RANLP 2017

Workshop in conjunction with Recent Advances in Natural Language Processing (RANLP) 2017

Proceedings of the First Workshop on Language technology for Digital Humanities in Central and (South-)Eastern Europe LTDH4CSEE

First Workshop on

Language technology for Digital Humanities in Central and

(South-)Eastern Europe

PROCEEDINGS

Varna, Bulgaria
8 September 2017

ISBN 978-954-452-046-5

Designed and Printed by INCOMA Ltd.
Shoumen, BULGARIA

The workshop was supported partially by the Volkswagen Foundation through the Project HercoRe.

On the annotation of vague expressions: a case study on Romanian historical texts

Anca Dinu
Center for Digital Humanities
University of Bucharest
anca_d_dinu@yahoo.com

Walther v. Hahn,
Research Group "Computerphilology"
University of Hamburg
vhahn@informatik.uni-hamburg.de

Cristina Vertan,
Research Group "Computerphilology"
University of Hamburg
Cristina.vertan@uni-hamburg.de

Abstract

Current approaches in Digital .Humanities tend to ignore a central aspect of any hermeneutic introspection: the intrinsic vagueness of analyzed texts. Especially when dealing with historical documents neglecting vagueness has important implications on the interpretation of the results. In this paper we present current limitation of annotation approaches and describe a current methodology for annotating vagueness for historical Romanian texts.

1 Introduction

Digital Humanities (henceforth „DH") nowadays tend to use huge corpora („big data") to achieve reliable results with computer-based technologies. However, behind all interpretations, such as reliability discussions, stands a hermeneutic approach, which is always qualitative in nature. Such research can be backed up by quantitative descriptions of the material, which is basically the classical annotation method in DH. The scientific use of annotations is usually a positive ascription of features, such as „is reliable" or „is not reliable" and a statistics of the corresponding feature. This kind of approach ignores a fundamental aspect of the data, the vagueness of many assertions and thus the drawbacks of such crisp choice "is/is not".

In this article we describe recent on-going research activities in which we investigate to which extent assertions found in historical texts original texts or in their translations are:

- consistent within the same text and across the originals;
- reliable with respect to author's annotations or the annotations of further translators;
- consistent and reliable across different language versions.

We propose to support the hermeneutic traditional approach through the following computer-based methods:

1. an annotation formalism which allows the mark-up of different types of vagueness and its source;
2. a set of inference rules for the combination of such vague features to calculate an overall result of their reliability;
3. a similarity measurement of the inferred results obtained for the same queries on different translations.

2 Vaguenss theoretical an practical considerations in DH-projects

2.1 Theoretical approaches to vagueness in natural language.

Since a long time vagueness is described in theory (for German: Pinkal 1980) and processed in various conceptual and technical environments (Zadeh 1965). Data in humanities (history, arts, literature, music etc.) are subject to interpretation of the researcher and therefore their possible or real vagueness must be kept for final resolution.

In historical texts - more than in modern texts - many vague expressions are standard for describing events, attitudes or even factual data (v.Hahn 2015). Writing them into a standard data base would distort the entries, because its later processing in inferences will treat them as true facts.

Indefiniteness (Unbestimmtheit) is a central feature of natural language. Any type of text (even the very specialised ones (v. Hahn, 1983)) includes indefinite expressions. According to Pinkal (Pinkal

Proceedings of the LT4DHCSEE in conjunction with RANLP 2017, pages 24–31,
Varna, Bulgaria, 8 September 2017.
http://doi.org/10.26615/978-954-452-046-5_004

1980), (Pinkal 1985) one can evaluate the degree of indefiniteness ("Unbestimmtheit") in an expression according to three criteria

4. Semantic vagueness
5. Unclearness of the illocutive role
6. 3 the communicative expression can be unclear, when from the point of view of the situation and the recipient it is too less informative

An expression is indefinite when only by means of imposing other conditions one can assign to it the value "true" or "false". Such conditions can have a semantic or pragmatic nature. We will refer here strictly at the semantic nature. According to Pinkal, the semantic indefiniteness can be either vague or ambiguous. Vagueness has several sources: either porosity of language or imprecision at expression's borders, inexactity, one or multidimensional relativeness. Ambiguity is due to several natural language features like: Homonymy, polysemy, syntactical ambiguity, multiple referential meaning, and dual metaphorical meaning

Vagueness is related more to the conceptual backbone of the language, while ambiguity relates to words and terms. Vagueness can be preserves across languages, while ambiguity can be present in just one language.

Computer linguistics concentrates often more on ambiguity, by means of resources like Wordnet[1]. Vagueness detection is in strict correlation with conceptual modelling of the text. In the current proposal we will investigate to which extent vagueness influence the hermeneutic interpretation of historical sources. However translation can be often of source of transforming ambiguous expression in source language in vague expressions in the target language, especially if the knowledge base is reduced and the source and the target language belong to different language families.

2.2 Computer-based approaches for historical document analysis

Processing multilingual (historical) texts beyond digital reproduction of paper versions, implies several obligatory steps:

Definition and formal representation of concepts which are relevant for the document(s);

Annotation of these concepts within the documents also by means of automatic processes implying text minning and natural language processing (named entity recognition, lemmatization, PoS tagging and parsing);

definition an implementation of a reasoner to be applied on the annotations;

choice of a query language compatible with the reasoner, i.e. the query language should be powerful enough so that it can exploit the entire inference mechanism of the parser.

There are few projects in digital humanities which employ semantic representation of the data. One of the most prominent example is the CIDOC-CRM Ontology[2] a conceptual reference model for representing cultural heritage objects. Unfortunately the ontology is used mainly for encoding meta-data about the objects, but less for deep annotation of the content.

Another project using semantic relation between objects is Averroes[3], in which a corpus of all editions and translations of the philosopher Averroes are not only reproduced, but aligned by means of RDF-formalism.

The project „Inschriften im Bezugssystem des Raumes"[4] uses also RDF as formalism to represent different topologies of epigraphs and their interconnections.

We should mention also that currently, the PERSEUS[5] project containing the biggest collection of digitizes classical texts, is starting to release the data as LOD (Linked Open Data)[6] (Bridget et. al. 2014).

The projects mentioned above reached a certain degree of completion, and obviously there are some recent activities started. However, they represent a very small percent from the huge number of current digital humanities projects focused mainly on digitization and meta-data cataloguing.

Even the above mentioned projects do not consider a central aspect of humanities and in particular historical data: representation of vagueness. Meanwhile there are strong theoretical developments:

1 https://wordnet.princeton.edu/

2 http://www.cidoc-crm.org/official_release_cidoc.html

3 http://www.thomasinstitut.uni-koeln.de/11610.html?&L=0

4 http://www.digitale-akademie.de/projekte/inschriften-im-bezugssystem-des-raumes-ibr.html

5 http://www.perseus.tufts.edu/hopper/

6 http://linkeddata.org/

- inclusion of a module on certainty and precision in TEI;
- development of an ontology language including vagueness and
- corresponding implementations of reasoners and query languages.

However, to our knowledge, at least at the current moment they are not exploited by any project in digital humanities, although this is the only way to fully support humanists with new interpretations and analysis of their texts.

Manfred Thaller discusses already in (Thaller 1984) how relevant for historical research are the quantitative approaches, and insists for more computer-based formalisms which allows investigations lead by questions like "why fact X happened" (qualitative research) and not "how often fact X happened?" (quantitative research).

In (Thaller 2007) it is stated very clear that digitized texts as they are now realised are: not-ambiguous, context free and contain just the information embedded in the code, whilst historical texts are sequence of symbols, each carrying a meaning, which co-exist in a multidimensional space. These spaces are independent one of the other which makes possible to attach to each of them a metric.

These two works are seminal but with exception of them, digitization of data remains just a way of raw preservation: documents which can be read online. Search is related to words or in best case to words clusters, called wrongly concepts. This makes the computer just a static support for reading or in best case flat search, but does not imply it dynamical in the hermeneutic research. Progress in Computer science in the last years allows a change of paradigm.

2.3 Vagueness and Mark-up annotations

TEI[7] is currently the main standard used for encoding historical texts. The P5-Guidelines follow the XML-mark-up formalism and thus:

do not allow concurrent mark-up and

enable connections between annotated segments through Xlink-like pointers and unique identifiers.

TEI has a modular architecture: there is a core module with elements which can be used in all texts (most dealing with basic text structuring and formatting), as well as more than 20 modules covering almost all fields of textual humanities. The price paid for this broad coverage is an increased complexity in data representation, which triggers also difficulties in the parsing process.

Automatic Text processing tools cannot interpret TEI-Tags, thus these are filtered out, together with useful information contained by their semantics (Piotrowski 2012, pag.66).

TEI offers three possibilities for encoding vagueness:

1. the <note> element: the user can write unstructured text, mentioning the degree and scope of the vague aspect identified);
2. the <certainty> element: this offers the possibility to structure the information about vagueness. 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). It indicates the numerical accuracy associated to some aspects of a text mark-up. If a standard value is precise and known, one can express it with the element <precision> and the attribute @stdDev, which represents its standard deviation.

Additional TEI offers the possibility to indicate the responsible for the whole content or partial annotators. In this way one can specify if the vagueness is due to the author, the quoted source or the editor.

TEI-P5 specifications mention that. "The certainty element allows for indications to be structures with at least as much detail and clarity as appears to be currently required in most ongoing text projects" As mentioned before, TEI is mainly used for encoding text as near as possible to the original and display it. Rarely deep queries are performed on the annotation, also because it is practically impossible to have a general parser, which allows

7 http://www.tei-c.org/Guidelines/P5/

complex queries. TEI parsers are usually dedicated, e.g. deal with the core module and a certain domain.

There are several drawbacks of the TEI approach for annotating vagueness:

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 these sources of vagueness can be specified with the <respons> tag which can be attached just to individuals.

3. <precision> can be specified just for numerical values. An expression "some kilometres south from the city" introduce a non-numerical vague coordinate. When we speak about historical documents, sometimes even geo-location of the place is not possible.

4. There is no reasoner which can be applied to the TEI annotation.

2.4 *2.4. Vagueness in Ontological modelling*

A semantic model for historical data should imply a mapping to at least a domain ontology. OWL[8] (Web Ontology language) is the current standard used for expressing ontological knowledge. One can specify classes, subclasses, properties and subproperties, roles and can relate all these together through logical statements from Description Logic. OWL assertions are specified following the RDF-triple formalism (Subject-Predicate-Object). The OWL was used intensively in the first generation applications of Semantic Web. However, it became obvious that it is a common requirement in real world applications that the system is able to deal with imprecise /vague knowledge, which cannot be modelled with OWL (Bobillo et al 2012).

In order to simulate vague knowledge, methods as Reification[9] or Named Graphs[10] were used. However they have two drawbacks:

- Increase (sometimes dramatically) the number of RDF-triple

- At the end, they rely again on crisp description logic.

The new OWL 2 standard offers the possibility of designing fuzzy Ontologies and realize inferences with fuzzy logic. The principle is to use an OWL 2 ontology, extending its elements with annotation properties representing the features of the fuzzy ontology that OWL 2 cannot directly encode (Bobillo and Straccia 2010). With this formalism one can define vague expressions as fuzzy modifiers and apply them to data-types and concepts. In OWL 2 Concepts can receive also weights.

For example to define the concept (0.8 A + 0.2 B): one creates the atomic "Sum08Aplus02B" and annotate it:

```
Class ( Sum08Aplus02B Annotation
        ( fuzzyLabel
                < fuzzyOwl2
fuzzyType =" concept " >
                < Concept type ="
weightedSum " >
                < Concept type ="
weighted " value ="0.8" base =" A " /
>
                 < Concept type ="
weighted " value ="0.2" base =" B " /
>
  ))
```

For the creation of a fuzzy ontology a Protégé[11] API the Fuzzy Ontology Editor[12] is freely available. The plug-in is generic and not specific to any reasoner. In the next section, we show how this can be pipelined with a reasoner.

2.5 Vagueness and reasoning

Most used reasoner for fuzzy ontologies is the DELOREAN reasoner (Bobillo et al. 2013). The reasoning algorithms within this system, are based on the computation of a crisp ontology that preserves the semantics of the original fuzzy ontology and therefore reasoning with the former is equivalent with the latter. The developers of the DELOREAN reasoner applied successfully the same principle for Zadeh as well as Gödel fuzzy description logic.

8 https://www.w3.org/2001/sw/wiki/OWL
9 https://www.w3.org/DesignIssues/Reify.html
10 https://www.w3.org/2004/03/trix/

11 http://protege.stanford.edu/
12 http://www.umbertostraccia.it/cs/software/FuzzyOWL/

The equivalent crisp ontology is larger than the fuzzy one, as additional axioms have to be added in order to keep the semantics.

The DELOREAN reasoner can be used as a standard application or through the provided API integrated in a larger system.

An important contribution to mathematical modelling of vague data is given in (Schlarb 2008). The definition of certain operators still has to be compared with the formalism offered by fuzzy OWL.

2.6 Automatic processing of historical texts

Language technology (LT) reached a certain maturity during last years. Industrial applications use now LT component for modern languages, like lemmatisers, PoS Tagging, Named entity Recognizers (Vertan and v Hahn 2012). The picture is different when referring to historical languages. Moreover, for one modern language, there are several historical variants and the borders between them are not really clear. Additionally, languages became standardized in the late XVIIth-XIXth century, so there are not clear rules to be encoded. A big problem is also the orthography which was not completely standardized, so many variants may occur for the same word. Without any pre-processing step, no modern language processing tool can be applied to historical variants.

Minimal transformations imply orthographic normalization and in some cases syntactic translation rules (Piotrowski 2012). Less attention is paid to semantics and the conceptual space (thus implicit knowledge) which changed during the years (Vertan 2010) (Vertan and v. Hahn 2014a).

Many historical documents present a document multilinguality: there are words or paragraphs at least in Latin or classical Greek. These paragraphs have to be identified and isolated prior to any other processing.

.

3 Rationale of the corpus

Dimitrie Cantemir (1673-1623) was prince of Moldavia (historical region including regions from current eastern part of Romania, Republic of Moldavia and some parts from Ukraine), man of letters-philosopher, historian, musicologist, linguist, ethnographer and geographer. He received education in classical studies (Greek and Latin in his country of origin), then he lived for several years in Istanbul where he learned Turkish, and familiarized himself with the cultural traditions of the ottomans, meet important persons around the sultan and learned a lot about history of the Empire. After a very short period of being prince of Moldavia he was forced to immigrate to Russia, where he became an important person at the court of Tsar Peter the Great. During this period, his works gained attention in the Western countries. He became member of the Royal Academy in Berlin and, at their request, he produced the two books which are the subject of this 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, the political and administration system. Local denominations and troponins, as well as names are written in Romanian with Latin script as his intention is to demonstrate the Latin origin of his folk. The transcriptions are not standardized and one retrieves for the same troponin several name variations. Quotations as known today are very rare, there is no bibliography. According to (Lemny 2010), as there was practically no consistent previous work about the region, Cantemir himself was not particularly careful with indicating sources of knowledge. The work is accompanied by a map, the first detailed cartography of the region. The names on the map are in Romanian language. The Latin original was translated for the first time in German, and only later at the middle of the XIXth century in Romanian. The Latin manuscript seemed to be lost for a long time, so that the first Romanian translation was following the German one. The German translation is containing editorial notes of the translator.

Historia incrementorum atque decrementorum Aulae Othomanicae, the history of the Ottoman Empire. In contrast with the previous work about Moldavia, here Cantemir indicates very carefully the sources of information. (Lemny 2010) supposes the existence of previous works, known in the western countries, behind this decision. This work was written also at the request of the Academy in Berlin. Cantemir follows the same principle: text in Latin, while the troponins and local denominations are written this time in Ottoman Turkish. Although there were already some previous works about the Ottoman Empire, the novelty of his approach is the quotation of Turkish sources. The reliability of these sources is untrusted sometimes by Cantemir himself. The manuscript reaches the western world

after Cantemir's death, carried by his son to London. Here, a first translation in English is produced: The history of Raise and Decay of the Ottoman Empire. The translator reinterprets the texts, probably also being confused by the presence of Turkish information sources, which were perceived in that time as completely unreliable. The Latin original remains lost for centuries and is rediscovered only at the end of the XXth century in the USA. Thus, the German translation is based on the English one and inherits the same alterations, and presumably adds new ones. The Romanian translations use in contrast the Latin original. The last translation (Costa 2015a) will be used in this proposal.

Until now there is no systematic study on the reliability of the text sources in Cantemir's works, nor the degree of alterations produced by the translations of the two works.

Given the fact that both works became standard reference for western authors until the middle of XIXth century, it is expected that their reception influenced also following historical material. There is no reprint / new edition of his works in German or English. There are however, several reprints of the Romanian versions. Recent Romanian translations of Decriptio Moldaviae are done after the original Latin manuscript.

A lot of works were dedicated to the personality of Dimitrie Cantemir and its perception in different parts of Europe. A study of the reliability and consistency of the historical facts as they are described in originals and their translations is practically impossible to be done only with traditional hermeneutic methods. One needs expertise in the same time in Latin, German, English, Romanian, Turkish, just to enumerate the main languages used in the two books, which additionally sum up to a volume of about 1000 pages. Both German editions are printed in "Fraktur" script, which is nowadays very difficult to be read. A recent digitalization done by the BBAW for the History of the Ottoman Empire13, makes the text more accessible. The digital version is freely available in TEi-P5 format. However, the TEI-P5 concentrates only on a diplomatic transcription and a flat linguistic annotation (lemma and part of speech) and does not touch any aspects of vagueness or reliability of sources.

Cantemir's texts are a real challenge with respect to multilinguality: in Descriptio Moldaviae,

[13]http://www.deutschestextarchiv.de/book/show/cantemir_geschichte_1745

4 Workflow for annotation of vagueness

For the particular corpus presented in section 3 we decided to represent vagueness and other types of uncertainty at least five levels

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

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

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

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

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

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

A clear indicator of vagueness are also named entities like persons and places, especially when they differ in transliteration, spelling within the text or across similar historical sources. Thus the annotation of named entitiees is of central role.

However the unclear person, time, place identification is even more difficult to automatize or at least assist by computer techniques, being more of a matter of hermeneutical research for humanists and historians.

5 Conclusions and further work

Annotation and interpretation of vagueness is a central issue in digital processing of historical texts. However this issue was completely neglected until now, and has as consequence often distorted interpretation of digitized historical texts. In this article we presented the current state of the art on vagueness annotation and introduce the first approached for considering vague expressions as part of the annotation process. Further work concerns the automatic annotation of such expressions, the construction of the ontology and the implementation of the interpretation layer.

Acknowledgements

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

Bibliography

Fernando Bobillo, and Umberto Straccia, 2010, *"Fuzzy Ontology Representation using OWL 2",* http://arxiv.org/pdf/1009.3391.pdf (last retrieved 15.02.2016)

Fernando Bobillo, and Miguel Delgado, and Juan Gomez-Romero, 2013, *"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

Almas Bridget and Alison Babeu, Alison and Anna Krohn, 2014, *„Linked data in the erseus Digital Library",* ISAW Papers 7.3, http://dlib.nyu.edu/awdl/isaw/isaw-papers/7/almas-babeu-krohn/ (last retrieved 15.02.2016)

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

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

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

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

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

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

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

Michael Piotrowski 2012, „Natural Language Processing for Historical Texts“, Morgan & Claypool Publishers, Synthesis Lectures on human Language Technologies.

Sven Schlarb, 2008 *Unscharfe Validierung strukturierten Daten –ein Model auf basis unscharfer Logik*, Köllner Beiträge zu einer geisteswissenschaftlichen Fachinformatik, Band 1,

Manfred Thaller 1984, *„Ungefähre Exaktheit. Theoretische Grundlagen und Praktische Möglichkeiten einer Formulierung historischer Quellen als Produkte ‚unscharfer' Systeme",* in Neue Ansätze in der Geschichtswissenschaft, Conceptus Studien 1, Wien, pp. 77-100

Manfred Thaller 2007 *„Was macht einen Quellentext für die Informatik ‚historisch'?"*, in De litteris, manuscriptis, inscriptionibus … Festschrift zum 65. Geburtstag von Walter Koch, Kölzer, Theo and Friedl, Christian and Vogeler, Georg (Eds.), Wien-Köln-Weimar 2007, pp. 543-557

Cristina Vertan and Walther v. Hahn 2012, "*Multilingual Processing in Eastern and Southern EU Languages: Low-Resourced Technologies and Translation"*. 396 pages. Cambridge Scholars Publishers. Newcastle

Cristina Vertan and Walther v. Hahn 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.

Cristina Vertan and Walther v. Hahn 2014*, Making historical texts accessible to everybody*. Dublin Workshop für Text Simplification. Dublin 2014

Cristina Vertan and Walther v. Hahn 2014, *Das Balkanbild in Deutschland während der letzten 300 Jahre - Eine digitale Plattform zu Analyse und Erschließung multilingualer Dokumente über die Balkanländer und das Osmanische Reich*, In: Proceedings of the Digital Humanities- Germany Conference, 25-28 March 2014, Passau, Germany

Cristina Vertan 2010, *Towards the Integration of Language Tools Within Historical Digital Libraries"*, in Proceedings of the international Conference for Language Ressources and Evaluation LREC 2010, Malta, 19-21May 2010

Cristina Vertan and Walther v. Hahn 2015, *Multilinguality in Historical Texts - Applying Language Technology for Cultural Heritage* , Charles university of Prague. (http://ufal.mff.cuni.cz/events/seminar-35th-anniversary-cooperation-between-charles-university-prague-and-hamburg-university)

Walther v. Hahn 2015 *Preserving Vagueness: The Central Mission of Next Generation Digital Humanities*. Charles university of Prague. (http://ufal.mff.cuni.cz/events/seminar-35th-anniversary-cooperation-between-charles-university-prague-and-hamburg-university