Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

UNIVERSITATEA DIN BUCUREȘTI
VIRTUTE ET SAPIENTIA

DEPARTMENT INFORMATIK
NATÜRLICHSPRACHLICHE SYSTEME (NATS) AB.
„Computerphilologie“ -Arbeitsstelle

Dimitrie Cantemir‘s historical heritage: a digital analysis of his international discourse and quotation style

Ioana Costa, Alptuğ Güney, Walther v. Hahn, Cristina Vertan
{alptug.gueney,cristina.Vertan}@uni-hamburg.de
vhahn@informatik.uni-hamburg.de
ioana.m.costa@gmail.com

Outline

- Computer Science methods and analysis of historical texts (W. v. Hahn)
- Pilot project on Dimitrie Cantemir's works (C. Vertan)
- Research questions from the humanities (A. Güney)
- From Text to digital Texts, challenges (I. Costa)
- Technological approach (C. Vertan)

DH Chalenges

- Preservation of vagueness and uncertainty,
- Semantic analysis based on ontologies vs. individual languages and
- Preservation of intermediary results and features,
- Sustainability of research and results,
- Comparison of geographically distributes objects,
- Support for historical text variation,
- Analysis and fusion of multilingual historical background information.

Preservation of vagueness and uncertainty

If your DH aim is to analyse historical subjects, a proper treatment of vagueness and uncertainty is mandatory, because in the past most data are uncertain:

"Da aber nachher dieses Volk seinen durch den tapfern Nerva Trajan überwundenen König Decebalus verlohr, und theils vertilget, theils hin und her zerstreuet ward, so wurde das ganze Land, das sie bewohnet hatten, in eine römische Provinz verwandelt,… " (Descr.Mold 1,1)

după aceea

Curajosul Nerva Traian

Intreaga țară

În parte

Even if the data are "exact" *(Moldavia became Roman province on Oct.3rd. 155*) the text is vague because of specific adjectives, adverbs or other phrases

However even if you analyse every day's news or newspapers:

"No 10 tries to keep Sturgeon off televised Brexit debate."

Only if you know the continuation:

"Downing Street is trying to exclude Nicola Sturgeon and other political leaders from TV debates on Brexit." (The Times, Nov.28th 2018)

some vague expressions become exact, others more vague than before.

Ontologies vs individual languages

- Still in the project DBR-Mat (German-Bulgarian-Romanian Machine aided translation) in the 1980ies applied an ontology-based semantics of three languages and their formal fusion with feature intersections.
- Todays attempts with Wordnets synses, but without ontology and onlz by translations from American English, fall back behind DBR-Mat.
- Meanwhile ontologies are acknowledged as the most powerful means of fusion for multilingual cultural heritage objects.
- As a prerequisite you need only a PoS-tagger and a morphological analyser.

Preservation of intermediary results and features

- In a process sequence or even more in a program pipe line of any type you must assure that no intermediate result is destroyed by follow-up processes, because it means an overall loss of knowledge.
- Even during the hermeneutic analysis of cultural items no positive information (factual, linguistic) detail nor any negative feature (vagueness, ambiguity, polysemy, controversy) may get lost.
- As a consequence, at least a thorough list of intermediary findings should be available with any result, at least the user interface must keep track of any vestige from earlier steps of processing or earlier knowledge states.
- You may need an intelligent summarizer for this sort of tracking.

Support for historical text variation

- DH research in historical fields has a central difficulty, which shows up in two forms:
- Historical documents are different from today's documents:
  - the grammar changed,
  - the lexicon entries changed their meanings, words were added or became obsolete,
  - Missing parts make texts incomprehensible,
  - the writing or the script changed (like in Romania),
  - the language changed (like in Norway), most handwritten texts are difficult to read or use a different script (like black letter in German).
- For the analysis of texts there is no suitable corpus for comparison or statistical processing.:
- Tools from computationa linguistics (tagger, parsers, etc) do not work for modern languages

Multilingual historical background information

- To analyse a historical situation we need a lot of contemporary material, in the best case material from the time before as well as material from the following time. Otherwise we cannot evaluate the importance of a person, an activity, or a text.
- E.g. in Dimitrie Cantemir's case the attitude of the Ottoman empire towards Russia, of Moldavia towards the Ottoman empire, the attitude of Germans and Englishmen towards Moldavia and the Ottoman empire.
- E.g. An important part in the puzzle is Dimitrie Cantemir's way of assimilating the Ottoman history, his memory of literature in Constantinopel, his own notes, and comments of his Moskow friends.
- For this one may need a quite detailed knowledge base, including the semantic conceptual space of the time, sources of quotation, persons, places usw.

HerCoRe – Hermeneutic and Computer based Analysis of Reliability, Consistency and Vagueness in historical texts

- Illustrated through two main works of Dimitrie Cantemir-

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

UNIVERSITATEA DIN BUCUREȘTI
VIRTUTE ET SAPIENTIA

Funded by
VolkswagenStiftung

April 2017 –March 2020

„Mixed Methods in Humanities“

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

H

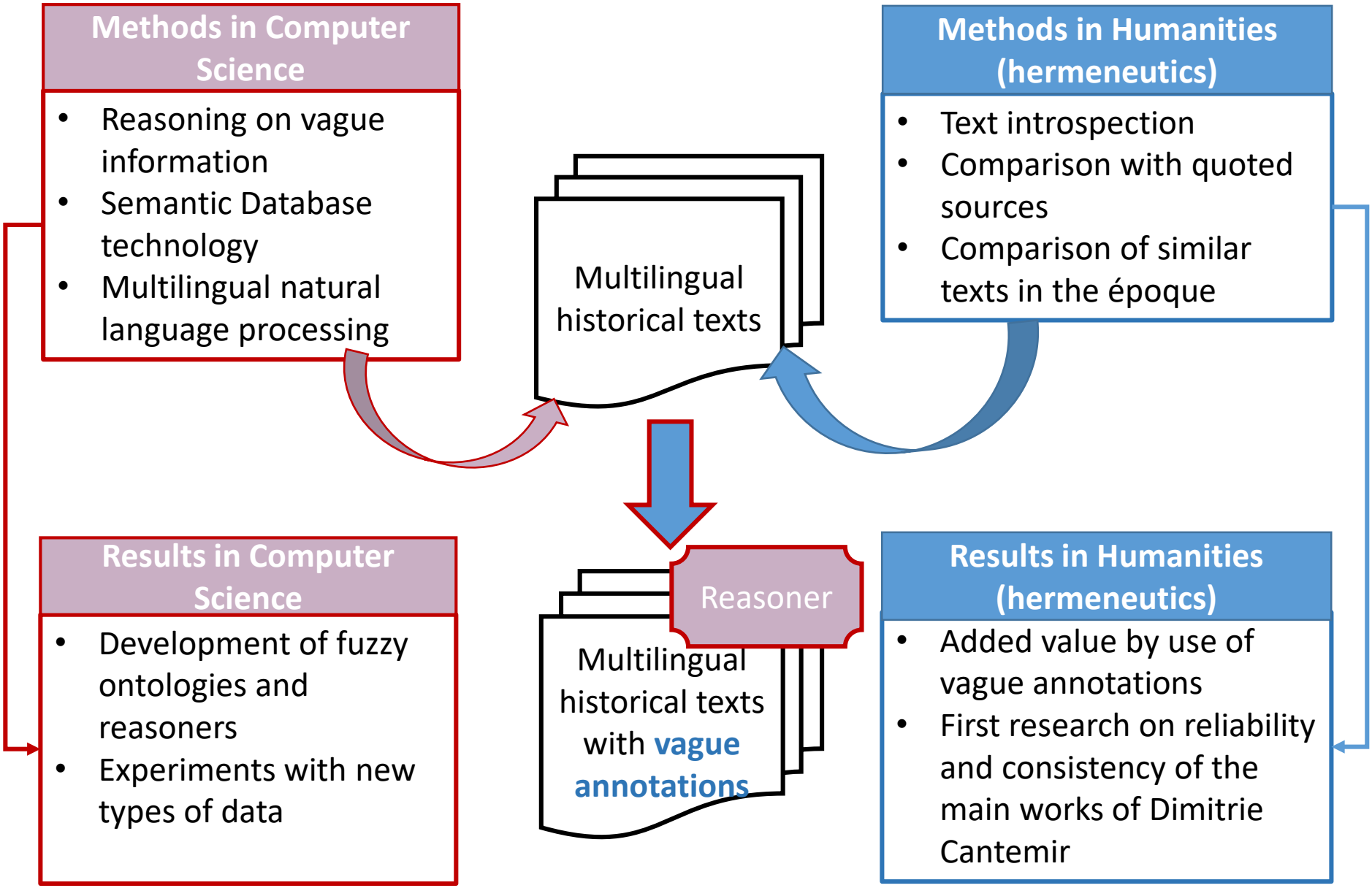
Compare for the first time “original” with translations done in the 18th- 19th century

(In)Validate assumptions about source quotations in original text

CS

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

Progress work in automatic recognition of vague expressions


Methods in Computer Science
• Reasoning on vague information
• Semantic Database technology
• Multilingual natural language processing
Multilingual historical texts
Methods in Humanities (hermeneutics)
• Text introspection
• Comparison with quoted sources
• Comparison of similar texts in the époque
Reasoner
Multilingual historical texts with vague annotations
Results in Computer Science
• Development of fuzzy ontologies and reasoners
• Experiments with new types of data
Results in Humanities (hermeneutics)
• Added value by use of vague annotations
• First research on reliability and consistency of the main works of Dimitrie Cantemir

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

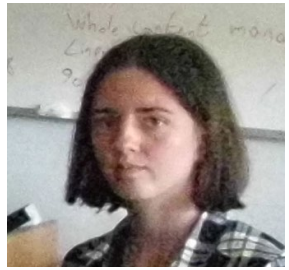
Prof. Dr. Yavuz Köse, UHH
Turcology

CS

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

H

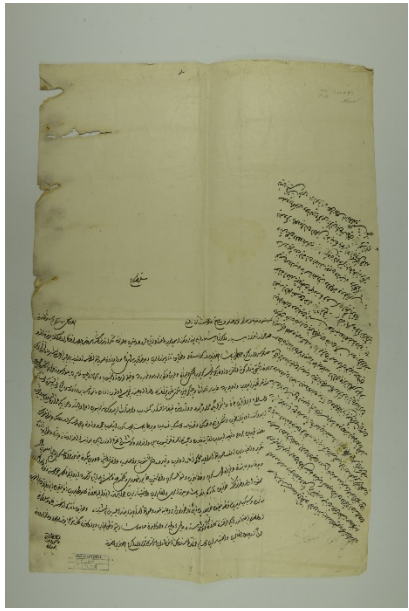
Alptug Güney, UHH
Turcology



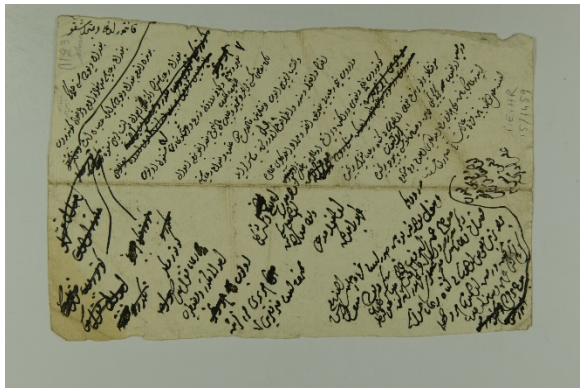
CS

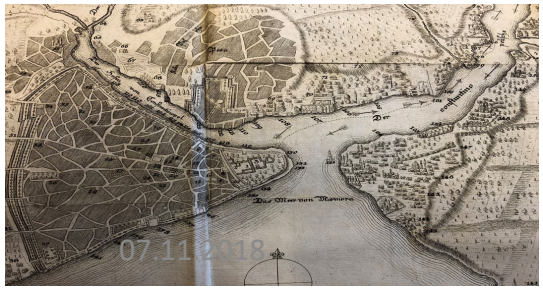
Alina Ciobanu, UB
CL, CS

Research on Cantemir and Evaluation of the Historia Othmanica



- Research in Ottoman and Patriarchal archives in Istanbul on Cantemir
- Study on Cantemir‘s life in Istanbul to complete the missing parts of his biography
- Determining the Turkish sources of Cantemir
- Comparatively study of these sources and of the HO
- Determining the translation failures or missing parts in the translations
- Examination of reliability of the text
- Finding the vague parts in narration





Historical Dissonance: Battle of Ankara or Bursa?

- «Most Christian writers tells us that this battle was fought upon the banks of the Euphrates. But the unanimous confent of all the Turkish writers that Tamerlane immediately after the Battle enter'd Prusa, the metropolis of Bithynia, clearly proves it to have been fought in the plains of that place.» s. 54, English Edition, 1734

Hoca Saadeddin Efendi (1536/7-1599), Tac üt-Tevarih (1520?): Timur bu tutumu öğrenince, Kayseri yolu açık ve geniş bulunduğundan, bu yönden Ankara'ya yürümeyi öngördü. […] Bu sırada Yıldırım Han da Tokat'tan kalkıp Ankara üzerine hareket ederek, Timur ile savaşı burada vermeyi ve onu karşılamayı kararlaştırmış […] s. 261

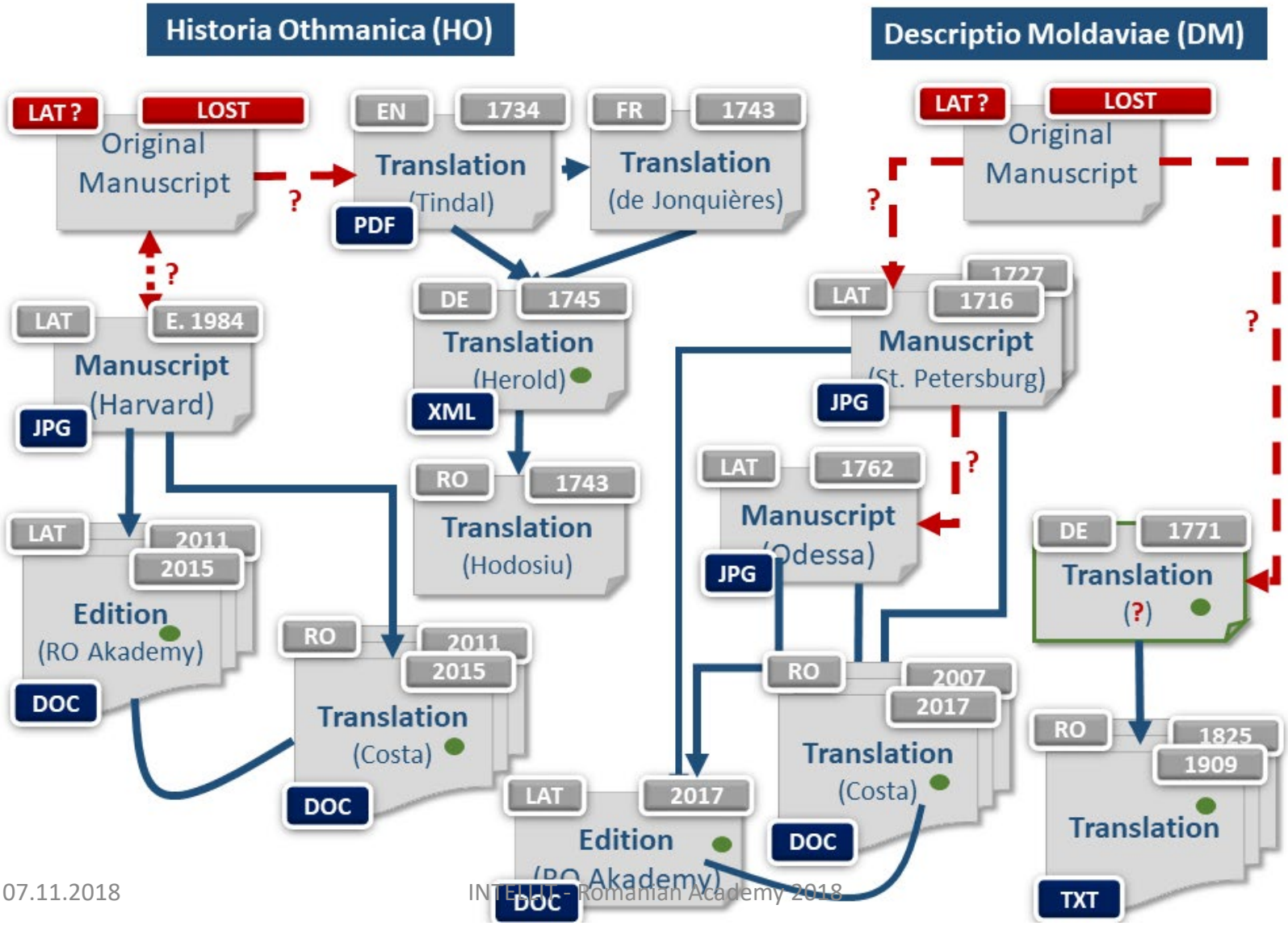
Solakzade Mehmet Hemdemi Efendi (1590-1657), Tarih-i Solakzade (1660?): O zamanda Yıldırım Bayezid Han 90000 asker ve düşman avlayan sipahi ile Ankara'ya gelip Timur'un karşısında karar kıldı. S. 98

Mehmed Neşrî (Hüseyin bin Eyne Bey?) (?-1520), Kitâb-ı Cihannümâ (1485?): […] Engüri'de, Sivriler'de Timur'a mukabil olup kondı. S. 351

Links to scientific articles/ webpages/ archives/ sources?

- www.islamansiklopedisi.info/ (TDV Encyclopedia of Islam)
- www.deutsche-biographie.de (German Biography)
- www.dnb.de/DE/Home/home_node.html (Deutsche Nationalbibliothek)
- www.suleymaniye.yek.gov.tr/ (Süleymaniye Manuscrip Library)
- http://olden.rsl.ru/en (Russian State Library)
- https://katalog.devletarsivleri.gov.tr/ (Ottoman Archives

The Corpus
Manuscripts, editions, translations- unclear tradition



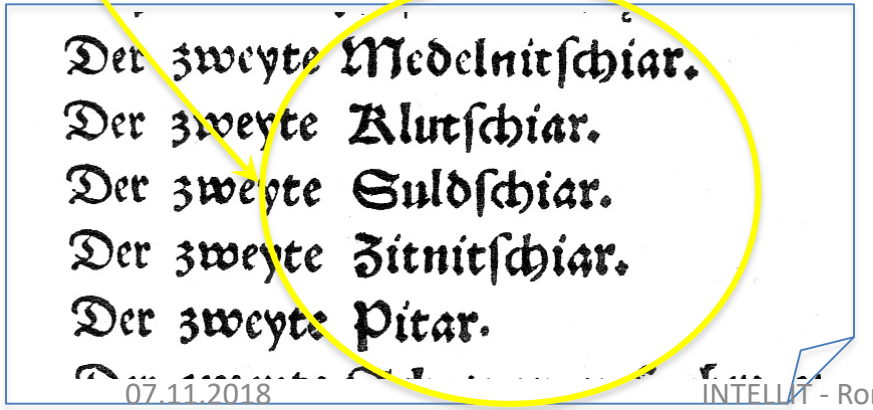
Directions of investigation in DH terms

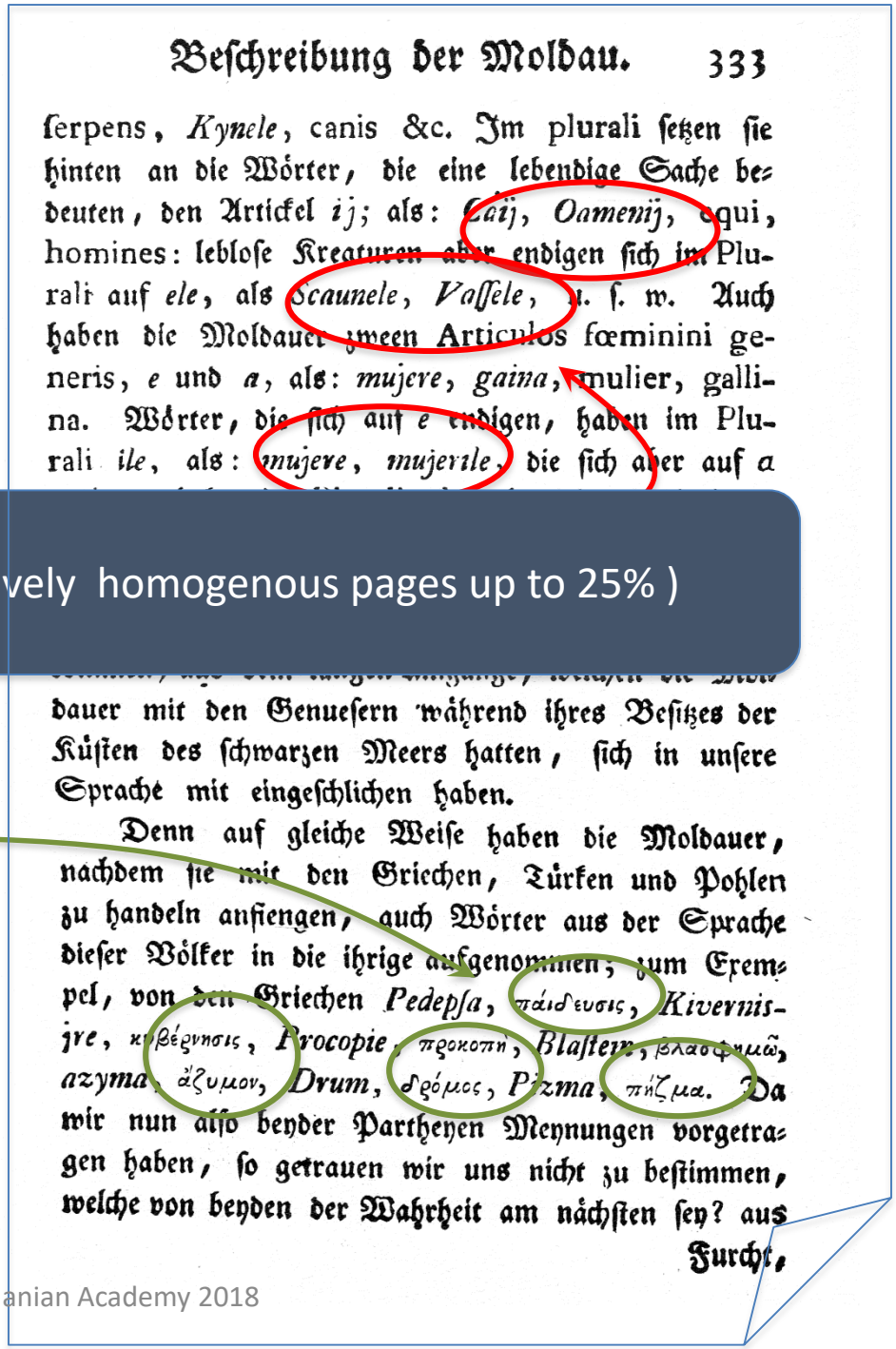
- **Reliability:**
  - Of the original: are the quotations made by Cantemir grounded? Is there a concordance between his degree of trust in these sources and the current knowledge about them (e.g. is there any evidence that a person which Cantemir claims to have spoken to, really lived in that time?)
  - 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

Corpus creation – challenges

- Surface form – level
  - German texts are in black-letter typeface
  - Mixed typefaces
  - Mixed scripts

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


Der zweyte Medelnitschiar.
Der zweyte Klutschiar.
Der zweyte Suldschiar.
Der zweyte Zitnitschiar.
Der zweyte Pitar.


Beſchreibung der Moldau. 333

ſerpens, *Kynele*, canis &c. Im plurali ſetzen ſie hinten an die Wörter, die eine lebendige Sache bedeuten, den Artickel *ij*; als: *Cãij*, *Oamenij*, equi, homines: lebloſe Kreaturen aber endigen ſich im Plurali auf *ele*, als *Scaunele*, *Vaſſele*, u. ſ. w. Auch haben die Moldauer zween Articulos fœminini generis, *e* und *a*, als: *mujere*, *gaina*, mulier, gallina. Wörter, die ſich auf *e* endigen, haben im Plurali *ile*, als: *mujere*, *mujerile*, die ſich aber auf *a*

dauer mit den Genueſern während ihres Beſitzes der Küſten des ſchwarzen Meers hatten, ſich in unſere Sprache mit eingeſchlichen haben.

Denn auf gleiche Weiſe haben die Moldauer, nachdem ſie mit den Griechen, Türken und Pohlen zu handeln anfiengen, auch Wörter aus der Sprache dieſer Völker in die ihrige aufgenommen; zum Exempel, von den Griechen *Pedepſa*, παίδευσις, *Kivernisire*, κυβέρνησις, *Procopie*, προκοπή, *Blaſtem*, βλασφημῶ, *azyma*, ἄζυμον, *Drum*, δρόμος, *Pizma*, πίζμα. Da wir nun alſo beyder Partheyen Meynungen vorgetragen haben, ſo getrauen wir uns nicht zu beſtimmen, welche von beyden der Wahrheit am nächſten ſey? aus

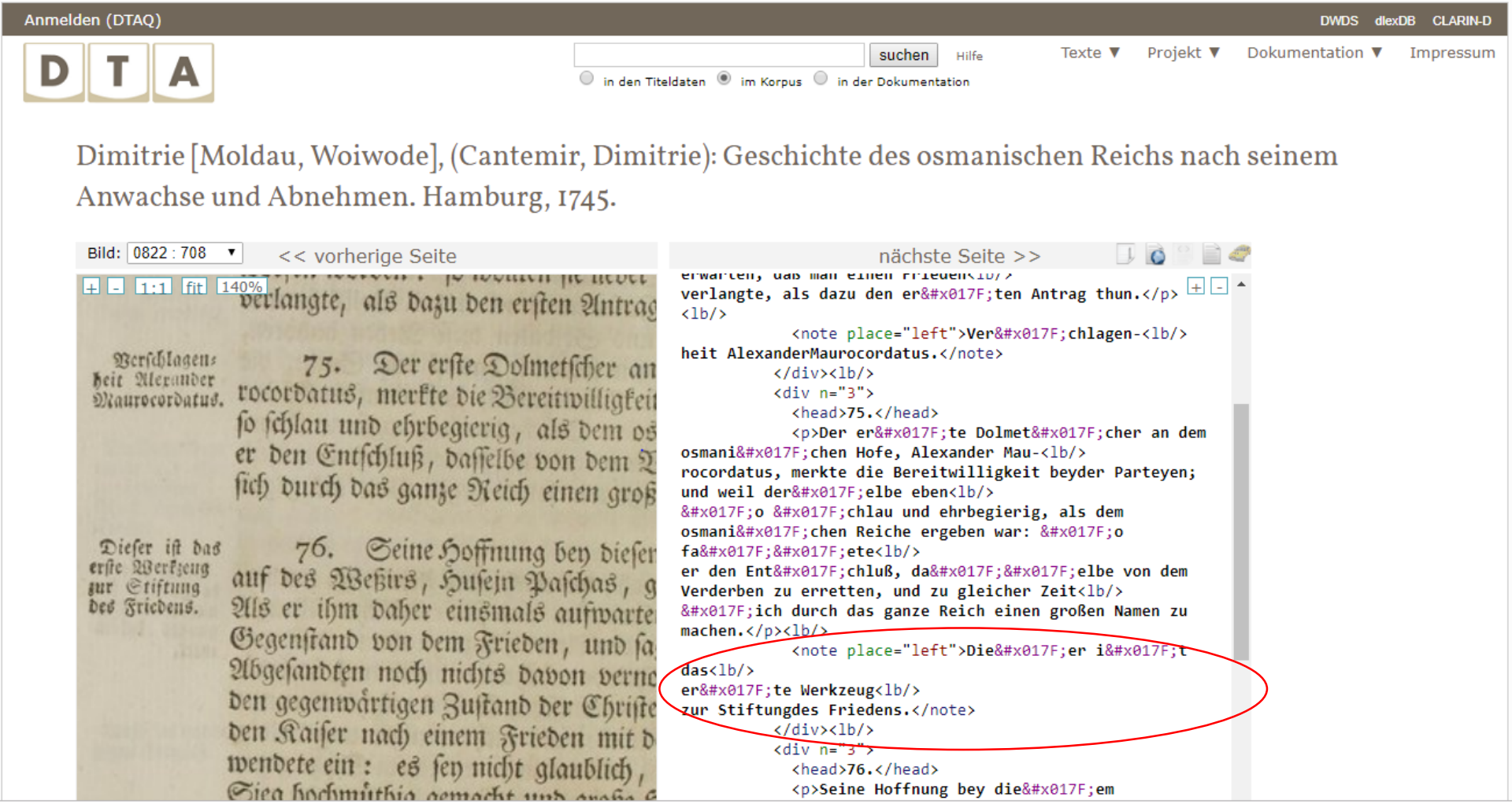
Furcht,

Current state-of- the art methods after digitization

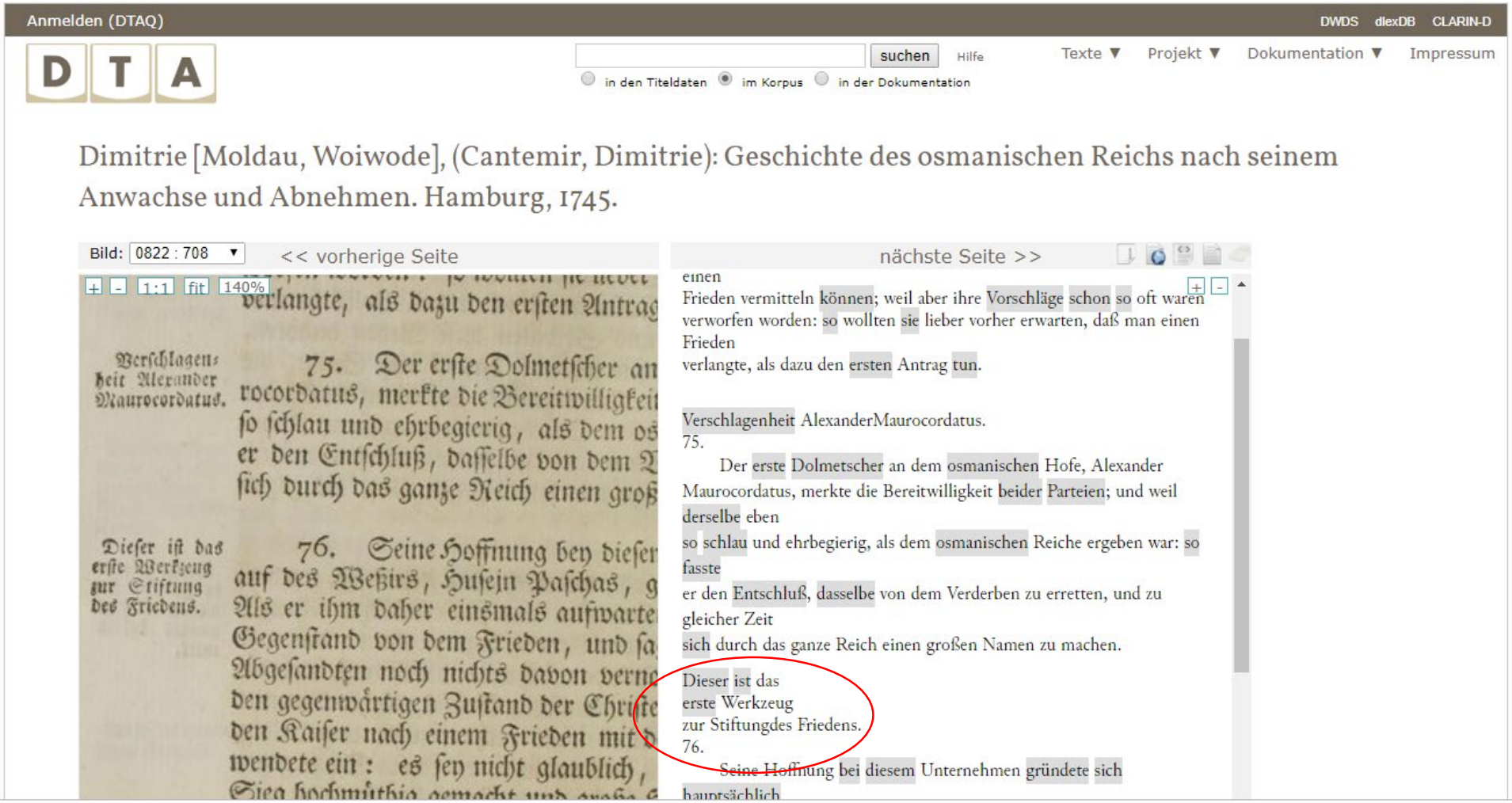
- Linguistic preprocessing :
  - Tokenisation- splitting into words (what happens with abreviations, special characters with a meaning
  - Lemmatisation – reducing word to its quotation form (ambiguities?)
  - PoS Tagging (ambiguities ?)
  - Parsing (syntactic structure)
  - Rarely lexical semantics
- Quantitative Evaluation (corpus linguistics)
- Machine Learning(black box) to extract some relations (on exactly what?; what and how correct is in fact preserved from the initial text)

Digitization and implications for further processing… -1-

Digitization and implications for further processing… -2-



Digitzation and implications for further processing… -3-



Editorial issues regarding the Latin text of Cantemir's *Historia Othmanica* and *Descriptio Moldaviae* from the standpoint of digital humanities

Ioana Costa

DIMITRIE CANTEMIR

Istoria măririi şi decăderii Curţii othmane

ACADEMIA ROMÂNĂ

Fundaţia Naţională pentru Ştiinţă şi Artă

2015

Editarea textului latinesc, aparatul critic şi indicii
OCTAVIAN GORDON, FLORENTINA NICOLAE, MONICA VASILEANU

Traducere din limba latină IOANA COSTA

Studiu introductiv ŞTEFAN LEMNY

Academia Română

Fundaţia Naţională pentru Ştiinţă şi Artă

Bucureşti, 2015

(2 vol.)

CONSPECTVS SIGLORVM NOTARVMQVE (1)

( )	Cantemirii abbreviationis expansio
< >	Litterarum / verborum ab auctore vel a scriba oblitarum supplementum
[ ]	Litterarum secus a scriba insertarum expunctio
[[ ]]	Litterarum, verborum, paragraphorum ab ipso auctore vel a scriba expunctio
{ }	Inter- vel intralineare auctoris vel scribae additamentum
{{ }}	Auctoris (rarius scribae vel alicuius lectoris) marginale additamentum

CONSPECTVS SIGLORVM NOTARVMQVE (2)

…	Spatium plurium litterarum a scriba vacuum relictum, vel illegibilium
vere Italicae	litterae lectorem de erroris correctione certiorem faciunt
(sic)	Errorum peculiari sensu praeditorum nota
I, V, 5, 41	Ad Cantemirii Textum remittit, e.g. ad Librum I., Caput V., Sectionem 5, paginam 41 (Cantemirii H(arvardiensis) Codicis).
I, V (a) 513	Ad Cantemirii Annotationes remittit, e.g. ad Librum I., Caput V., notam (a), paginam 513 (Cantemirii H(arvardiensis) Codicis)

Paginae Cantemirii H Codicis in margine nostri textus notatae sunt.

Imp. Oth., p. 6

- illa Graeci Imperii periodus fuit, in anno quo acciderit, gravissimi dissentiunt scriptores, aliis illam ad annum 1452., aliis ad 1453. referentibus. Vnde colligi potest, quanti in reliquis Turcicae historiae capitibus, v(erbi) g(ratia) Sultanorum diebus natalibus, emortualibus etc. definiendis, commissi fuerint errores. His ut occurramus, haud inconsultum duximus, collatis Christianis Turcisque autoribus, de hoc computo paulo accuratius inquirere, eoque ipso / probare recte nos in ea, quam iam edimus, historia annos Hegirae cum aera Christiana conciliasse. Non vero opus esse existimamus de nomine „Hegirae“ <…> litem movere, atque disquirere, utrum ea a fuga Falsoprophetae Mecca Medinam, an ab ipsius obitu (ut nonnullis Muhammedanis placet) ducat initium.

DIMITRIE CANTEMIR

Despre numele Moldaviei: în vechime şi azi • Istoria moldo-vlachică • Viaţa lui Constantin Cantemir • Descrierea stării Moldaviei: în vechime şi azi

ACADEMIA ROMÂNĂ

Fundaţia Naţională pentru Ştiinţă şi Artă

2017

Editarea textului latinesc, aparatul critic şi indicii FLORENTINA NICOLAE

Traducere din limba latină IOANA COSTA

Studiu introductiv ANDREI EŞANU, VALENTINA EŞANU

Academia Română

Fundaţia Naţională pentru Ştiinţă şi Artă

Bucureşti, 2017

CONSPECTVS SIGLORVM NOTARVMQVE (1)

• add.	*addidit, addiderunt*	a(u) adăugat
• a. m.	*alia manu*	altă mână
• ant.	*anterior*	anterior,-oară
• cf.	*confer*	compară cu
• corr.	*correxi, correxit, correxerunt*	am corectat, a(u) corectat
• del.	*delevi, delevit, deleverunt*	am eliminat, a(u) eliminat
• D-Rom	*Daco-Romanice*	româneşte
• ed.	*editio*	ediţia
• emend.	*emendavi*	am emendat
• f.	*folium*	foaia

CONSPECTVS SIGLORVM NOTARVMQVE (2)

• i.e.	*id est*	adică
• interl	*inter lineas*	între rânduri
• loc. cit.	*loco citato*	în locul citat
• ms(s).	*manuscriptum, manuscripta*	manuscrisul, manuscrisele
• mut.	*mutavit*	a schimbat
• n.	*nota*	nota
• om(m).	*omisit, omiserunt*	a(u) omis
• p.	*pagina*	pagina

CONSPECTVS SIGLORVM NOTARVMQVE (3)

• p. m.	*prima manus*	prima mână a autorului
• ref.	*refeci(t)*	a(m) refăcut
• s. a.	*sub anno*	sub anul
• scr.	*scripsi, scripsit, scripserunt*	am, a(u) scris
• s. m.	*secunda manus*	a doua mână (revizuirea, de către autor, a textului)

CONSPECTVS SIGLORVM NOTARVMQVE (4)

- s. v. *sub voce* sub cuvântul
- tempt. *temptavit* a încercat
- t. m. *tertia manus* a treia mână (revizuirea de către autor, a textului)
- v. *vide* vezi
- 147v (etc.) *147 verso* pagina 147 verso (numerotaţie în MN)

Descr. Mold., B2, C6

- Labefactata post illa tempora Romanorum dominatione, oppressere Moldaviam crebrae Barbarorum, Sarmatarum, Hunnorum, Gothorum, invasiones coactaeque fuere Romanae coloniae superare Alpes, ac in montana regione Maramoris contra Barbarorum / furorem quaerere receptaculum. Ibi, posteaquam, per aliquot s<a>ecula, locorum difficultate defensi, suis legibus regibusque vixissent, tandem, circa annum Domini < >, cum premi se animadverterent incolarum multitudine, Regis Bogdani filius **Dragosz**, cum 300 saltem hominibus, venantis specie, montium transitum Ortum versus tentare constituit. Hoc in itinere casu invenit bovem sylvestrem, Moldavis ‘Dzimbr<u>’ dictum, et, dum eum persequitur, ad montium radices descendit.

Descr. Mold. p. 1134

9 Huius amnis antiquum nomen investigare non potuimus *mg. A, in textu B, C* / 10 *Ante* Suczava, *del.* In hunc se exonerant *A* / 11 ubi...Metropoli: urbs…Metropolis *p. m. A* / 12 Moldova] Moldava *p. m. A; Post* Moldova, *in textu* cuius nominis *(*nomen *p. m.)* rationem supra explicavimus *et supra* a quo tota feudis regio nomen ducit *del. A*, *postea infra* rationem supra explicavimus Cap. 1mo *alia manus scripsit; supra* Moldava*, signum inserendae notae, ulterius deletae, exstat. Haec nota ab ipsa manu Cantemirii scripta est*: de cuius nomine, licet incerto authore, talis apud Moldavos fertur fabula: Dragoszum *(*-um *ulterius add.),* Bogdani filium (filius *p. m., inde* de quo in priori libro fuse dictum est *del.*), cum 300 saltim hominibus, venantis specie, ex Marmarusio *(*Maramorusio *DS)*, Transylvaniae regione montium transitum tentare constituisse, hoc in itinere casu invenisse bovem sylvestrem, Moldavis 'Dzimbru' dictum et, dum eum persequitur, ad montium radices descendisse. Porro, cum catula quaedam venatica, quam prae caeteris diligebat, 'Molda' dicta, fortius ferae instaret, aestuans fera (*sic*), in profluentem se proiecisse et telis ibi confectam. Canem vero, quae in ipsis aquis venatum quaesiverat fugientem, rapidis fluvii undis absorptam *(post* absorptam*, del.* fuisse *A)* in huius *(post* huius, *del.* me *A)* itaque memoriam, fluvium 'Moldovam' a Dragosso Principe appellatum fuisse. Loco etiam, ubi haec acciderunt, suae gentis nomen, 'Roman' (quae civitas Bonfinio 'Forum Romanorum' est), indidisse. / 13 Transylvaniae] Transylvanie *p. m. C* //

C20 1 et montibus B, *supra* et *lineatum, s .m.* ex *corr.* / 2 Bistriza] Bistri…a *p. m. A* / 3 *Post* miscentur, *em.* (s(ive) miscetur) *B* / 4 Molnitza] Molniza *p. m. A* / 5 Valeniagra] Valenagra *p. m. A;* Valeneagra *B;* Valenjagra *ex* Valeneagra *C* //

Caps (or not)

- Vesirium v/**V**esirius
- Kiehaiam k/**K**iehaia
- Tefterdarium t/**T**efterdarius
- Baronum b/**B**aronus

I/J

- iussit VERB 1 jubeo iubeo
- iniuriis NOUN 1 injuria iniuria
- iniustam ADJ 1 injustus iniustus

V/U

- Vrbis NOUN 1 **Vrbis** urbs/Vrbs

X-*que*

- eam**que** PRON PRON+**CONJ**
- tibi**que** NOUN PRON+**CONJ**
- eos**que** PRON PRON+**CONJ**

Editorial marks

• i	NOUN	1	eo	1
• (	PUNCT	1	(	1
• d	X	1	d	1
•)	PUNCT	1	)	1
• e	ADP	1	ex	1
• (	PUNCT	1	(	1
• st	VERB	1	st	1
•)	PUNCT	1	)	1

id est: PRON is, VERB sum

False full stop

Itaque, anno 923. {{**H. 923**.}}, maioribus quam ante copiis executurus destinata...

Current state-of- the art methods after digitization

- Linguistic preprocessing :
  - Tokenisation- splitting into words (what happens with abreviations, special characters with a meaning
  - Lemmatisation – reducing word to its quotation form (ambiguities?)
  - PoS Tagging (ambiguities ?)
  - Parsing (syntactic structure)
  - Rarely lexical semantics
- Quantitative Evaluation (corpus linguistics)
- Machine Learning(black box) to extract some relations (on exactly what?; what and how correct is in fact preserved from the initial text)


Demetrii Kantemirs
ehemaligen Fürsten in der Moldau,
historisch- geographisch- und politische
Beschreibung
der
Moldau,
nebst
dem Leben des Verfassers
und
einer Landcharte.
EX LIBRIS
BIBLIOTHECA
ACADEMIEI
ROMÂNE
Frankfurt und Leipzig,
1771.

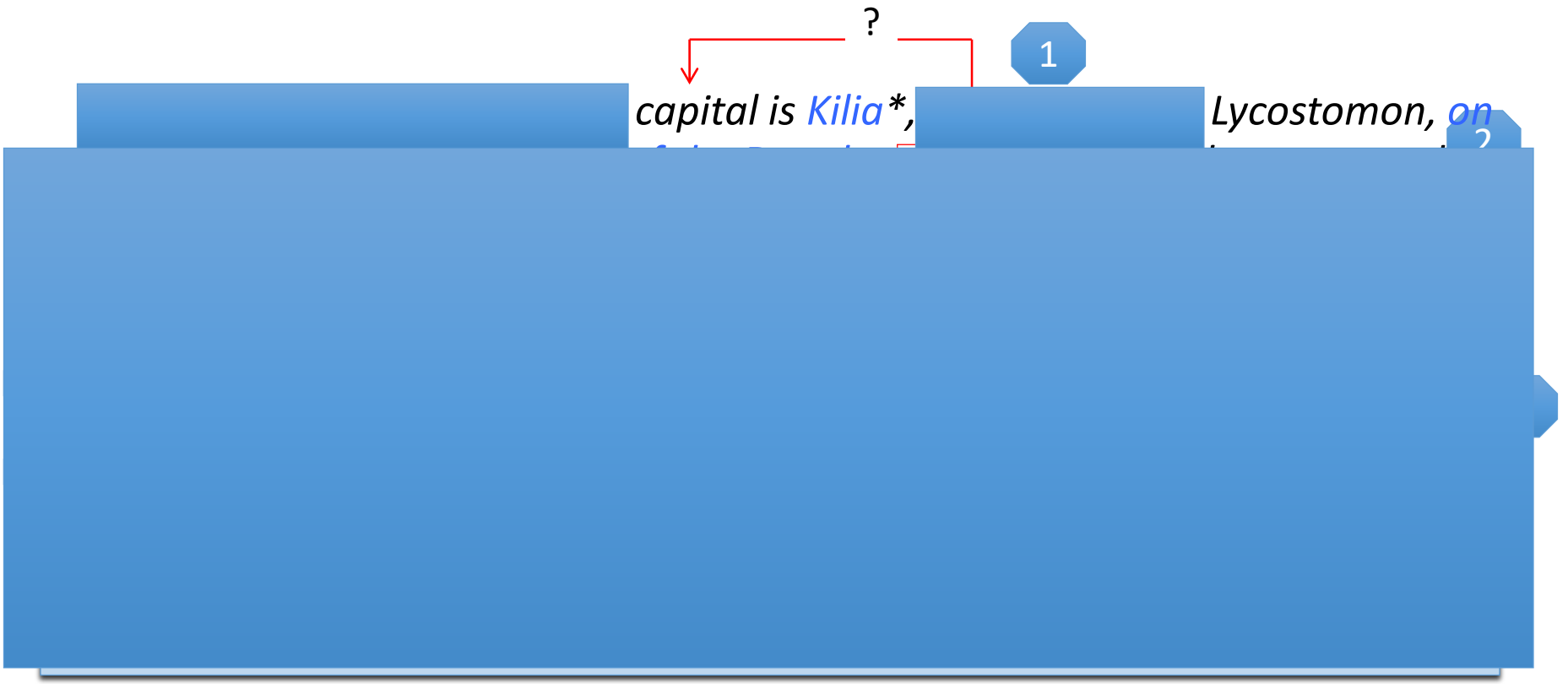
„Nu îndrăznim să spunem ce e adevărat şi ce e fals într-o asemenea întunecime a istoriei. “

(Dimitrie Cantemir, Descrierea stării Moldaviei in vechime si azi, traducere Ioan Costa 2017)

„I do not dare to decide what is the truth about this matter, given the high darkness of this story“

A Complicated Explicit Example: Ambiguity

(Cantemir, Descriptio Moldaviæ, p.73 transl.)



Although our books do not record his descendants, it is a wellknown legend for us that he is coming from the moldavian kings

Enriched Classical Markup

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** [illegible]**her von gemeiner Herkunft mit einem so** [illegible] Gefolge auf die Jagd (welche die Molda[illegible]cken Gelegenheit gegeben,).

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		

One should trust even more this opinion, as one can hard think that….

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 sich 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.)

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

Sources /levels of vagueness/uncertainity to be annotated

1. Linguistic markers for vagueness

2. Factual uncertainity

2.1 References to external written materials (publicatiosn)

2.2 References to external persons, places, names

2.3 References to events

2.4. References to other external knowledge (e.g. legends, folk beliefs)

3. Editors

3.1. Editorial marks

() pretty sure extensions

< > correction

[ ] deletion

{ } marginals /between line

3.2 „Footnotes“

4. Metadata

4.1 genre

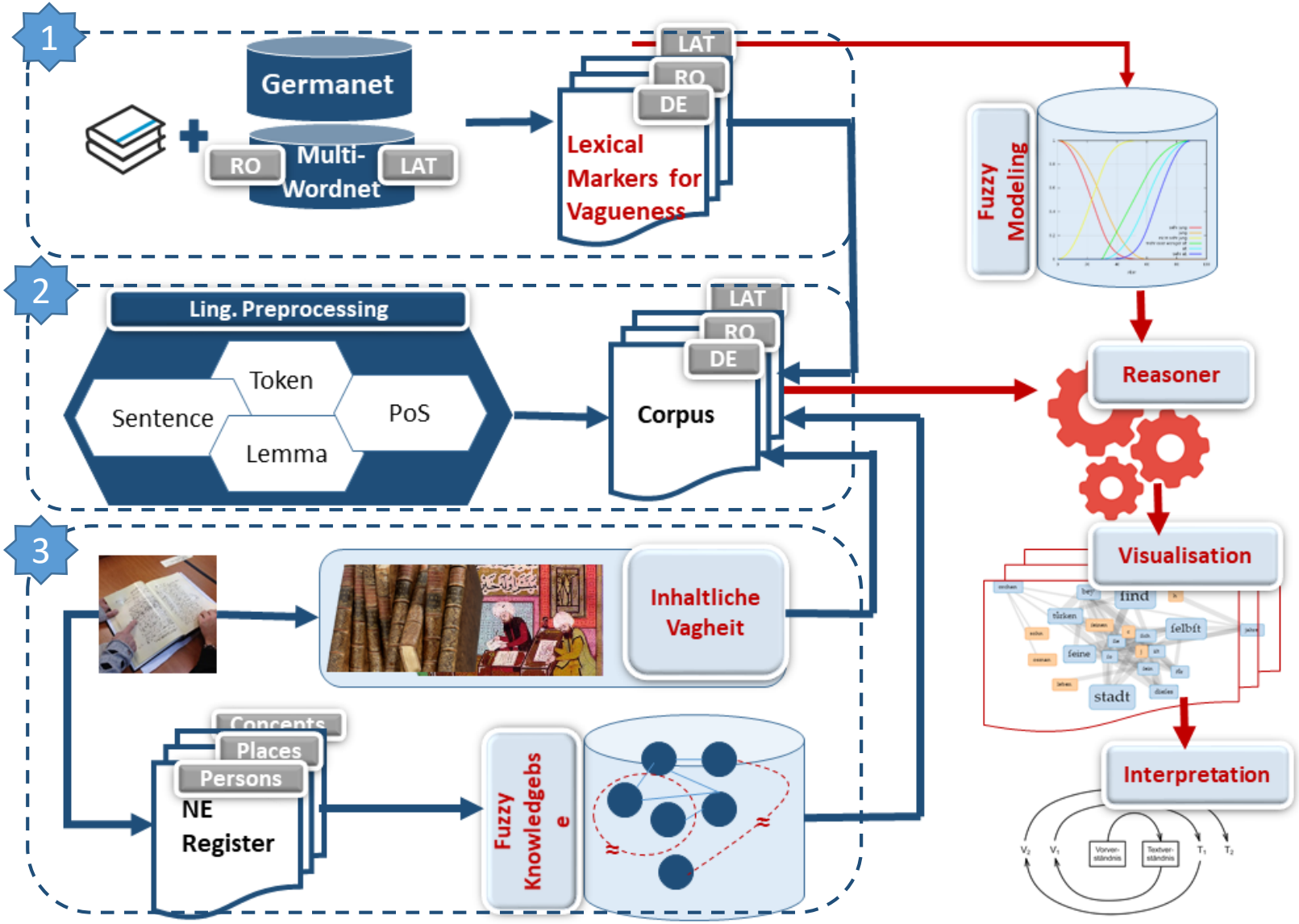
4.2. author

4.3 translation

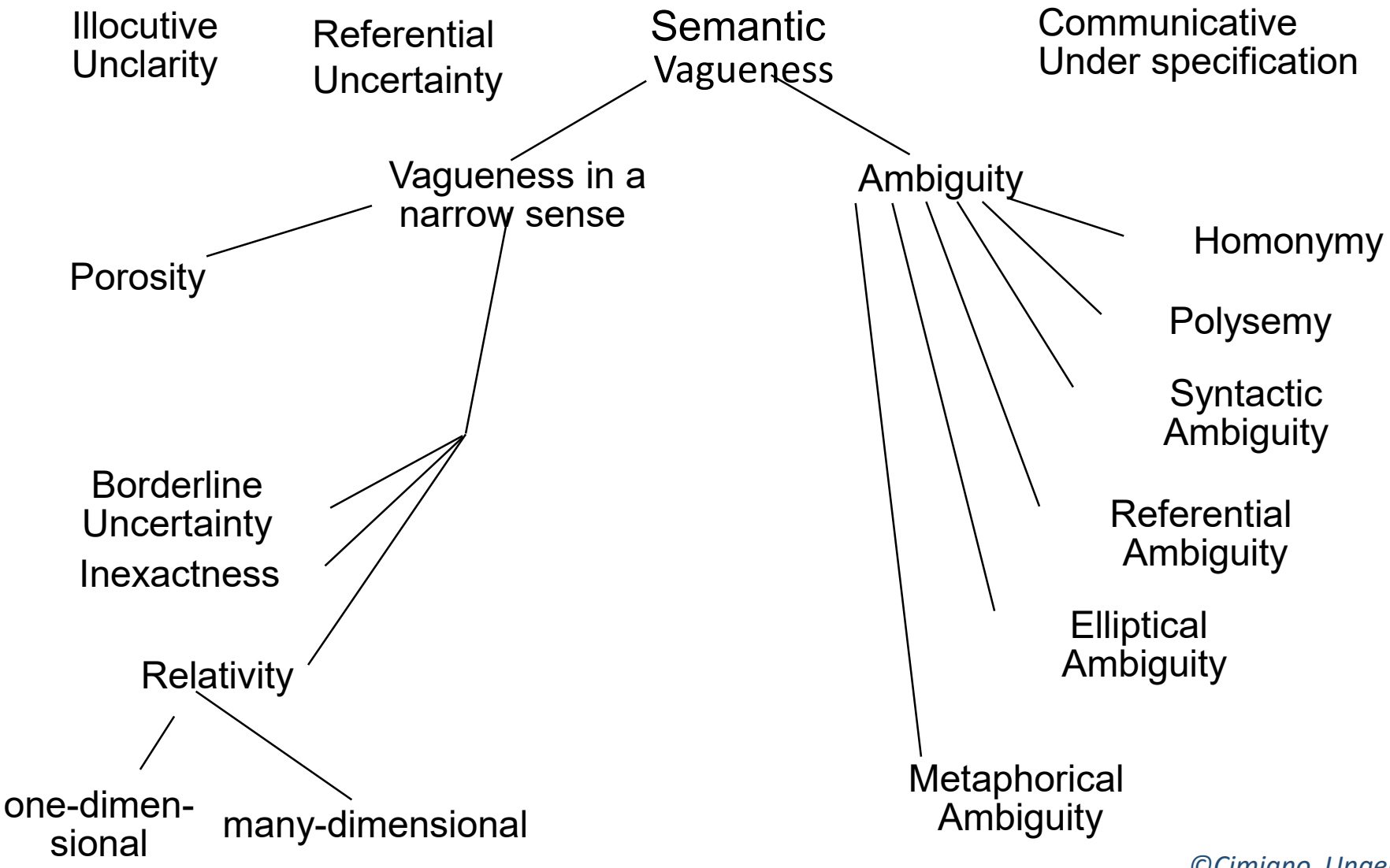
4.4. tradition path

Vagueness annotation is useful only if it is accompagned by inference rules and adequate ontological knowledge-base

System Architecture



M.Pinkal's Schema of Semantic Vagueness



©Cimiano, Unger and McCrae

Selected markers for linguistic vagueness

1. comparatives, inexact adjectives e.g. *“hostile”, “near”, “distant”*
2. non-intersectives e.g. „supposend“, „so-called“
3. Hedges e.g. „rather“, „more or less“
4. inexact measures *a 4 days’ journey, 10 feet”*
5. modals (attitudes) e.g. probably, hopefully
6. subjonctives v+analyse
7. lexical quotation markers
8. vague quantifiers e.g. „many“, „most“
9. complex quantifiers e.g. *"roughly half of the 20-30 thousand soldiers”*
10. numbers
11. range expressions e.g. *“The beginning of the 18. century”*
12. unclear place *„the former prince“, „Moramor“*
13. unclear person e.g. *„the former prince“*
14. unclear time e.g. *„in prehistoric times“*

Annotating vague expressions

- 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.

Processing the lists of historical persons, terms and geographical indications

DHO	LHO	Turkish Sources	
Eskjischehri Altstadt	Eskiszehr	اسکی شهر Eskişehir	Stadt in Zentralanatolien
Gjermijan Phrygien	Giermian	کرمیان Germiyan	Landschaft in West-Anatolien/ Kütahya und ihre Umgebung

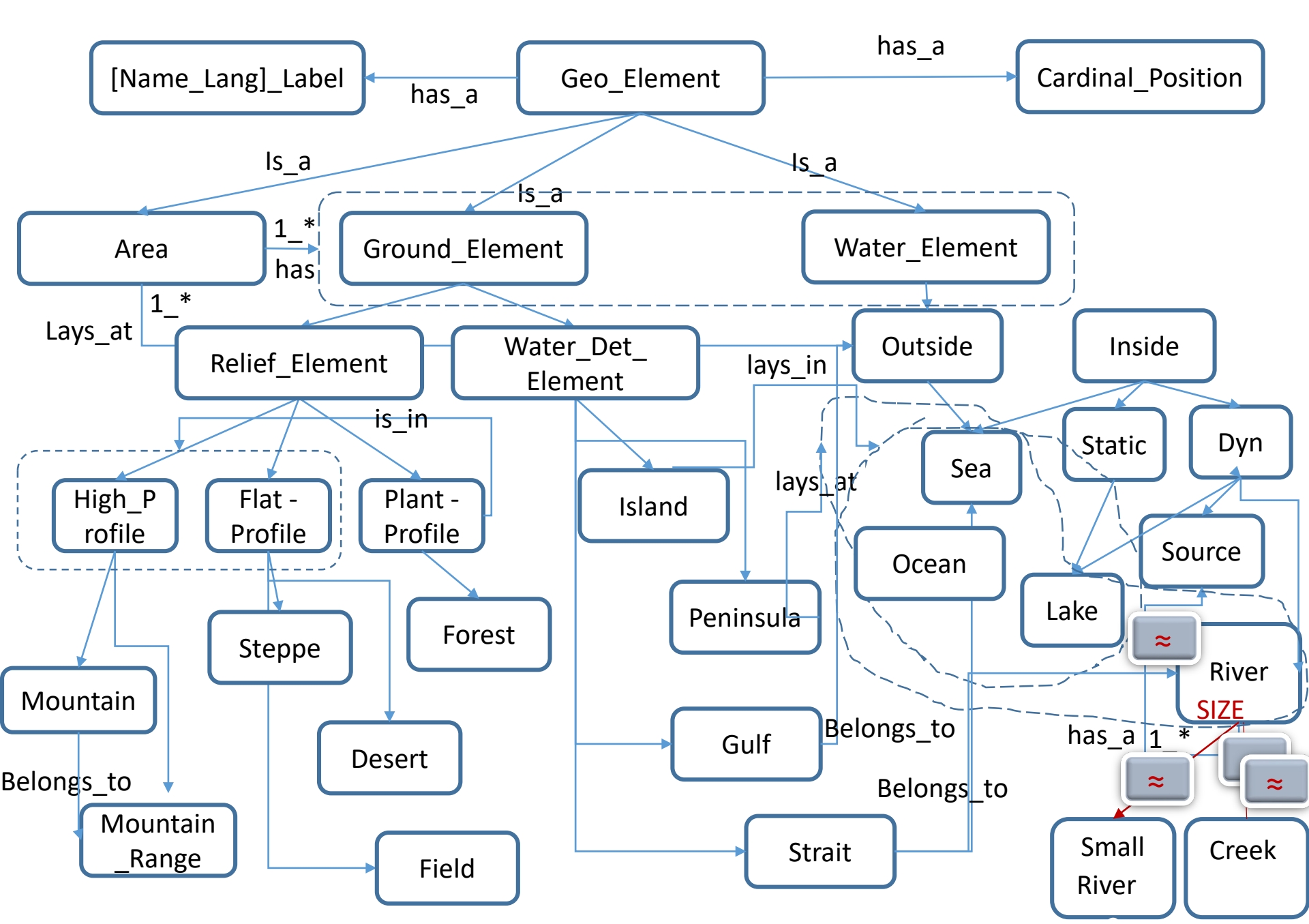
Places Database (initial ≈500)

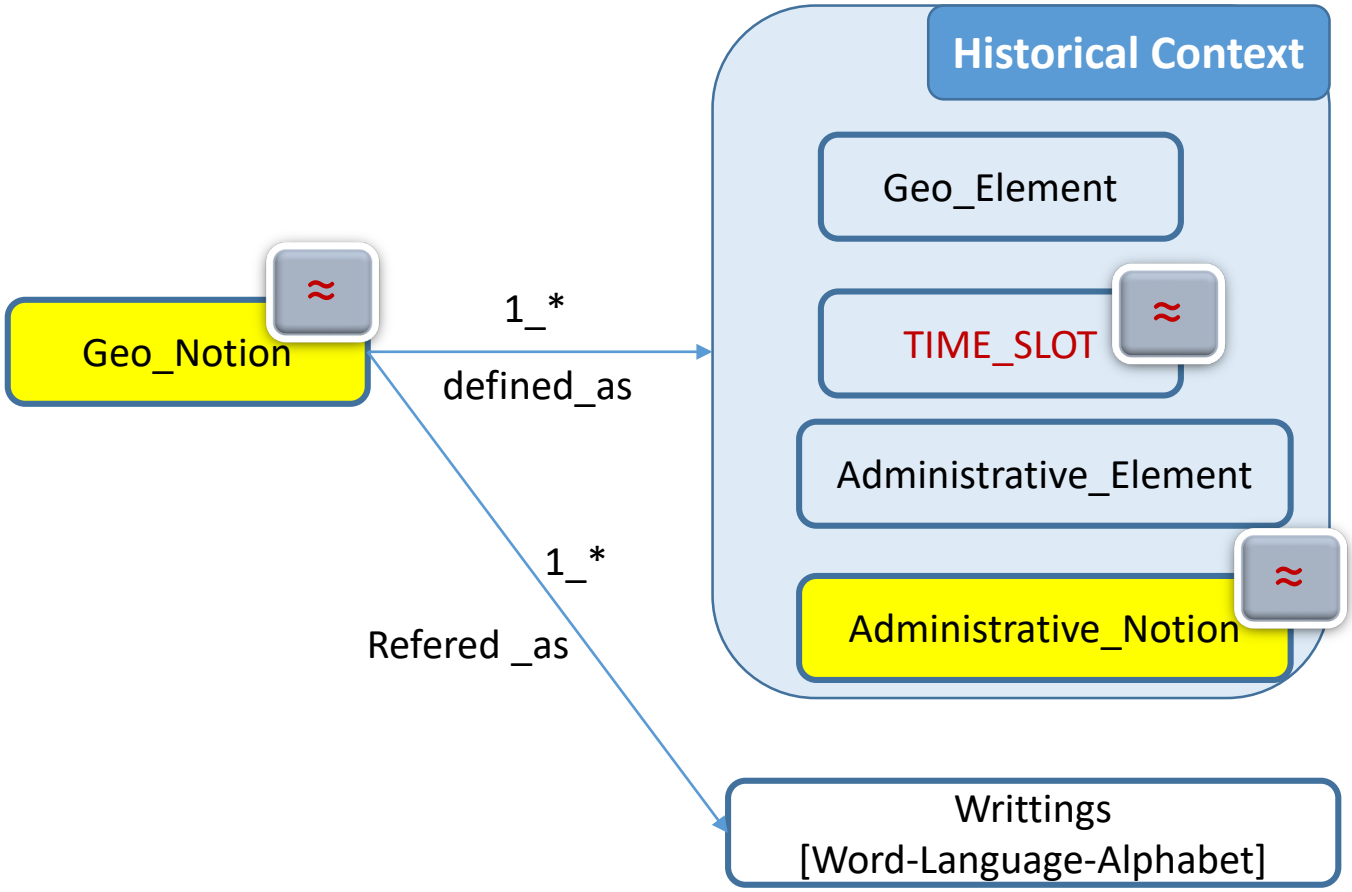
Begjlerbegj (Bsp. Begjlerbegj von Rumilien)	Beglérbèg/ Beglèrbègi Beglerbeg بکلربکی /بکلربك (Rumeli Beglerbeg)	بکلربکی Beylerbeyi بکلربك Beğlerbeğ روم ایلی بکلربکی (Rumeli Beylerbeyi)	Politik/ Staatsmann/ Oberster Staathalter in Rumelien und Anatolien/ Rang des Staathalters in wichtigen Ländern im Reich (Bsp. Beglerbeg von Budin), Nur die beiden höchsten Beglerbegs (von Anatolien und Rumelien) waren Mitglieder des kaiserlichen Diwans (Rat)

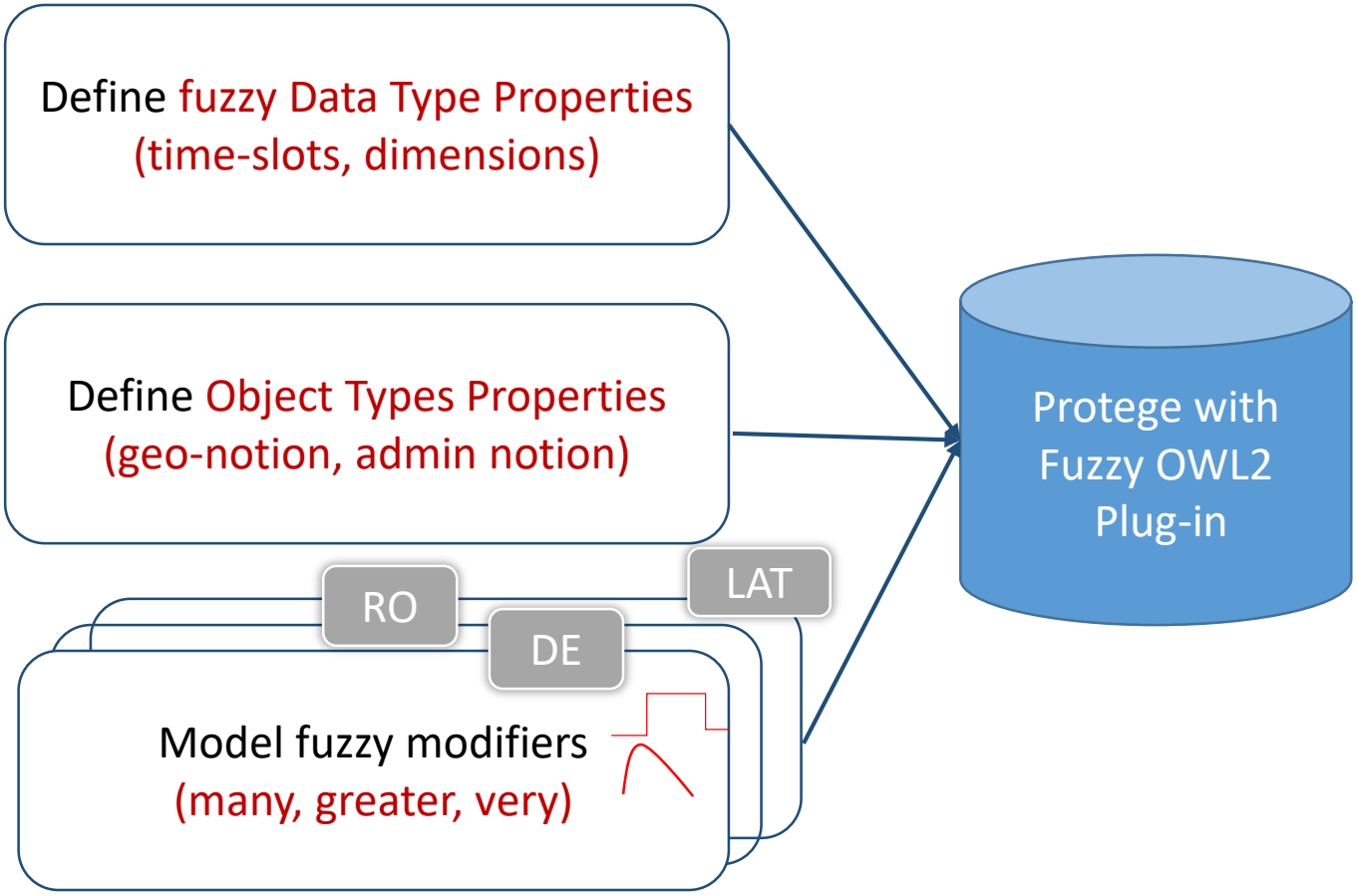
Processing the lists of historical persons, terms and geographical indications

DHO	LHO	Turkish Sources	
Aeladdin, Sultan von Ikonien Sultan Saʿladdin	الالدين Alaiddin/ Alaiddìn Alaiddinum Iconiae Sultanum Alaiddin Sultano	الالدين سلطان Sultan Alaaddin Konya Sultanı Alaaddin	Herrscher/ Seldschukischer Sultan, Alaeddin Keykubad II. (?-1254), reg. 1249-1254, Sohn von Gıyaseddin Keyhusrev II. Alaeddin Keykubad I. (?-1237), reg. 1220-1237, Sohn von Gıyâseddin Keyhusrev I. Keykubad I. oder II.? Borrowed vagueness?

Persons Database (initial≈300)

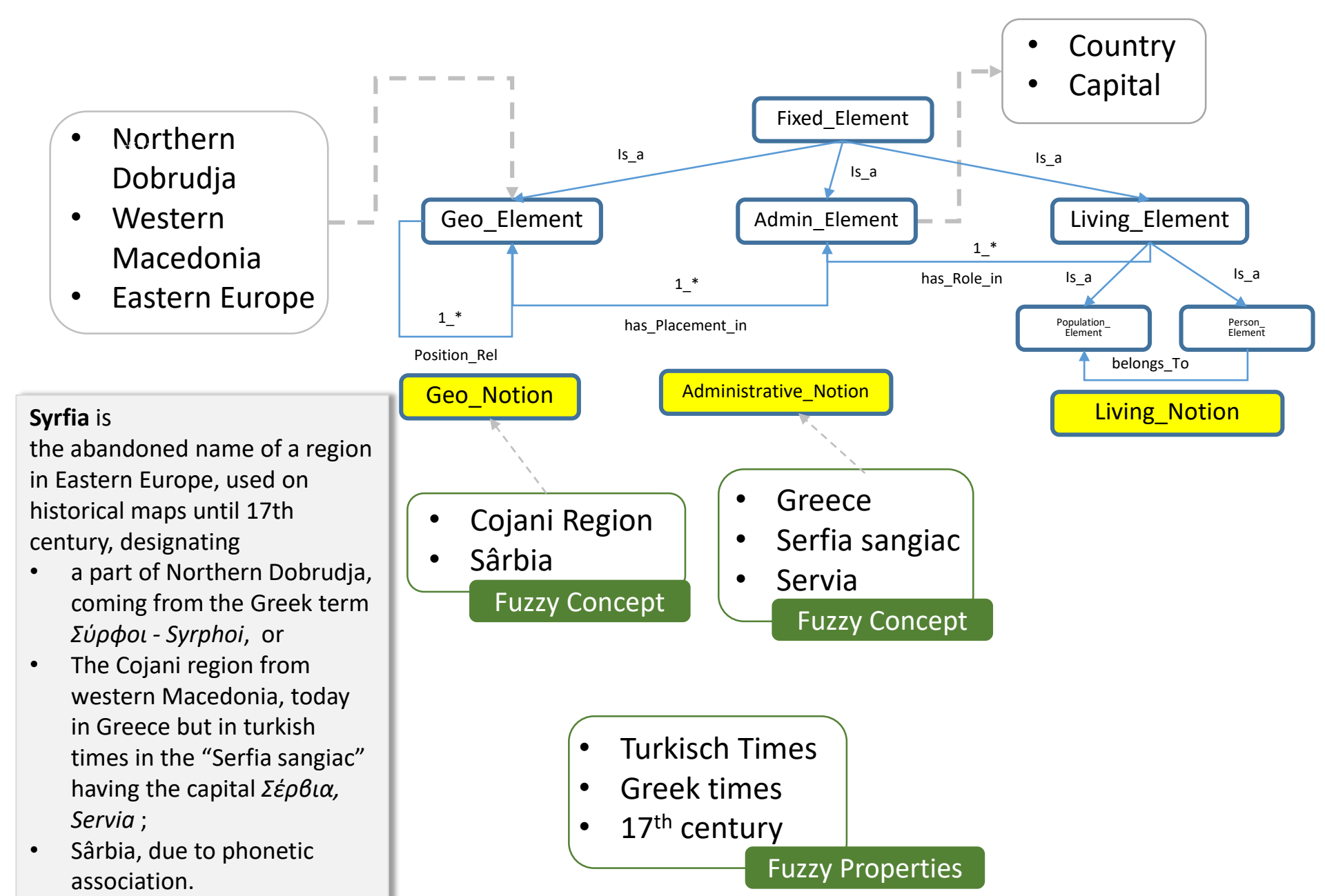

[Name_Lang]_Label
has_a
Geo_Element
has_a
Cardinal_Position
Is_a
Is_a
Is_a
Area
1_*
has
Ground_Element
Water_Element
1_*
Lays_at
Relief_Element
Water_Det_
Element
lays_in
Outside
Inside
is_in
Static
Dyn
Sea
High_P
rofile
Flat -
Profile
Plant -
Profile
Island
lays_at
Ocean
Source
Peninsula
Lake
Steppe
Forest
≈
River
Mountain
SIZE
Gulf
Belongs_to
has_a
1_*
Desert
Belongs_to
Belongs_to
≈
≈
Mountain
_Range
Strait
Small
River
Creek
Field


Historical Context
Geo_Element
TIME_SLOT
Administrative_Element
Administrative_Notion
Geo_Notion
1_*
defined_as
1_*
Refered _as
Writtings
[Word-Language-Alphabet]


Define fuzzy Data Type Properties
(time-slots, dimensions)
Define Object Types Properties
(geo-notion, admin notion)
RO
DE
LAT
Model fuzzy modifiers
(many, greater, very)
Protege with
Fuzzy OWL2
Plug-in

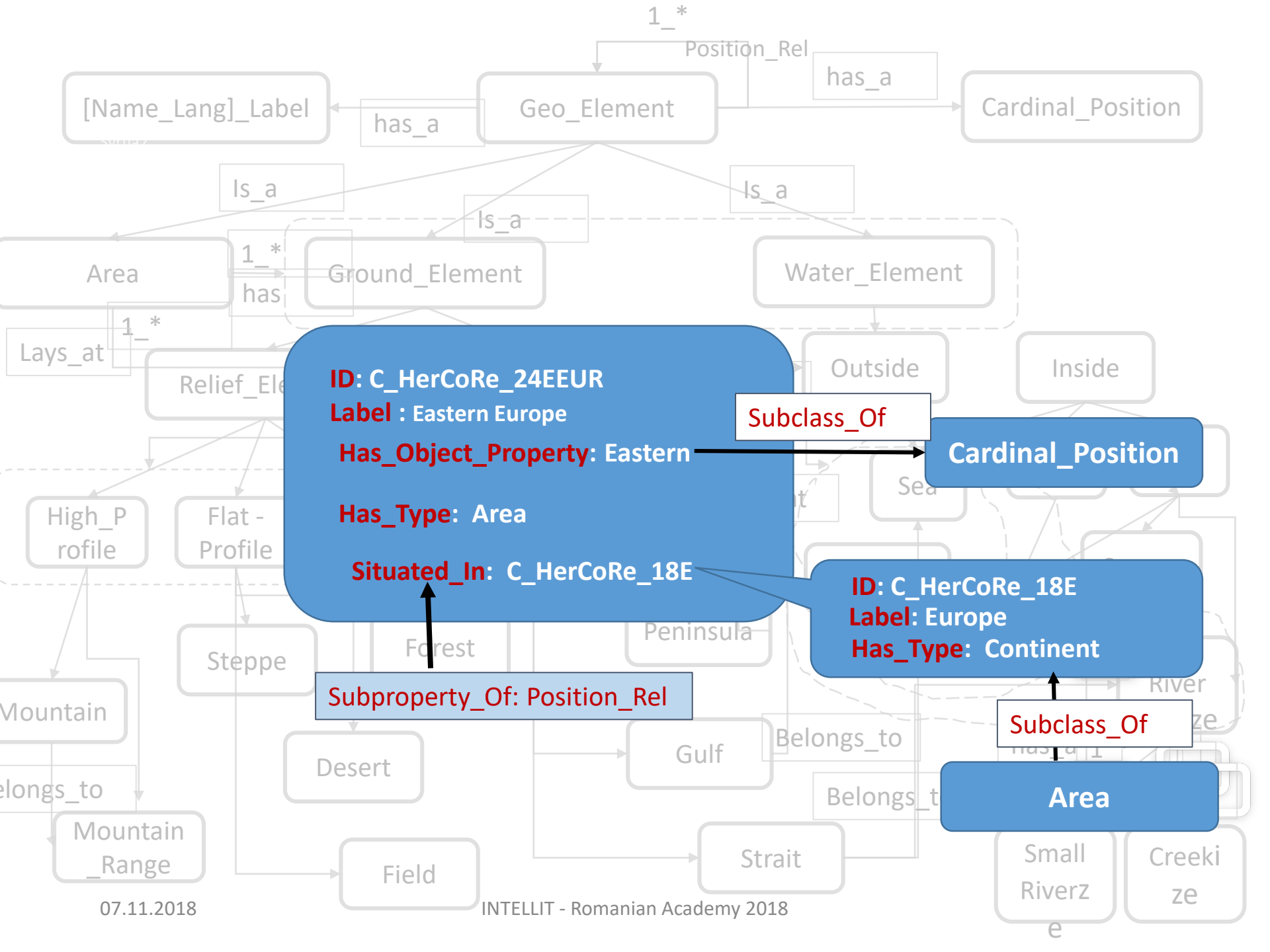
Example: Annotation of ambiguous places

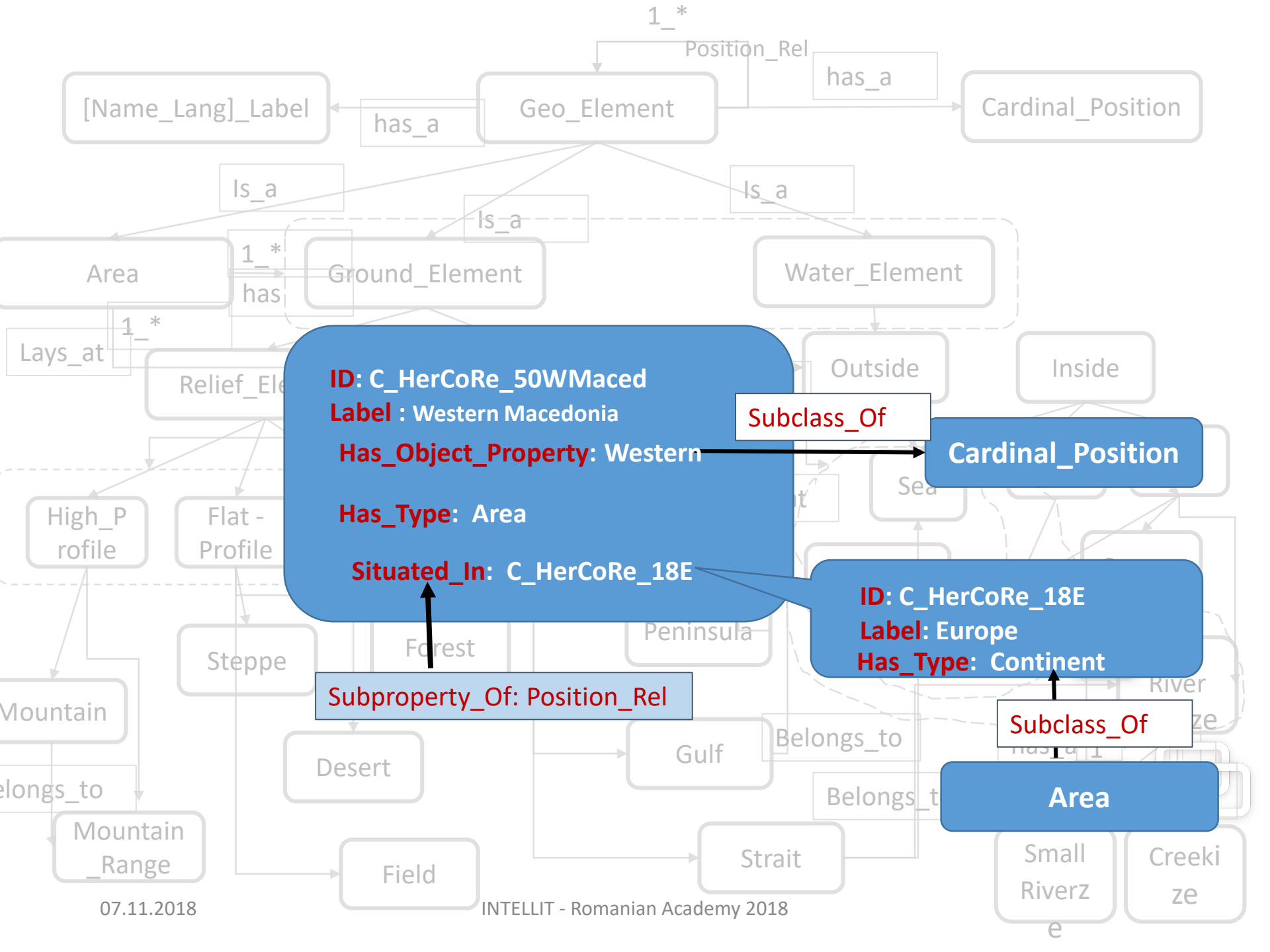
Ortelius
Map 1570

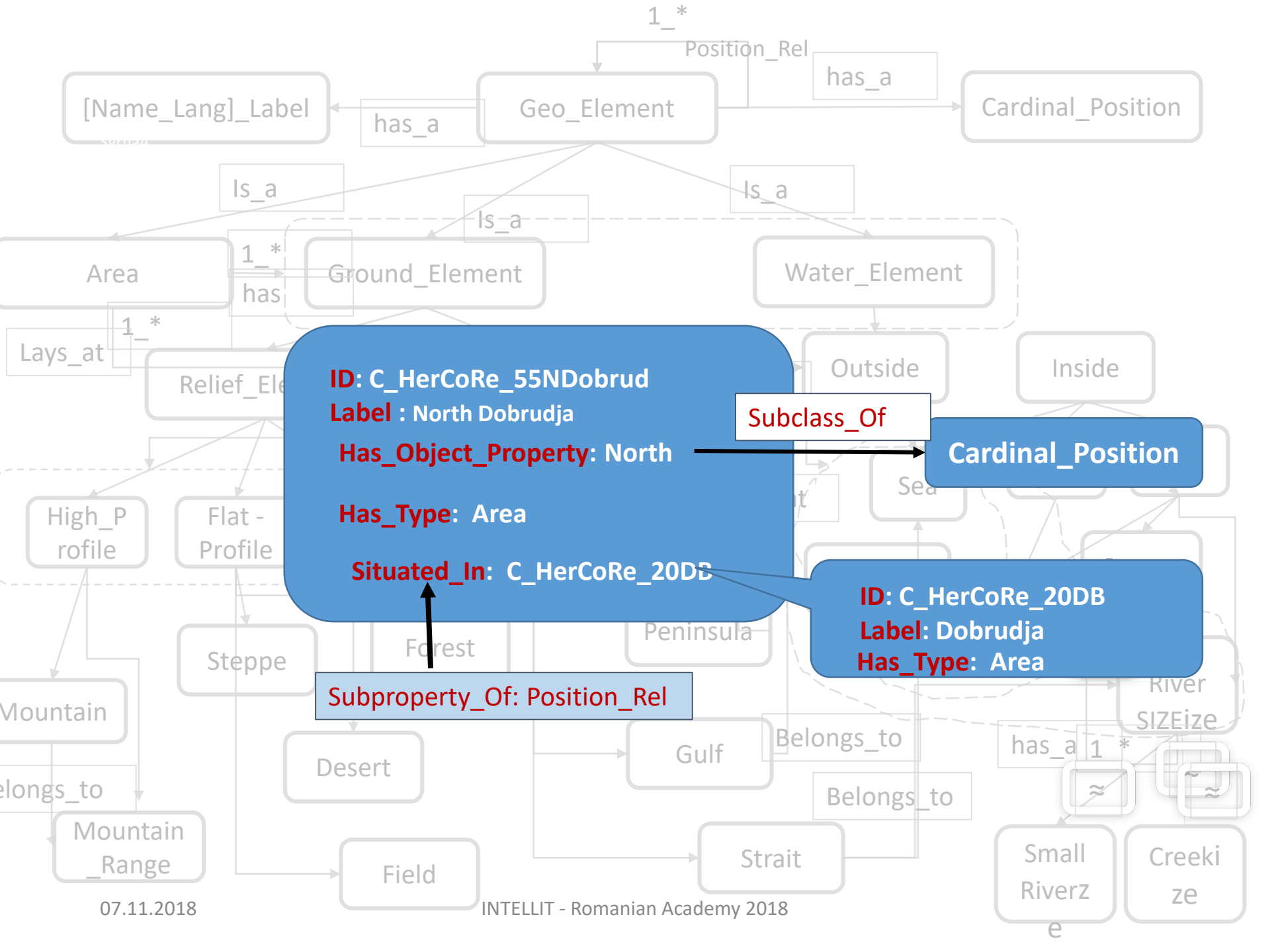


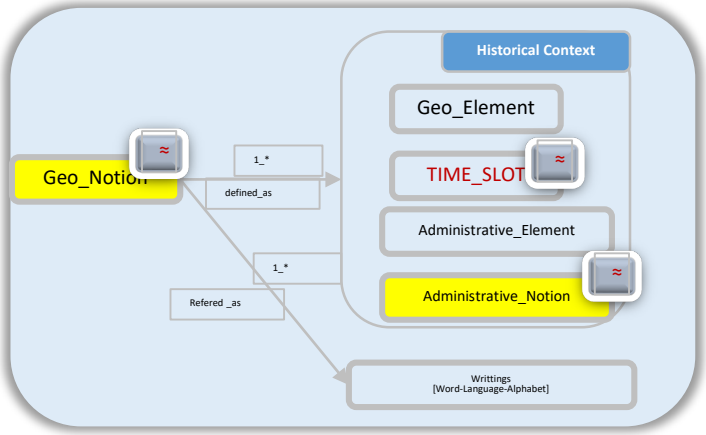
Syrfia is
the abandoned name of a region in Eastern Europe, used on historical maps until 17th century, designating

- a part of Northern Dobrudja, coming from the Greek term *Σύρφοι - Syrphoi*, or
- The Cojani region from western Macedonia, today in Greece but in turkish times in the "Serfia sangiac" having the capital *Σέρβια, Servia* ;
- Sârbia, due to phonetic association.


1_*
Position_Rel
has_a
[Name_Lang]_Label
has_a
Geo_Element
has_a
Cardinal_Position
Is_a
Is_a
Is_a
Area
1_*
has
Ground_Element
Water_Element
1_*
Lays_at
Relief_El
Outside
Inside
ID: C_HerCoRe_24EEUR
Label : Eastern Europe
Has_Object_Property: Eastern
Subclass_Of
Cardinal_Position
Has_Type: Area
Situated_In: C_HerCoRe_18E
ID: C_HerCoRe_18E
Label: Europe
Has_Type: Continent
High_P rofile
Flat - Profile
Sea
Steppe
Forest
Peninsula
Subproperty_Of: Position_Rel
Subclass_Of
Mountain
River
Belongs_to
Gulf
Desert
Belongs_t
Area
elongs_to
Mountain _Range
Strait
Field
Small Riverz e
Creeki ze


1_*
Position_Rel
has_a
[Name_Lang]_Label
has_a
Geo_Element
has_a
Cardinal_Position
Is_a
Is_a
Is_a
Area
1_*
has
Ground_Element
Water_Element
1_*
Lays_at
Relief_El
Outside
Inside
ID: C_HerCoRe_50WMaced
Label : Western Macedonia
Has_Object_Property: Western
Subclass_Of
Cardinal_Position
Has_Type: Area
Situated_In: C_HerCoRe_18E
ID: C_HerCoRe_18E
Label: Europe
Has_Type: Continent
High_P rofile
Flat - Profile
Sea
Steppe
Forest
Peninsula
Subproperty_Of: Position_Rel
Subclass_Of
Mountain
Belongs_to
Gulf
Desert
Belongs_t
Area
elongs_to
Mountain _Range
Strait
Field
Small Riverz e
Creeki ze


1_*
Position_Rel
has_a
[Name_Lang]_Label
has_a
Geo_Element
has_a
Cardinal_Position
Is_a
Is_a
Is_a
Area
1_*
has
Ground_Element
Water_Element
1_*
Lays_at
Relief_El
Outside
Inside
ID: C_HerCoRe_55NDobrud
Label : North Dobrudja
Has_Object_Property: North
Subclass_Of
Cardinal_Position
Has_Type: Area
Situated_In: C_HerCoRe_20DB
ID: C_HerCoRe_20DB
Label: Dobrudja
Has_Type: Area
High_P rofile
Flat - Profile
Sea
Steppe
Forest
Peninsula
Subproperty_Of: Position_Rel
Mountain
River SIZEize
Belongs_to
has_a 1 *
Desert
Gulf
Belongs_to
elongs_to
Mountain _Range
Strait
Field
Small Riverz e
Creeki ze



Hist. Context 1

Geo_Element

Area

Situated_In

ID: C_HerCoRe_50MACEd
Label : Western Macedonia

F_TODAY

≈ Time_Slot

Admin_Notion

Id: C_HercoRe_17GR
Label: Greece

Hist. Context 2

Geo_Element

Area

Situated_In

ID: C_HerCoRe_50MACEd
Label : Western Macedonia

Turkish time

≈ Time_Slot

Admin_Notion

Id: C_HercoRe_71 SarfS
Label: *Serfia sangiac"*

Defined_As

Defined_As

Id: C_HercoRe_71 CojR
Label: Cojani Region:

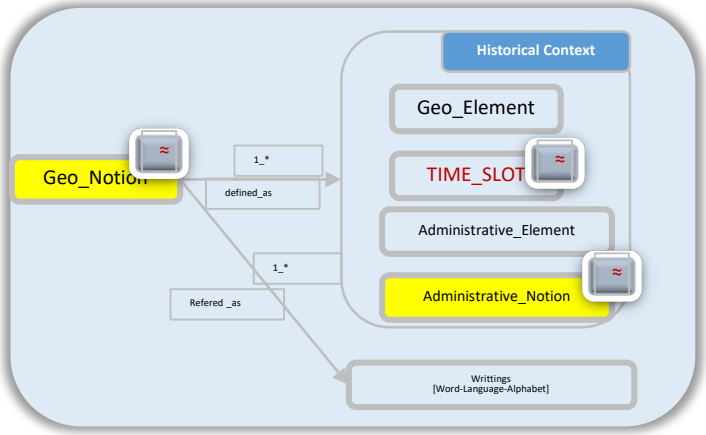
Geo_Notion

Situated_in

Admin_Notion

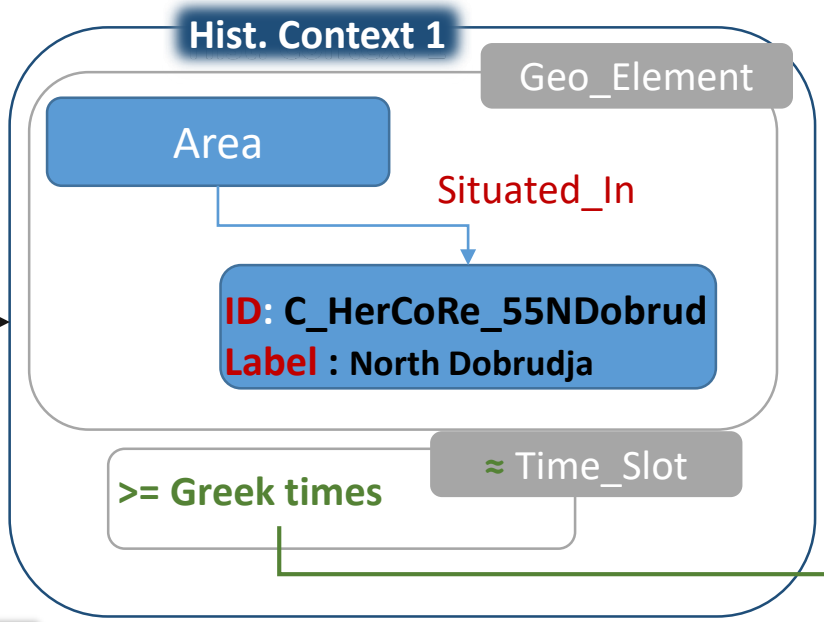
Id: C_HercoRe_74 SarfC
Label: *Σέρβια,Servia*

The Cojani region from western Macedonia, today in Greece but in Turkish times in the "Serfia sangiac" having the capital Σέρβια, Servia ;




```
<DatatypeDefinition>
    <Datatype IRI='#GreekTimes'/> <DataIntersectionOf>
    <DatatypeRestriction>

          <Datatype abbreviatedIRI='xsd:integer'/>
          <FacetRestriction facet='&xsd;minInclusive'>
                  <Literal datatypeIRI='&xsd;integer'>-750</Literal>
           </FacetRestriction>
       </DatatypeRestriction>
       <DatatypeRestrictionOf>
</DatatypeDefinition>
```



Geo_Notion

Defined_As

Id: C_HercoRe_10 DSyrf

Label: *Σύρφοι, Syrphoi*

Part of Northern Dobrudja, coming from the Greek term Σύρφοι --Syrphoi;

```
Class ( Syrfia Annotation
 (fuzzyLabel
    < fuzzyOwl2 fuzzyType =" concept " >
    < Concept type =" weightedSum " >
    < Concept type =" weighted " value ="0.33" base ="C_HercoRe_71CojR " / >
     < Concept type =" weighted " value ="0.33" base =" C_HercoRe_10DSyrf " />
     < Concept type =" weighted " value ="0.33" base =" C_HercoRe_11Srb " />
))
```

Syrfia is
the abandoned name of a region in Eastern Europe, used on historical maps until 17th century, designating

- a part of Northern Dobrudja, coming from the Greek term *Σύρφοι - Syrphoi*, or
- The Cojani region from western Macedonia, today in Greece but in turkish times in the "Serfia sangiac" having the capital *Σέρβια, Servia* ;
- Sârbia, due to phonetic association.

Source: Wikipedia

Geo_Notion

Id: C_HercoRe_70Syrfia
Label: Syrfia
Used_for: Map
<= 17 Century
≈ Time_Slot

Defined_As
Confidence 0.33

Hist. Context 1
Id: C_HercoRe_71 CojR
Label: Cojani Region:

Defined_As
Confidence 0.33

Hist. Context 2
Id: C_HercoRe_10 DSyrf
Label: *Σύρφοι, Syrphoi*

Defined_As
Confidence 0.33

Hist. Context 3
Id: C_HercoRe_11Sarb
Label:*Sarbia*

Orteliusmap 1570

Id: I_C_HerCoRe_Sy
Label: Syrfia_Ortelius
Used_for: Map_Ortelius
Time_slot: 1570

Instance_Of

Syrfia is
the abandoned name of a region in Eastern Europe, used on historical maps until 17th century, designating

- a part of Northern Dobrudja, coming from the Greek term *Σύρφοι - Syrphoi*, or
- The Cojani region from western Macedonia, today in Greece but in turkish times in the "Serfia sangiac" having the capital *Σέρβια, Servia* ;
- Sârbia, due to phonetic association.

Source: Wikipedia

Geo_Notion

Id: C_HercoRe_70Syrfia
Label: Syrfia
Used_for: Map

<= 17 Century
≈ Time_Slot

Defined_As
Confidence 0.2

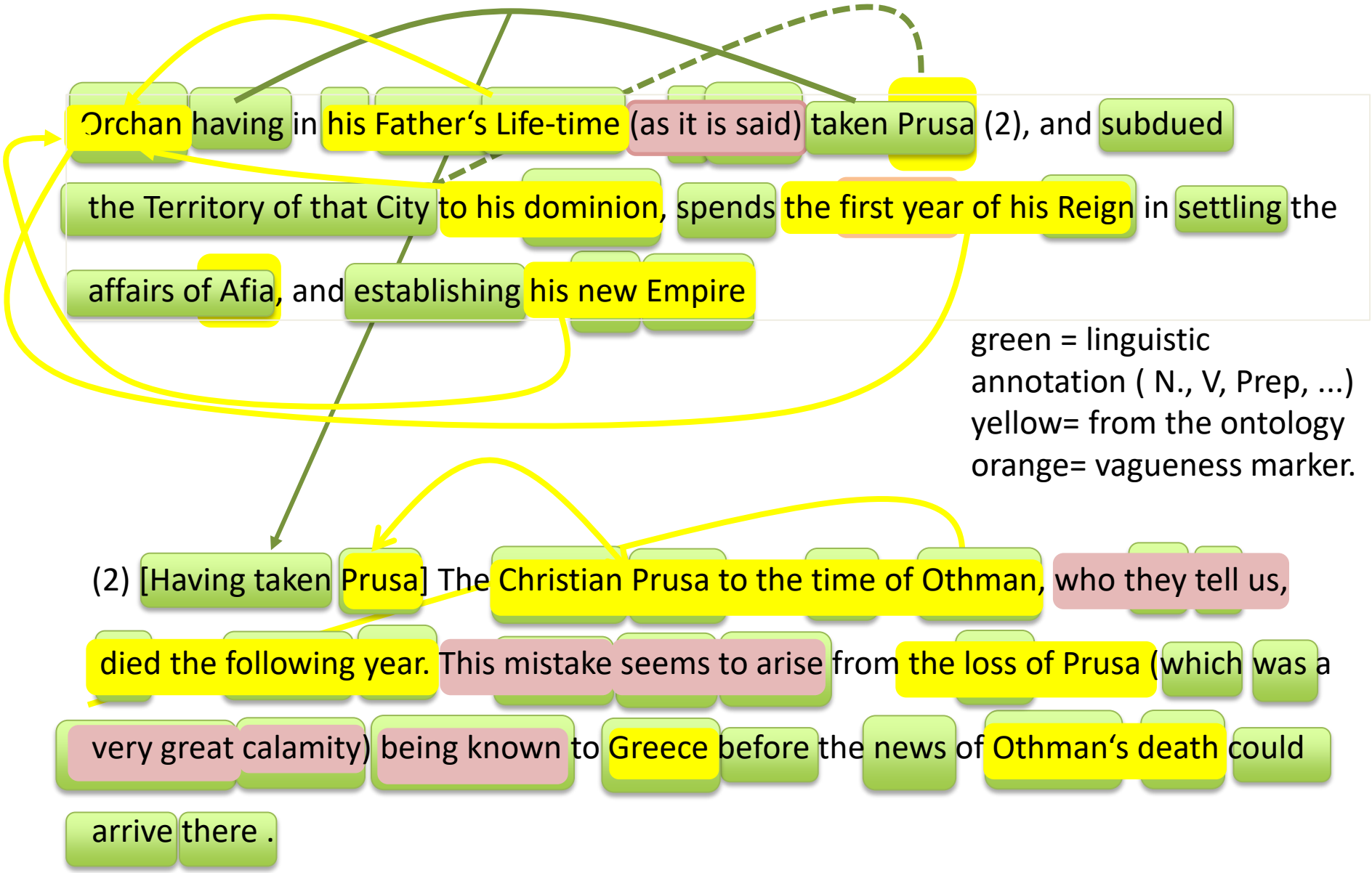
Hist. Context 1
Id: C_HercoRe_71 CojR
Label: Cojani Region:

Defined_As
Confidence 0.7

Hist. Context 2
Id: C_HercoRe_10 DSyrf
Label: *Σύρφοι, Syrphoi*

Defined_As
Confidence 0.1

Hist. Context 3
Id: C_HercoRe_11Sarb
Label: *Sarbia*



History of Growth and Decay Ottoman Empire, English Translation, pag. 24

Annotation-Environment Functionality

- Several Layers of Annotation (e.g. linguistic, editorial, text structure, domain specific).
- Annotation layers are interconnected
- Synchronisation between different text variants (original, translation editorial remarks)
- Discontinuous annotation segments
- Controlled automatisation of manual annotation
- Need of user-friendly annotation interfaces
- Modular Architecture flexible at changes (new layers, new annotation categories)
- Import of automatic annotation (basic linguistic, basic vagueness)

Work ahead

- Semantic Linking between the map and „Geo_notions“
- Formal representation of vagueness markers
- Adaptation of existent fuzzy reasoners
- Specification of user scenarios
- Visualisation of results
- Reuse and enhance the ontology
- Model the Cantemir‘s Network in Istanbul and link it with the documents