A mixed hermeneutic/ machine aided investigation of reliability early modern historical sources

Cristina Vertan, Walther v. Hahn, Alptug Güney

University of Hamburg

Dimitrie Cantemir was one of the most prominent figures at the end of XVIIth and beginning of the XVIIIth century in the cultural space of Central Eastern Europe. His manuscripts about the history of the Ottoman Empire and his own country (Moldavia), were translated soon after his death in English, German and French and remained most quoted works about these areas until the end of XIXth century (Cantemir 1771), (Cantemir 1745), (Costa 2015). The originals were lost and partially retrieved only in the last fifty years. Although it is generally accepted that the translations deviate from originals, no rigorous investigation was performed until now due to several barriers: geographical distribution of the data, multilingual character of the texts, inaccessibility to sources quoted by the author.

Digital methods can facilitate analysis on the reliability of translations but also of the historical facts claimed by the author (Vertan and v. Hahn 2014). In order to be effective these methods must consider an intrinsic feature of natural language: the ability of producing vague utterances. The project HerCoRe[1]- Hermeneutic and Computer –based approached for investigating reliability, consistency and vagueness in historical sources 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, general recognition),
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:

- 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 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).

For the detection of linguistic vagueness we follow a multilingual approach. We collected initially listed indicators in the three languages involved in the project (Latin, German and Romanian). Based on (Pinkal 1980), (Pinkal 1985) we distinguish between:

- Vague quantifiers, e.g.: some, most of, a few, about, etc.

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

- 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.)

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

In this contribution we will present the general set-up of the system, the annotation framework as well as the compute-based approach for marking linguistic vagueness.

References

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

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

(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)

(Pinkal 1980) Pinkal, Manfred, Semantische Vagheit: Phänomene und Theorien. In: Linguistische Berichte 70. 1980. 1-26. und 72. 1981. 1-26.

(Pinkal 1985), Pinkal, Manfred, 1985 Logik und Lexikon: Die Semantik des Unbestimmten.

(Vertan and v. Hahn 2014) 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.