in Turkey give access for the first time to documents from the 18th century, which may support Cantemir's claims.

We focus our research on three directions:

1) Reliability: Questions to be investigated here are: are the quotations made by Cantemir himself grounded in authenticity? Is there concordance between his degree of trust in these sources and the current knowledge about them (e.g. is there any evidence that a person which Cantemir claims to have spoken to, really lived in that time?). In the translations, too, the insertions of the editors or translators have to be investigated.
2) Consistency: Does Cantemir keep a constant opinion about persons, events, facts across the same work (e.g. in the main text and his own annotations)? The two main works of the corpus have a common pool of persons, events and places; thus we want to investigate also if the author keeps to his opinion or describe things according to the target public and possible political context.
3) Vagueness: Cantemir's works are full of vague expressions. He states that even 'I do not dare to decide what the truth is about this matter, given the high darkness of this story'. The corpus analysis tried to emphasize these expressions and analyse if they were used on purpose (political or tactical reasons) or just as a stylistic tool. Also, we have a look at the translations in order to see to what extent they preserve the degree of vagueness stated by Cantemir.

The project HerCoRe, therefore, addresses a real important research problem for Ottoman studies. At the same time, the complexity of the corpus and of the user requests (i.e. the historians' demand for a digital tool) calls for a different approach to the current state-of-art in digital humanities. We will discuss it in detail in section *Mixed-Method Investigation Approach*.

Corpus

As mentioned in section *Rationale*, our research focuses on the two books of Dimitrie Cantemir covering the history of the Ottoman Empire (HO) and the description of his country Moldavia (DM). From the numerous translations and editions, we chose:

For HO: the edition of the Latin manuscript found in Harvard University, the translation in Romanian of this edition, and the historical translation in German from 1771.[5]

For DM: the edition collating the three existent Latin manuscripts from St. Petersburg and Odessa,[6] the translation in Romanian of this edition[7] and the historical translation in German from 1745.[8]

The choice had a scientific and a pragmatic reason. From the scientific point of view, we used in both cases the most recently available editions which collated all historical manuscripts known until now. The Latin editions and the Romanian translations were realized by well-known specialists, one of them being part of the HerCoRe team. The German historical translation is relevant for two reasons. Firstly, it was done after the English translation of Tindal preserved most of his errors but corrected some of them. Secondly, it was one of the most used in the 18th and 19th centuries, especially for further translations. In this respect, it offers a different narrative to those of the recent editions and translations.

From a pragmatic point of view, we chose those documents that were also available in machine readable versions (with one exception[9]). Tindal's translation is for the moment available only in PDF format. Given the number of foreign words in the text and the relatively low PDF-quality of existent files, any PDF-to-text conversion would have failed. Given the time span of the project, digitization of further documents was not considered.

5 Dimitrie Cantemir, *Beschreibung der Moldau*, Faksimiledruck der Originalausgabe von 1771 (Frankfurt und Leipzig: 1771).

6 Florentina Nicolae and Ioana Costa (eds.), *Descrierea stării Moldaviei: în vechime şi azi* (Bucureşti: Academia Română-Fundaţia Naţională pentru Ştiinţă şi Artă, 2017), 2019.

7 Nicolae and Costa, *Descrierea*.

8 Dimitrie Cantemir, *Geschichte des osmanischen Reichs nach seinem Anwachse und Abnehmen* (Hamburg: Herold, 1745).

9 Cantemir, *Beschreibung*.

The HerCoRe Corpus contains three types of texts, each of them with its own particularities:

- 1. Editions of Latin manuscripts[10]
  These documents include not only the original text but also markup and footnotes of the editor. The markup (e.g. different bracket types) has to be removed from the basic text and reinserted as annotations. They represent a challenge for any tool processing natural language. The main language of the manuscripts is Latin. However, Cantemir reproduces names of persons, places and quotations in original Ottoman Turkish written with Arabic characters and also provides a Latin transliteration of the same. Often the manuscript contains a translation of these paragraphs into Latin. Also, Greek quotations are encountered. Romanian words are written (in contrast with the tradition in the 18th century) with the Latin alphabet. As far as it is known, it is the first attempt of using Latin instead of Cyrillic alphabet for this purpose. The transcriptions are however not standardized and, as a consequence, a name has several transcriptions. For example, the name of the Moldavian capital (Iași) is encountered in 32 transcription forms (Iassi, Jassy etc.).
- 2. Modern translations of Latin manuscripts into Romanian[11]
  In these documents the translators inserted footnotes with explanations. They retained the transcriptions of Romanian names as in the Latin manuscript and at places offered explanations in footnotes. Turkish and Greek phrases were preserved in the original alphabets and transcriptions. The translator kept Cantemir's transcription style even though it did not correspond to the modern Romanian language. The translation is, however, adapted to suit the modern reader and uses to a great extent current Romania grammar and lexicon.
- 3. Historical translations into German[12]
  These translations mainly came after the historical English translation by Tindal and (as the editors mention in the foreword) with an additional consultation in case of DHO of the French translation (which however had as basis this English version as well). The Ottoman Turkish paragraphs were omitted completely (Arabic original and Latin transcription). If Cantemir offered a Latin translation of the paragraphs then they consequently appeared in German. In all other cases the Turkish quotations were simply omitted. Romanian and Turkish names and

10 Octavian Gordon, Florentina Nicolae, Monica Vasileanu and Ioana Costa (eds.), *Cantemir Dimitrie, Istoria măririi şi decăderii Curţii othmane*, 2 volumes (Bucureşti, Academia Română-Fundaţia Naţională pentru Ştiinţă şi Artă, 2015), Nicolae and Costa, *Descrierea*.

11 Costa et al., *Istoria*; Nicolae and Costa, *Descrierea*.

12 Cantemir, *Geschichte*; Cantemir, *Beschreibung*.

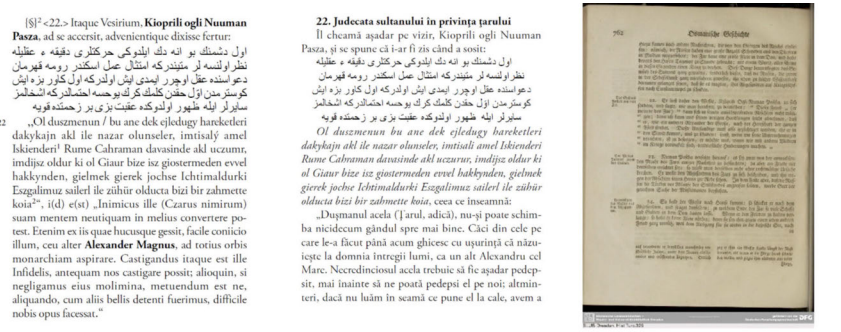
denominations of occupations were modified and adapted to German phonetics. For example, 'Kioprili ogli Nuuman Pasza' in the Latin original was preserved in the Romanian modern translation (although the modern form of Pasza would be in Romanian Paşa), but changed to 'Kjüprili Ogli Numan Pascha' in the German version.

These variations make practically impossible a consistent character-based search over the entire corpus.

The German translation also includes footnotes by the editor, marked with (V). Both German translations were printed with black-letter fonts (Fraktur). This is an additional challenge both for an untrained person and for the preparation of machine readable versions.

The example illustrated above can be retrieved in Figure 1.

Fig. 1: Same paragraph in Latin manuscript and translations. From left to right: Latin edition,[13] Romanian edition,[14] German edition.[15]



{§}[2] <22.> Itaque Vesirium, **Kioprili ogli Nuuman Pasza**, ad se accersit, advenientique dixisse fertur:
اول دشمنك بو انه دك ايلدوكى حركتلرى دقيقه ء عقليله
نظر اولنسه لر متيندركه امتثال عمل اسكندر رومه قهرمان
دعواسنده عقل اوجرر ايمدى ايش اولدركه اول كاور بزه ايش
كوسترمدن اوّل حفدن كلمك كرك بوحسه احتمالدركه اشخالمز
سايرلر ايله ظهور اولدوكده عقبت بزى بر زحمتده قويه
522 „Ol duszmenun / bu ane dek ejledugy hareketleri dakykajn akl ile nazar olunseler, imtisalý amel Iskienderi[1] Rume Cahraman davasinde akl uczumr, imdijsz oldur ki ol Giaur bize isz giostermeden evvel hakkynden, gielmek gierek jochse Ichtimaldurki Eszgalimuz sailerl ile zühür olducta bizi bir zahmette koia[2*], i(d) e(st) „Inimicus ille (Czarus nimirum) suam mentem neutiquam in melius convertere potest. Etenim ex iis quae hucusque gessit, facile coniicio illum, ceu alter **Alexander Magnus**, ad totius orbis monarchiam aspirare. Castigandus itaque est ille Infidelis, antequam nos castigare possit; aliaquin, si negligamus eius molimina, metuendum est ne, aliquando, cum aliis bellis detenti fuerimus, difficile nobis opus facessat."

22. Judecata sultanului în privinţa ţarului
Îl cheamă aşadar pe vizir, Kioprili ogli Nuuman Pasza, şi se spune că i-ar fi zis când a sosit:
اول دشمنك بو انه دك ايلدوكى حركتلرى دقيقه ء عقليله
نظر اولنسه لر متيندركه امتثال عمل اسكندر رومه قهرمان
دعواسنده عقل اوجرر ايمدى ايش اولدركه اول كاور بزه ايش
كوسترمدن اوّل حفدن كلمك كرك بوحسه احتمالدركه اشخالمز
سايرلر ايله ظهور اولدوكده عقبت بزى بر زحمتده قويه
Ol duszmenun bu ane dek ejledugy hareketleri dakykajn akl ile nazar olunseler, imtisalî amel Iskienderi Rume Cahraman davasinde akl uczurur, imdijsz oldur ki ol Giaur bize isz giostermeden evvel hakkynden, gielmek gierek jochse Ichtimaldurki Eszgalimuz sailerl ile zühür olducta bizi bir zahmette koia, ceea ce înseamnă:
„Duşmanul acela (Ţarul, adică), nu-şi poate schimba nicidecum gândul spre mai bine. Căci din cele pe care le-a făcut până acum ghicesc cu uşurinţă că năzuieşte la domnia întregii lumi, ca un alt Alexandru cel Mare. Necredincios ul acela trebuie să fie aşadar pedepsit, mai înainte să ne poată pedepsi el pe noi; altminteri, dacă nu luăm în seamă ce pune el la cale, avem a

These particularities are real challenges for the preparation of the machine readable corpus. In the following section, we will describe which steps had to be taken for an appropriate digital representation and the workflow plan we prepared in order to combine information technology and hermeneutic methods.

13 Nicolae and Costa, *Descrierea*.
14 Costa et al., *Istoria*.
15 Cantemir, *Geschichte*.

Mixed-Method Investigation Approach

The complexity of the research questions presented in section *Rationale* as well as the heterogeneity of the data as described in *Corpus* made an exclusively hermeneutic approach and a fully automatic processing exercise impossible. Thus a mixed method paradigm seemed to suit this research project. Our aim was to involve both computer scientists and researchers from relevant humanities fields (Ottoman history, German and Romanian linguistics, Romanian history and Latin studies) in all phases of the investigation. Our research built on three pillars: data modelling, data processing and data visualization as well as investigation. In this section, we will explain how a mixed approach was used for each of the three pillars.

Data Modelling

This step involves data acquisition followed by recording of data in an appropriate model. Additionally, we scrutinized here which additional knowledge was needed and how it could be provided in a digital form.

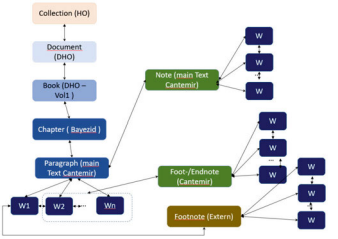
Four axioms guided our work:

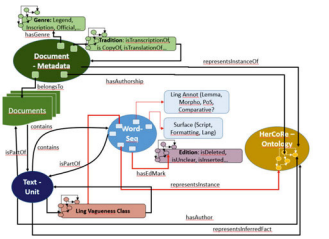
- A1. At any time, human researcher or automatic process should be able to identify the text source: main text, side- or end-annotation done by the author of the main text, editor's/translator's annotation.
- A2. Editorial markup in the Latin manuscript has to be preserved as long as it concerns insertions, removals or empty spaces in the manuscript.
- A3. Layout information (e.g. italic or bold formatting) underlies a meaning and has to be preserved.
- A4. Text normalization is done only in cases where rules can be applied without exceptions and they do not interfere in further analysis. This means we would not normalize across languages the transcription of proper names or specific occupations as these variations could be a source of misinterpretations and wrong translations. On the other side, we normalize the 18th century German diacritics style (e.g. $\overset{e}{u}$ -> ü, $\overset{e}{a}$->ä, ->s) as well as old writing rules for which non-ambiguous transformations are possible (e.g. bey->bei).

We use two models for our corpus. The first one, the Document Collection Model, defines the corpus structure. A Collection represents one work (HO or DM) including the Latin manuscript as well as the Romanian and German translations. Each collection contains several documents; in our case, one for each considered language. Substructures such as book, chapter and paragraph follow. The smallest unit in this structure is a word-unit (a string separated by empty spaces). Each word is recorded

as an object identified by a unique ID. These IDs can be used in order to define non-hierarchical data structures such as notes written either by the author or by the editor/translator. This model is presented in Figure 2.

Figure 2: Document collection model. *Figure 3: Annotation model.*





All additional information (linguistic, editorial, layout, knowledge, vagueness) is attached to groups of word-units and is used in the following annotation. In this way, we allow annotations over discontinuous group of word-units and overlapping of several annotation layers. The annotation model is presented in figure 3 and shows different annotation types (corresponding to different layers) defined on sets of word-units. In the system, a word-unit is an object with a unique identifier and several labels (normalized writing, original writing) and pointers to annotation objects. This model comes with great advantages: it allows different ways of display of the text including the changes or corrections made in the initial text (e.g. OCR corrections).

In the following section, we will discuss in detail the knowledge and vagueness annotations which are central aspects of our mixed-method research paradigm.

Levels of Vagueness and Uncertainty in HerCoRe —Challenges and Solutions

As mentioned in the previous sections, in the HerCoRe corpus we encounter different types of vagueness and uncertainty which need to be annotated in order to ensure a proper qualitative/hermeneutic analysis. The following classes of uncertainty /vagueness can be defined:

A. Linguistic

- *Linguistic uncertainty* refers to expressions of temporal or geographical information, for example, 'some years before', 'more or less 5 kilometres', 'two days far

from' or 'at the beginning of 13th century'. These expressions assume that a certain location exists or that a certain event has taken place but their geographical or temporal coordinates cannot be fixed. This type of uncertainty is witnessed mostly at the lexical level.

- *Linguistic vagueness* refers to those expressions that induce a subjective position of the narrative, for example, 'it was said', 'I am not sure', or 'it is supposed'. Such vagueness indicators may be lexical, or they may even be syntactic, for example, through the use of language-specific verb modalities or time: 'would have been', 'should have taken place'. This syntactic vagueness is challenging because it is difficult to say if it is just a matter of narrative style or a vagueness marker used intentionally.

B. Knowledge

- Proper names (geographical, persons). Uncertainty is often related to writing variants or political renaming of places over time which may lead to confusions (e.g. Izmir vs. Izmit, two different cities in the Ottoman Empire; Sultan Bayezid—there were two rulers with this name). Vagueness is connected with geographical areas with borders changing over time (e.g. Ottoman Empire, Principality of Moldavia) or regions for which the exact coordinates cannot be defined (e.g. Syrfia).
- Concepts. They are domains and in our case also country specific. From the linguistic point of view, the lexicalization of these concepts is not vague or uncertain. We can take for example, the word 'Vizier' (a political functionary in the Ottoman Empire and equivalent to a minister in modern times). Used in the plural, it denominates all such ministers, but when it is used in the singular it refers usually to the Great Vizier (the leader). Depending on the context, it can also point to a certain person named in a previous sentence (anaphora). Another example may be a 'place of pilgrimage' which has different definitions depending on the religion in question.

C. Sources (factual uncertainty)

- Uncertainty of sources refers to good documented sources (e.g. chronicles) for which certain parameters are not precise (e.g. author or publication date). However, the physical objects being referred to exist in many cases and they can be consulted.
- We classify as vague sources with oral transmissions (e.g. legends, quotation of spoken opinions).

D. Editorial markup usually describes uncertain parts (word which cannot be deciphered from the manuscript, damaged places, partially reconstructions, etc.).

Processing Linguistic Vagueness and Uncertainty

For the annotation of linguistic vagueness and uncertainty we use a mixed approach based on collaboration between linguists (one for each language involved) and a computer scientist.

The linguists prepared a list of language-dependent vagueness following Pinkal's classification[16]

- 1. Comparatives, inexact adjectives, for example *'mehr/more,' 'größer/bigger,' 'älter/older'*
- 2. Non-intersectives, for example 'vermeintlich/supposed,' 'so-genannt/so-called'
- 3. Hedges, for example 'ziemlich/quite,' 'einigermaßen/approximately,' 'etwa/about'
- 4. Inexact measures, such as '*4 Tagereisen/4 days' trip,' '10 Fuß /10 feet'*
- 5. Modals (attitudes), for example 'vielleicht/maybe,' 'hoffentlich/hopefully;' and subjunctive verbs
- 6. Lexical quotation markers, such as 'es wurde gesagt /it is said'
- 7. Vague quantifiers, such as 'viele,' 'meistens /mostly'
- 8. Complex quantifiers, for example '*etwa die Hälfte von den 20–30 tausend Soldaten / about a half of the 20–30 thousand soldiers*'
- 9. Numbers
- 10. Range expressions, for example '*Anfang des 18. Jhds./beginning of the 18th century'*
- 12. Unclear person, such as *'der ehemaligen Herzog / the former duke*'
- 13. Unclear time, for example *'in alten Zeiten /in olden times*'

The list of vagueness indicators was created manually and then enriched semi-automatically with elements of synsets (synonyms) extracted from the corresponding language-specific WordNet. The word semi-automatic meant that for each term in the vagueness list, its synset was extracted automatically from the WordNet. Then a human evaluated if the synset elements were part of the vocabulary of the 18th cen-

16 Manfred Pinkal, Logik und Lexikon: *Die Semantik des Unbestimmten* (Berlin/New York: de Gruyter, 1985).

tury or, in a positive case, if the semantics was the same. Finally, we created a vagueness and uncertainty indicator list for each language encoded in XML format.[17]

The annotation of these vagueness markers in the text is facilitated by an automatic morphological annotation, assigning to each token a part-of-speech and a morphological value including, for example comparative degrees for adjectives. In order to ensure uniform annotation across languages, we decided to use the CoNLL-U (Universal Dependencies[18]) format. It was necessary as all the languages involved had a rich morphology and, therefore, inflected/derived/conjugated forms of the items in the vagueness list had to be detected.

Development of a Rich Knowledge Base including Vague and Uncertain Concepts and Relations.

The HerCoRe knowledge base is a type of Fuzzy OWL[19] Ontology trying to model the Ottoman Empire world with its administrative, social, geographical and religious facets.

The following aspects need particular attention:

The modelling of geographical identities and their respective political entities is one such aspect. Political entities (e.g. countries) often tend to share their names with some geographical entities. Political entities retain their names but often change their borders. Thus, we considered as fixed unambiguous individuals those geographical elements still visible today. Historically attested geographical zones which did not exist were modelled as fuzzy concepts.

Political entities were defined as a sum of several historical contexts. Additionally, we introduced the concept of 'historical zone' in order to model concepts such as 'Europe' which from the point of view of the Ottoman Empire, for example began at its borders with Hungary, or 'the Balkans' which for the Ottoman Empire was represented by the Wallachia and Moldavia principalities. In Figure 4, we illustrate the representation of the vague geographical entity of Syrfia which, depending on the historical sources, was situated in three different regions of Europe. Originally we describe three political contexts, each having the same weight. Depending on the additional information (in this case a map), the confidence level can be increased for

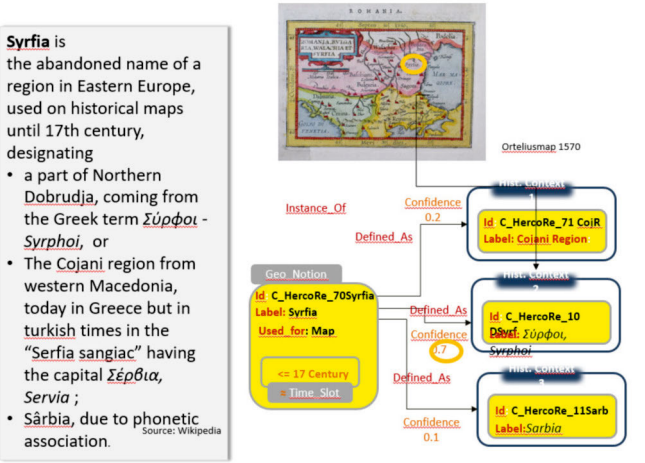
17 Alptug Güney, Walther von Hahn, Cristina Vertan, "Checking reliability of quotations in historical texts – A digital humanities approach", in: Steven Krauwer and Darjy Fiser (Eds.): *Proceedings of Twin Talks 2 and 3, 2020 Understanding and Facilitating Collaboration in Digital Humanities* (2020), https://ceur-ws.org/Vol-2717/.

18 https://universaldependencies.org/format.html.

19 OWL – Web Ontology Language.

one of the hypotheses. In the example in figure 4, the placement on the map near the Black Sea reinforces one of the three hypotheses.

Figure 4: Example of modelling multiple political contexts for one proper name (Syrfia).



Time intervals and geographical positions include fuzzy concepts which allow us to model uncertain dates and coordinates.[20] For the modelling of fuzzy time intervals, we follow the approach described by Métais et al.[21]

A particular challenge is posed by the representation of domain-specific fuzzy concepts. For example, the concept 'place of pilgrimage' is a geographical entity usually containing a monument and being visited by hundreds of people over the years. The complexity of the representation is revealed here by time dependency. A place cannot be an object of pilgrimage from the beginning, but it becomes one (or it loses this qualification) depending on the number of recorded visitors. The cooperation of a human researcher proves unavoidable to determine it: such knowledge cannot be inferred from elsewhere. In collaboration with the researcher in Ottoman studies we defined a threshold for the fuzzy representation of a place of pilgrimage. In

20 Cristina Vertan, "Annotation of vague and uncertain information in historical texts", in: M. Slavcheva et al. (Eds.): *Knowledge, Language, Models* (INCOMA Ltd. Shoumen, Bulgaria, 2020).

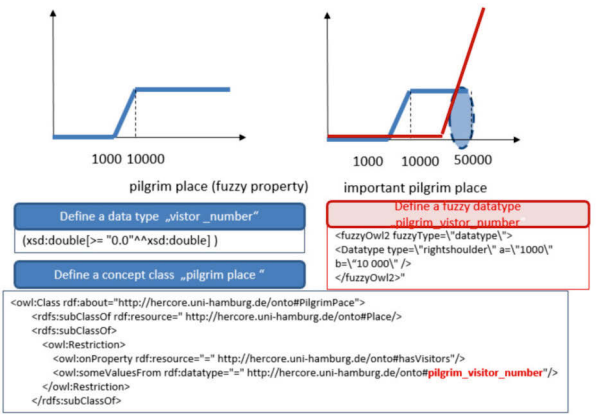
21 Elisabeth Métais, Fatma Ghorbelb, Fayçal Hamdia et al.: "Representing Imprecise Time Intervals in OWL 2", in: *Enterprise Modelling and Information Systems Architectures* Vol. 13, 2018, https://hal.archives-ouvertes.fr/hal-02467393/file/191-Article%20Text-478-1-10-20180322.pdf [accessed: 06.2021].

figure 5, we illustrate this step and show how this is expressed in OWL by using the fuzzy extension.

The ontology contains for the moment more than 500 classes, 250 object properties (relations between classes), and 3,000 individuals.

In order to deal with multilinguality, we attach to each individual the names used in the German translation, the official Ottoman documents and in the Romanian sources.

Figure 5: Fuzzy modelling and representation of the concept 'place of pilgrimage'.



A particular problem in the ontological representation of individuals (concrete representation of classes e.g. Istanbul is an individual of Class 'City') is represented by the multiple—sometimes contradictory—relationships between individuals according to different historical sources.

The model behind OWL is the triple association of <Subject><Predicate><Object> where the <Predicate> is represented by a relationship (e.g. <Constantinople><was_conquered_by><Sultan_Mehmet_II>). This underlying model does not allow the inclusion of a source of truth for this assertion. In our system we overcame this limitation with the following approach:

– the researcher in Ottoman studies identified three major Ottoman works that were used in Cantemir's time (and was also quoted by him) to write the history of the Ottoman Empire. If a certain relation such as 'was_conquered_by' is dependent on historical evidence, we define sub-relations such as 'was_conquered_by_Source1',

'was_conquered_by_Source2', or 'was_conquered_by_Source3'. If there is no historical evidence, we add an additional sub-relation 'was_conquered_by_unclear'. If no sub-relation is defined, we consider the relation as representing a historical fact generally accepted by the research community.

Each individual in the ontology received a unique identifier. This identifier was used for linking individuals with their particular realizations in the corpus.

Representation of Factual Uncertainty

Under 'factual uncertainty' come all paragraphs related to quotations (direct and indirect) from other sources. The quotation style differs dramatically from the modern one. Often paragraphs are quoted but sources are not given. Often quotations are rephrased. There is no bibliographical list. Thus a hermeneutic study is necessary.

Given the fact that DM (Description of Moldavia) practically lacks explicit quotations, the study concentrates on the analysis of HO (History of Ottoman Empire). It follows three directions:

- 1. Identification of the works quoted by Cantemir
- 2. Collection of expressions used by Cantemir when quoting
- 3. Comparison of related facts described by Cantemir against the sources identified in 1).

The study was extremely important as it revealed that not always there was one-to-one correspondence between Cantemir's quotation and the way he reported the event/fact. We identified several cases in which Cantemir had rated a certain event as 'sure' while the available sources reported it contrarily or omitted the fact altogether. On the other hand, there were cases in which Cantemir presented facts, which were reported in other sources, as quite vague or unsure.

This is a very important result which shows us that linguistic vagueness is not the only indicator for assessing the degree of truth for an utterance. The manual annotation, based on the hermeneutic investigation, cannot be avoided. A list of most important works quoted by Cantemir can be found in Vertan's further writings.[22]

Further Work and Discussions

The digital representation and annotation of the corpus was more challenging than expected. We identified the following challenges:

22 Vertan, "Annotation".

- 1. Existent digital versions of some parts of the corpus were not appropriate for our work (see section 3.1)
- 2. The development of the knowledge base took much longer than expected. The collaboration between the researcher in Ottoman studies and the computer scientist was more intensive than expected. The reasons for this are twofold:
  - The most appropriate tool to be used for the development of the ontology is Protégé[23]. The web-based version of the tool is meant to address a non-specialist difficulty and differs from the client version. Thus parallel work on parts of Ontology was impossible
  - The user-friendly fuzzy module for Protégé[24] runs only on an older version of Protégé. On the other hand, only the newest version allows some Description Logics features which we needed. Fuzzy classes and relations have to be written manually in OWL, this operation can be done only by the computer scientist
- 3. The query module for the fuzzy ontology is a programme called reasoner; it is able to use fuzzy logic and realize inferences (judgements). (Our assumption that such a system already existed was partially confirmed.) The reasoner[25] process successfully fuzzy concepts and relations. They, however, fail when processing complex classes modelled with traditional description logics

At the moment of completion of this article, the work concentrates on the development of an appropriate query /interrogation module for the corpus and the presentation of results. The general aim is to present to the user all possible solutions of the query even if not all of them have the same degree of truth (plausibility). The system should only display possible investigation paths; the researcher should use hermeneutic methods for the interpretation.

The development of the ontology turned to be a powerful and new tool for the hermeneutic research in Ottoman studies. At the beginning of the project, we expected that the use of the completed ontology would be of great benefit to the Ottoman studies research. It turned out that the efforts to structure the information—as demanded by the ontology—and the necessity to feed the knowledge base with extensive information triggered the need for further explorations in the

23 https://protege.stanford.edu/.

24 Fernando Bobillo, Miguel Delgado and Juan Gomez-Romero: "Reasoning in Fuzzy OWL 2 with DeLorean, in: Fernando Bobillo, Paulo C. G. Costa, Claudia d'Amato, Nicola Fanizzi, Kathryn B. Laskey, Kenneth J. Laskey, Thomas Lukasiewicz, Matthias Nickles and Michael Pool (Eds.): *Uncertainity Reasoning for the Semantic Web II* (Berlin/Heidelberg: Springer Verlag, 2013).

25 Fernando Bobillo and Umberto Straccia, "Fuzzy Ontology Representation using OWL 2", ht tp://arxiv.org/pdf/1009.3391.pdf [accessed: 15.02.2016].

archives. A PhD in Ottoman studies on Dimitrie Cantemir's works will be completed soon as a direct result of this project.

In parallel, we established a network of researchers working on applying fuzzy reasoning on real data, in particular humanities. It turned out that many of the reported solutions were tested on small, less complex data and that the systems indeed had limitations on how to handle more complex corpora as the ones presented here. We expect that parts of the HerCore corpus as well as the ontology will be used in the future for system improvements.

The annotation tool, the ontology and a limited part of the annotated corpus will be publicly available under CC BY licence.

Acknowledgments

The author thanks all participants in and contributors to the HerCoRe project: Alptug Güney (University of Samsun, formerly of University of Hamburg), Walther von Hahn (University of Hamburg), Ioana Costa (University of Bucharest), Yavuz Köse (University of Vienna), Rafael Quirros Marin and Miguel Pedregosa (University of Granada, formerly of University of Hamburg).

Bibliography

Bobillo, Fernando and Umberto Straccia, "Fuzzy Ontology Representation using OWL 2", http://arxiv.org/pdf/1009.3391.pdf [accessed: 15.02.2016].

Bobillo, Fernando, Miguel Delgado and Juan Gomez-Romero, "Reasoning in Fuzzy OWL 2 with DeLorean, In *Uncertainity Reasoning for the Semantic Web II* edited by Fernando Bobillo, Paulo C. G. Costa, Claudia d'Amato, Nicola Fanizzi, Kathryn B. Laskey, Kenneth J. Laskey, Thomas Lukasiewicz, Matthias Nickles and Michael Pool, (Berlin/Heidelberg: Springer Verlag 2013).

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

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

Costa, Ioana and Florentina Nicolae (eds.): *Descrierea stării Moldaviei* (Bucureşti: Academia Română-Fundaţia Naţională pentru Ştiinţă şi Artă, 2017).

Gordon, Octavian, Florentina Nicolae, Monica Vasileanu and Ioana Costa (eds.): *Cantemir, Dimitrie, Istoria măririi şi decăderii Curţii othmane*, 2 volumes, (Bucureşti, Academia Română-Fundaţia Naţională pentru Ştiinţă şi Artă, 2015).

Güney, Alptug, Walter von Hahn and Cristina Vertan: "Checking reliability of quotations in historical texts – A digital humanities approach", In *Proceedings of Twin Talks 2 and 3, 2020 Understanding and Facilitating Collaboration in Digital Humanities*

edited by Steven Krauwer and Darjy Fiser (2020), https://ceur-ws.org/Vol-2717/.

Métais, Elisabeth, Fatma Ghorbel, Fayçal Hamdi, Nebrasse Ellouze, Noura Herradi and Assia Soukane, "Representing Imprecise Time Intervals in OWL 2", In *Enterprise Modelling and Information Systems Architectures* Vol. 13, 2018, https://hal.archives-ouvertes.fr/hal-02467393/file/191-Article%20Text-478-1-10-20180322.pdf [accessed: 06.2021].

Pinkal, Manfred, *Logik und Lexikon: Die Semantik des Unbestimmten* (Berlin/New York: de Gruyter, 1985).

Vertan, Cristina, "Annotation of vague and uncertain information in historical texts", In *Knowledge, Language, Models* edited by M. Slavcheva et al., (INCOMA Ltd. Shoumen, Bulgaria, 2020).

von Hahn, Walther, „Vagheit bei der Verwendung von Fachsprachen", In: *Fachsprachen. Ein Internationales Handbuch zur Fachsprachenforschung und Terminologiewissenschaft* edited by Lothar Hoffmann, Hartwig Kalverkämper and Herbert Ernst Wiegand, (Berlin 1998), 383–390.