Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

UNIVERSITATEA DIN BUCUREȘTI
VIRTUTE ET SAPIENTIA

Research Center „Digital Humanities"

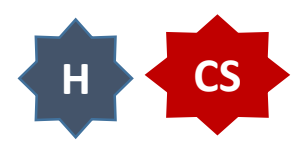Research Group „Computerphilology"

HerCoRe – Hermeneutic and Computer based Analysis of Reliability, Consistency and Vagueness in historical texts
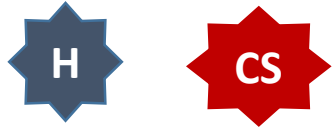
- Illustrated through two main works of Dimitrie Cantemir-

April 2017 –March 2020

CS

Dr. Cristina Vertan, UHH
Project coordination,
DH, CL, CS

H **CS**

Dr. Anca Dinu, UB
Team Leader UB,
Linguistics,CL

H **CS**

Prof. Dr. Walther v. Hahn,
UHH
Vagueness, CL, DH
German Linguistics,

H

Prof. Dr. Ioana Costa, UB
Cantemir Translations,
Classical philology

H

Prof. Dr. Yavuz Köse, UHH
Turcology

CS

Prof. Dr. Liviu Dinu, UB
Fuzzy Logic, CL, CS

H

Alptug Güney, UHH
Turcology

CS

Segiu Nisioi, UB
CL, CS

N.N.   Software developer **CS**

Combine hermeneutic approaches and methods from computer science for investigating reliability and consistency of original text from 18th century as well as their translation.s

CS

Demonstrate how to include vagueness and imprecision in annotations and interpretations engines

Progress work in automatic recognition of vague expressions

H

Compare for the first time original with translations done in the 18th,- 19th century

(In)Validate assumptions about source quotations in original text

Outline

- Rationale of the project
- Corpus' Insight
- Technological challenges
- HerCoRe Approach
- Position within „Mixed Methods in Humanities"

Dimitrie Cantemir (1673 -1723)

- Prince of Moldavia (historical province) as well as „universal" humanist (linguist, ethnographer, musicologist, historian, writer)

- As member of the Royal Academy in Berlin and at the request of this institution wrote two works :
  - Description of his own country („Descriptio Moldaviae")
  - History of ottoman empire (History of Growth and Decay of Ottoman Empire)

- Original material written in Latin; Both originals were lost already by the end of 18th century

- Several copies were used as basis for translations into German, English, French, Russian and later in Romanian

- Sometimes the translation relies on other translation (e.g. first Romanian translation of "Descriptio Moldaviae" was done after the German version from 1774.

These translations used as reference information about the Ottoman Empire and Romanian provinces until the middle of 19th century, i.e. they give an idea about the reception about this part of the world in Western Europe.

Analysis and interpretation of Cantemir's works

- Already in the 1920'ies, it was demonstrated using 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 XX century researchers claimed that some of the sources, persons and facts quoted by Cantemir were not existing

- BUT given the:
  - Geographic distribution of material (originals in libraries in USA and Russia; translations and copies 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 volume of data which has to be processed in parallel

no study about the reliability and consistency of the original and the translations could be performed until now

Computational methods could help in performing this study

State-of-the-art - Studies on Cantemir works

- The reception of Cantemir's works vary significantly depending of époque and geographical perspective of researchers

- (Lemny 2010) considers that most od the vague or imprecise assertions in Cantemir's texts were due to lack of previous work about the described regions.

- Franz Babinger in "Aufsätze und Abhandlungen zur Geschichte Südeuropas und der Levante" (1966) argues that some of the sources quoted by Cantemir did not even exist. However meanwhile new documents , especially in the Turkey were made available, so Babinger's assertions could be no longer true.

- In "Dimitrie Cantemir's Ottoman History and its Reception in England" (2010) Hugh-Trevor Cooper argues that not the text quality and accuracy made the work so well-known but the political connections of his son, who represented the Russian empire in London. The argumentation is supported by nomination of other works, claimed to be more accurate than Cantemir's.

- An opposite interpretation is done by M. Leezenberg in "the oriental origins of orientalism: the case of Dimitrie Cantemir" (2012) who insists on the Turkish sources to which no other western contemporary of Cantemir had access.

Directions of investigation

- **Reliability:**
    - Of the original: are the quotations made by Cantemir grounded? Is there a 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?)
    - Of the translation against the original; Here an important role have the inserted editorial annotations.

- **Consistency:**
    - Within the original: keeps Cantemir a constant opinion about persons, events, facts across the text? (see his own annex with annotations vs. the text)
    - Across the 2 originals: Are common persons and events described similarly?
    - Between original and translation: does the translation preserve the degree of vagueness /certainty stated by Cantemir?

- **Vagueness**
    - Political or tactical reasons for imprecise expressions

Manuscripts, editions, translations

Historia Othmanica

- The manuscript became available as *facsimile* in 1999; the first scientific edition of the Latin text was published in 2001 (Dan Sluşanschi) and was followed by two complete translations, in 2011 (Dan Sluşanschi) and 2015 (Ioana Costa).

- The English translation was published in 1735, by Nicholas Tindal; it remained for more than two and a half centuries the only direct descendant of the Latin text. Tindal's History of the Othman Empire was used in the following decades as a basis for new translations: in French (1743) and German (1745, Hamburg). Later on, this German translation of the English one was used to accomplish the first Romanian translation (1876-1878, Iosif Hodoş).

Descriptio Moldaviae

- For this work there are three manuscripts:

- The work was for the first time present in the Romanian culture in 1771, as a translation of the German translation. The first scientific Romanian translation was published in 1973 by Gheorghe Guţu, who edited the Latin text using the manuscripts A and B. In 2006 was published the first translation (Dan Sluşanschi) that incorporates the testimony of ms. C. It will soon be published a new translation (Ioana Costa), as part of the Cantemir's *Opera omnia*, meant to celebrate the tricentenary of his death (2023).

Author versus Source

- Dimitrie Cantemir composed his scientific works in close relationship with the available sources. The references are always assessed, either as correct or plausible perspectives or as errors to which he reacts stating his own opinion or simply by mentioning the error, leaving the reader to outline his personal point of view.

- His position toward the sources is explicit (quotation, comment) or suggested especially when he rejects them or expresses a certain reserve. This detachment from the sources he uses is materialized in the lexicon (nouns and verbs, nevertheless adjectives or adverbs meant to determine the basic terms) and/or syntactical means, *id est* the use of verbal forms that point to non-reality (potential or unreal nuance, expressed by the subjunctive mode: present/perfect tense or imperfect/pluperfect tense).

- Considering the fact that Latin syntax frequently involves non-real verbal forms (subjunctive mode) to point to an opinion that does not belong to the author (free indirect speech), we have to accept that the author's position toward the content of an excerpt that is reproduced in free indirect speech or simply restated (paraphrased) is not always obvious.

Lexical and Syntactic Sources of Vagueness

Quotation

More plausible

Would have been…

seems unlikely

equaly false

- Hactenus Gregoras: ad cuius verba observare haud extra propositum erit τὴν πρώτην, quam Gregoras vocat „**Tartaria**[...] esse, quam hodie vulgo „Ma[...] eiusque incolarum nomina, [...]storicis recenseantur, tamen adscita magis, aut ab exteris indita, quam propria eisque, dum in suis sedibus m[...] peculiaria fuisse. Ita, si quis [...]oni Praefatione legerit O[...] [...]cipes in duas stirpes fuisse divisos, „Aliothman" unam, et „Ali Dzengiz"[1] alteram, n[...] [...]ub ipsis horum generum condi[...] appellationem iam apud[...] [...]aluisse. Vti enim absonum videtur, Aliothmanos Suleimano parentes ab huius ne[...] integro post saeculo iis imperab[...] fuisse sortitos; ita non minus falso vulgo praedicantur Tartarorum Crimensium Principes ab ipso Dzengizchano „**Alidzengiz**" appellationem retinuisse.

- Până aici l-am citat pe Gregoras: faţă de cuvintele lui nu va fi nepotrivit să observăm că acea Tartaria „ἡ πρώτη", pe care o numeşte Gregoras, este chiar aceea pe care o numim îndeobşte cea „Mare", iar numele locuitorilor ei, chiar dacă sunt înregistrate de istorici, au fost totuşi mai degrabă împrumutate sau date de străini decât proprii lor, purtate întocmai pe vremea când se aflau în sălaşurile lor. Astfel, dacă va fi citit cineva în Prefaţa pusă înaintea acestui tratat că principii neamului oguzilor au fost împărţiţi în două stirpe, una „aliothmană", cealaltă, „alidzengiză", să nu creadă că denumirea aceasta era de-acum valabilă pentru întemeietorii acestor neamuri. Căci, după cum pare nepotrivit ca aliothmanizii care i se supun lui Suleiman să-şi fi ales numele de la nepotul acestuia, care a domnit peste ei după un secol întreg, la fel de fals se spune îndeobşte că principii tartarilor din Crimea şi-ar fi păstrat denumirea „alidzengiz" chiar de la Dzengizchan

Historia Othmanica, **I.1.k : Latin original text and Romanian translation**

Limitations of current computational approaches used by DH for such study

| **Digitalization** | → | **(Linguistic) Processing of the text** | → | **Annotation** | → | **Inferring the results** |

State of the art

- OCR or
- Double keying
- Diplomatic transcription (TEI)
- Transcription variants are recorded just for digital editions

Existent (blackbox) Pipelines:
Tokenization-> Sentence Recognizer-> Lemmatizer->PoS Annotation are applied

- Annotation of **crisp** facts (or some uncertainty marked with TEI mechanisms)
- If at all, one source of imprecision is marked
- Links to ontologies

If at all, **DL crisp Logic**, i.e. even if uncertainty is indicated it is not included in the inference process

Limitations

- OCR with poor performance (Fraktur script, multiling. text.)
- Transcription considered as an interpretation level: Annotation of variants is a must.

- Existent pipelines are not working (too many unknown words)
- One needs to decide first which PoS, chunks have to be identified.

Annotation for vague facts / assertions at several levels are required (and beyond TEI standards)

DL crisp logic will reduce to 0 any fuzzy annotation -> need of a fuzzy inference mechanism

Digital Resources available

- Cantemir's work about the history of ottoman empire:
  - German translation was digitized and annotated with PoS information by German Text Archive (DTA) at the Berlin Academy, and is available for academic purposes
  - Romanian and Latin Versions are in electronically form available for purposes of the project

- Cantemir's work "Description of Moldavia":
  - German version is partially digitized. Digitization and annotation will be completed during the project
  - Latin version is available as Wikisource, will be annotated during the project
  - Romanian Version is digitized, and annotated with PoS, and NE associated to a crisp ontology. The digitization and annotation was performed within a previous joint project between the University of Hamburg and University of Bucharest

- Shallow annotation tools for German, Romanian and Latin are as open source available.

- For German we will rely on the CDG-System developed at the university of Hamburg, which is robust to lexical gaps and syntactical variances; thus sentences with untranslated words can be processed easily

Corpus creation – challenges

- Surface form – level
  - German texts are in black-letter typeface
  - [Part of the foreign words are...]
  - [black-letter typeface]
  - Mixed typefaces
  - Mixed scripts

Higher error rate in OCR (even on relatively homogenous pages up to 25%)

Workflow

Original Raw Text

Transcription Variants

Cantemir Annot. LAT

Cantemir Annot. DE

Cantemir Annot. RO

Annot. Vague Expr. LAT → NLP Tools → Annot. Vague Expr. LAT

German Tr.

Annot. Vague Expr. DE → NLP Tools → Annot. Vague Expr. DE

Editor Comm. DE

Annot. Vagueness Ed.. DE

Romanian Tr.

Annot. Vague Expr. RO → NLP Tools → Annot. Vague Expr. RO

Editor Comm. RO

Annot. Vagueness Ed. RO

Fuzzy Reasoner

Extracted Results

Hermeneutic Interpret.

First steps: annotating vague expressions -1-

- To annotate vague expressions, we first need to (semi-automatically) identify them.

- Since there are two types of clues for uncertainty and vagueness (explicit lexical-semantical clues and implicit syntactic clues), two separate strategies are needed.

- The explicit lexical clues are words or expressions such as *some, most of, a few, probably, possibly, somewhere*, etc. and verbs like *to believe, think, prefer*, etc.

- The implicit syntactic clues are 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.)

M.Pinkal's Schema of Semantic Vagueness

Illocutive Unclarity

Referential Uncertainty

Semantic Vagueness

Communicative Under specification

Vagueness in a narrow sense

Ambiguity

Porosity

Homonymy

Polysemy

Syntactic Ambiguity

Borderline Uncertainty

Inexactness

Referential Ambiguity

Elliptical Ambiguity

Relativity

Metaphorical Ambiguity

one-dimen-sional

many-dimensional
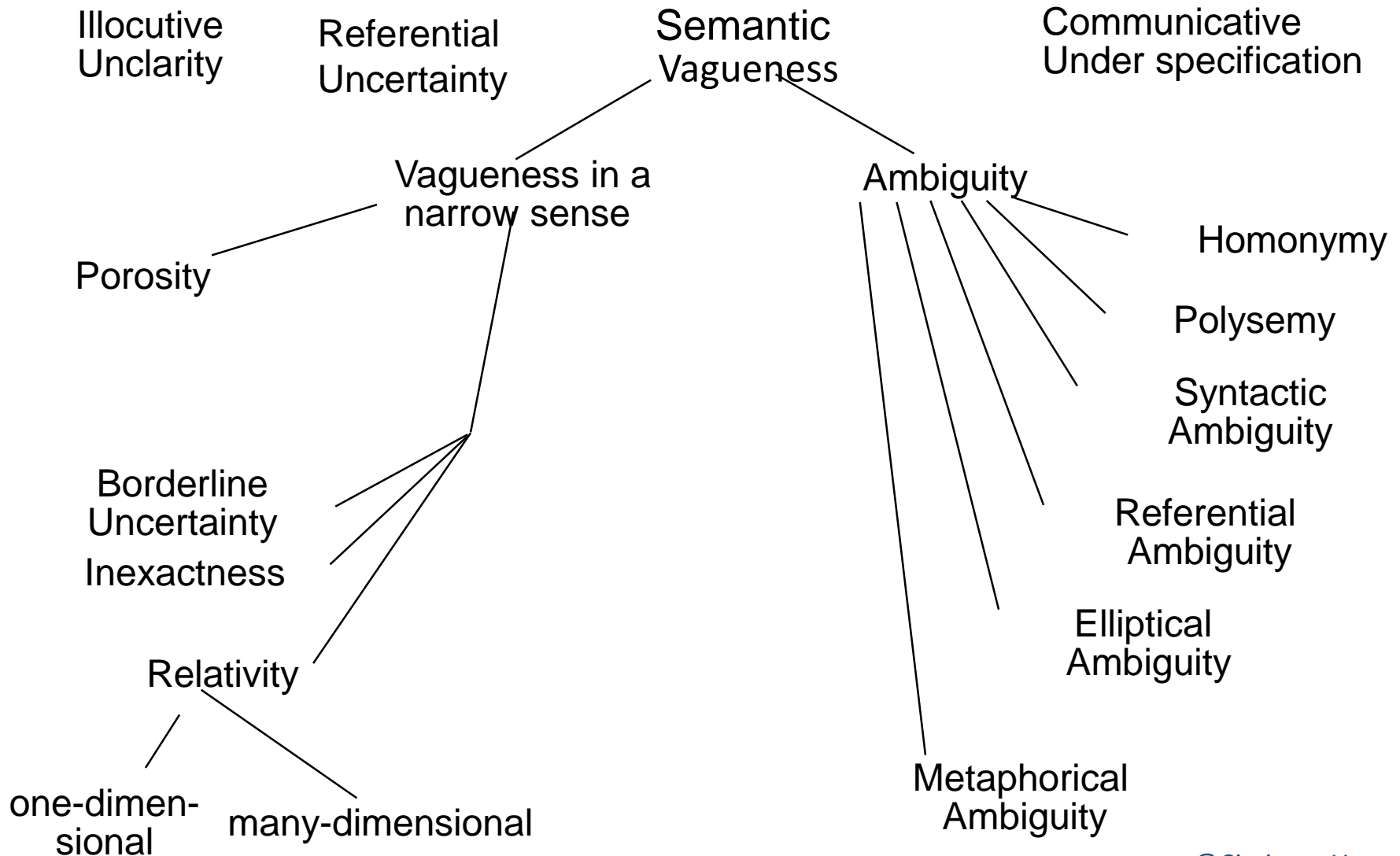
Sources of vagueness in German Texts (selection)

- Not specific (too much degree of abstraction) „Gegend"
- comparative   „mehr", „größer", „älter"
- Not exact adjectives: „eckig", „rund", „feindlich", „befreundet", „nah", „fern"
- Relative adjectives: „groß", „klein"
- Not exact numbers   „4 Tagereisen", „10 Fuß",
- Adverbs   "sehr", "ziemlich",, „etwa", "einigermaßen", „gegen"
- Homonyms   "Schloss", „Klage"
- Polysemy   "Gericht"
- Genitive-Ambiguity   "Der Vorwurf des Gerichts"
- PP-Attachment   „Der Vorwurf des Königs über seine Eltern"
- Deictic Ambiguity   "Wie vorher gesagt", „hier", „dort", „in diesem Buch",
- Elliptical ambiguity   „ demnächst stärker"

- unclear facts   „auf Befehl des Sultans"
- unclear time   „In grauer Vorzeit", „im 14.Jhd."
- unclear person   „der damalige Fürst"
- unclear action   „Unterwerfung der Barbaren"

First steps: annotating vague expressions -2

- 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 from both texts
  - Pre-processing the texts: chunking, lemmatizing, PoS tagging
  - Automatically finding and marking all the (inflected forms of) explicit vagueness terms
  - Manually checking the marking for a short part of text for evaluation (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.

Der erste demnach, der nach Batia Einfall (*) der Moldau ihren vorigen Glanz wieder verschafft hat, war

1. Dragosch. Obgleich unsre Jahrbücher sein Geschlechtsregister nicht angeben, so ist es doch eine beständige Sage bey uns, daß er aus dem alten königlichen moldauischen Stamme gewesen sey, und den Bogdan zum Vater gehabt habe, welcher ein Sohn des Johannis war, von welchem alle Fürsten den Namen Johannis in ihrem Titel zu führen pflegen; dieser Meinung ist desto mehr Glauben beyzumessen, weil man schwerlich glauben kan, daß einer von gemeiner Herkunft mit einem so großen Gefolge auf die Jagd (welche die Moldau zu entdecken Gelegenheit gegeben,) habe ausgehen, und seine übrigen Landsleute überreden können, ihn zu folgen.

(*) Diese Stelle bestätiget aufs neue, was wir oben schon vermuthet hatten, daß Dragosch erst nach des Tatars Bathy oder Batu Einfall, d.i. ungefähr nach 1250. aus Siebenbürgen ausgewandert ist; vielleicht aber lassen sích beede Meynungen vereinigen, wenn man zwey Auswanderungen annimmt, die eine in der lezten Hälfte des Zwölften, die andere in der ersten Hälfte des dreyzehenten Jahrhunderts (V.)

Dimitrie Cantemir, "Beschreibung der Moldau",
Faksimiledruck der Originalausgabe von 1771, Kriterion
Verlag Bukarest, 1973, pag. 111

Der erste demnach, der nach Batia Einfall (*) der Moldau ihren vorigen Glanz wieder verschafft hat, war

1. Dragosch. Obgleich unsre Jahrbücher sein Geschlechtsregister nicht angeben, so ist es doch eine beständige Sage bey uns, daß er aus dem alten königlichen moldauischen Stamme gewesen sey, und den Bogdan zum Vater gehabt habe, welcher ein Sohn des Johannis war, von welchem alle Fürsten den Namen Johannis in ihrem Titel zu führen pflegen; dieser Meinung ist desto mehr Glauben beyzumessen, weil man schwerlich glauben kan, daß einer von gemeiner Herkunft mit einem so großen Gefolge auf die Jagd (welche die Moldau zu entdecken Gelegenheit gegeben,) habe ausgehen, und seine übrigen Landsleute überreden können, ihn zu folgen.

(*) Diese Stelle bestätiget aufs neue, was wir oben schon vermuthet hatten, daß Dragosch erst nach des Tatars Bathy oder Batu Einfall, d.i. ungefähr nach 1250. aus Siebenbürgen ausgewandert ist; vielleicht aber lassen sích beede Meynungen vereinigen wenn man zwey Auswanderungen annimmt, die eine in der letzten Hälfte des Zwölften, die andere in der ersten Hälfte des dreyzehenten Jahrhunderts (V.)

Classical Markup based on crisp ontology

Dragosch	belongs	moldavian kings
Dragosch	son_of	Bogdan
Bogdan	son_of	Johannis
Dragosch	has_additional_name	Johannis
Bogdan	has_additional_name	Johannis
Drgaosch	discovered	Moldau
Dragosch	has_acitivity	hunting
Dragosch	has_activity	development
Development	takes_place after	Batia invasion

Dragosch	has_activity	moved
Movement	takes_place	2nd half 12th century
Movement	takes_place	1st half 13th century
Bathy invasion	takes_place	1250
Bathy	has_alternative_name	Batu
Bathy	is_a	Tatar

Der erste demnach, der nach Batia Einfall (*) der Moldau ihren vorigen Glanz wieder verschafft hat, war

1. Dragosch. Obgleich unsre Jahrbücher sein Geschlechtsregister nicht angeben, so ist es doch eine beständige Sage bey uns, daß er aus dem alten königlichen moldauischen Stamme **gewesen sey,** und den Bogdan zum Vater gehabt habe, welcher ein Sohn des Johannis war, von welchem alle Fürsten den Namen Johannis in ihrem Titel zu führen pflegen; dieser Meinung ist desto mehr Glauben beyzumessen, weil man schwerlich glauben kan, daß einer von gemeiner Herkunft mit einem so großen Gefolge auf die Jagd (welche die Moldau zu entdecken Gelegenheit gegeben,) habe ausgeben, und seine übrigen Landsleute überreden können, ihn zu folgen.

(*) Diese Stelle bestätiget aufs neue, was wir oben schon vermuthet hatten, daß Dragosch erst nach des Tatars Bathy oder Batu Einfall, d.i. **ungefähr** nach 1250. aus Siebenbürgen ausgewandert ist; vielleicht aber lassen sích beede Meynungen vereinigen wenn man zwey Auswanderungen annimmt, die eine in **der letzten Hälfte des Zwölften**, die andere in der **ersten Hälfte des dreyzehenten Jahrhundert** (V.)

Enriched Classical Markup based on crisp ontology

Dragosch	≈ belongs	moldavian kings
Dragosch	≈ son_of	Bogdan
Bogdan	≈ son_of	Johannis
Dragosch	≈ has_additional_name	Johannis
Bogdan	≈ has_additional_name	Johannis
Drgaosch	discovered	Moldau
Dragosch	has_acitivity	hunting
Dragosch	has_activity	development
Development	takes_place after	Batia invasion

Dragosch	has_activity	moved
Movement	takes_place	>=1150; <=1200
Movement	takes_place	>=1200; <=1250
Bathy invasion	takes_place	≈ 1250
Bathy	has_alternative_name	Batu
Bathy	is_a	Tatar

HerCoRe - Multiple level fuzzy annotation at document level

Der erste demnach, der nach Batia Einfall (*) der Moldau ihren vorigen Glanz wieder verschafft hat, war

1. Dragosch. Obgleich unsre Jahrbücher sein Geschle... ...es doc... ...den... ...gew... ...hab... ...wel... ...ihre... desto mehr Glauben beyzumessen, weil man schwerlich glauben kann, daß einer von gemeiner Herkunft mit einem so großen Ge... auf die Jagd (welche die Moldau zu entdeck... Gelegenheit gegeben,) habe ausgehen, und seine ü... ...hen zu fo... ...uthet hatten, daß Dragosch erst nach ...athy oder Batu Einfall, D.i. ungefähr ...us Siebenbürgen ausgewandert ist; ...er lassen sích beide Rechnungen ...wenn man zwey Auswanderungen ...e eine in der letzten Hälfte des ...e andere in der ersten Hälfte des ...Jahrhundert (B.)

Legend

Dragosch	most_probable belongs	moldavian kings
Dragosch	discovered	Moldau
Dragosch	has_acitivity	hunting
Dragosch	has_activity	development
Development	takes_place	after Batia invasion

Cantemir

Dragosch	$\approx$ son_of	Bogdan
Bogdan	$\approx$ son_of	Johannis
Dragosch	$\approx$ has_additional_name	Johannis
Bogdan	$\approx$ has_additional_name	Johannis

Translator

Dragosch	has_activity	moved
Movement	takes_place	$> \approx 1150; < \approx 1200$
Movement	takes_place	$> \approx 1200; < \approx 1250$
Bathy invasion	takes_place	≈ 1250
Bathy	has_alternative_name	Batu
Bathy	is_a	Tatar

Innovative Research

- Humanities : possibility of compare easily multiple multilingual sources; re-specification of hermeneutic circle
- CS: reasoning on data annotated at different vagueness layers. This type of data does not occur in traditional CS applications
- Innovative joint work by CS and Humanities Scientists

HerCorRE Approach

Traditional Approach

- HR: List of requirements (natural language)

- CSR: System specification

- CSR: System implementation

- HSR: Use of the system;

- CSR: error correction; light improvements of the system

- STEP I:
  - HR- text introspection and extraction of indicators for vague expressions
  - CSR+HR: system specification
  - CSR: shallow annotation of text

- STEP II:
  - CSR + HR : Fuzzy Ontology and Formalization
  - HR: annotation and first evaluation
  - CSR: ontology and system improvement

Generalization / Development for other research questions, languages ?

- The pilot software application will be developed modular and parts, like extraction of vague expressions, annotation interface will be reusable for other scenarios

- Ontology and vagueness layer are domain dependent, and have to be re-specified for each application

- Best practice documents containing description considered vagueness levels, ontology development will be provided.

- During the project contacts with other disciplines will be realized and explicit adaptation strategies will be discussed.

HerCorRe aims to (re)-integrate part of hermeneutic research into DH. By means of Computer Methods we will model vague facts and assertions, present them to the user, and let the researcher to draw conclusions and lift one or another veil of history

In Digital Humanities, is there any space left for the hermeneutic perspective?

YES: HerCoRe will provide an automatic identification of possible candidates for „Crux philologorum"

„Crux philologorum" = question mark regarding the correctness of the content/form, raised in the editorial process when the original is missing and based on background knowledge and experience)

Automatic methods can replace partially the background knowledge and support inquiring faster big amount of data and overcome the language barrier. Experience and Interpretation is still a matter of hermeneutic research
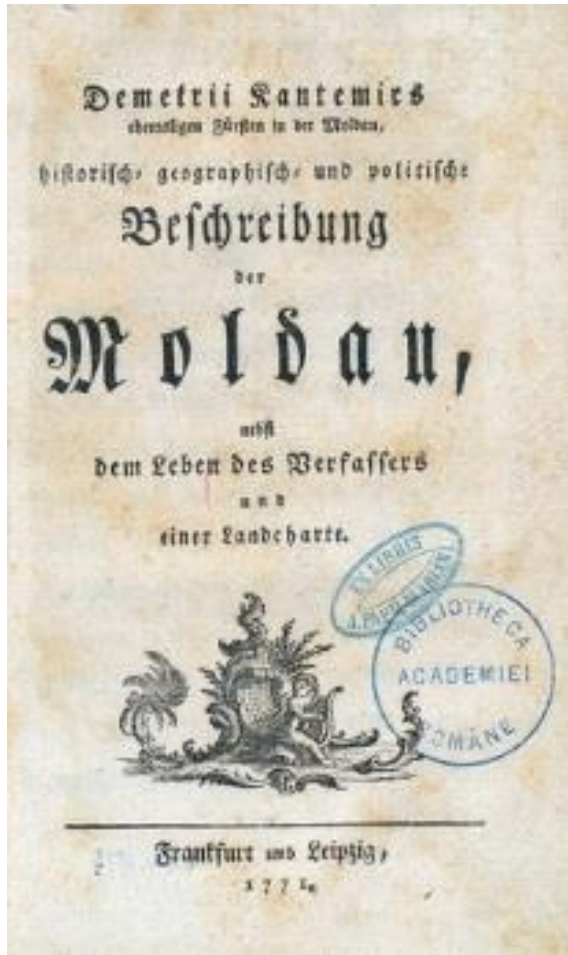
„I do not dare to decide what is the truth about this matter, given the high darkness of this story"

(Dimitrie Cantemir, History of Moldavia, 1752)

We would like to thank the Volkswagen Stiftung for funding not only our research but also an international project !

HerCoRe was already the motor for creating the **„Center for DH"** at the University of Bucharest (the 2nd one in Romania)

HerCoRe will be presented at:

- **DCH 2017** – Interdisciplinary Conference on Digital Culture Heritage, Staatsuniversität Berlin, 30 Aug. -1 Sept. 2017, Berlin

- **RANLP 2017**- Recent Advances in Natural Language Processing as part of the workshop **„Digital Humanities in Central and Eastern Europe"**, 4-8 Sept. 2017