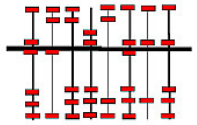
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG



Research Group
„Computerphilology“

Analysing Big Data in Social History

Cristina Vertan

Cristina.Vertan@uni-hamburg.de

What does it mean „big data“ in social history ?

- Usually it is not „big data“ in the classical sense

BUT

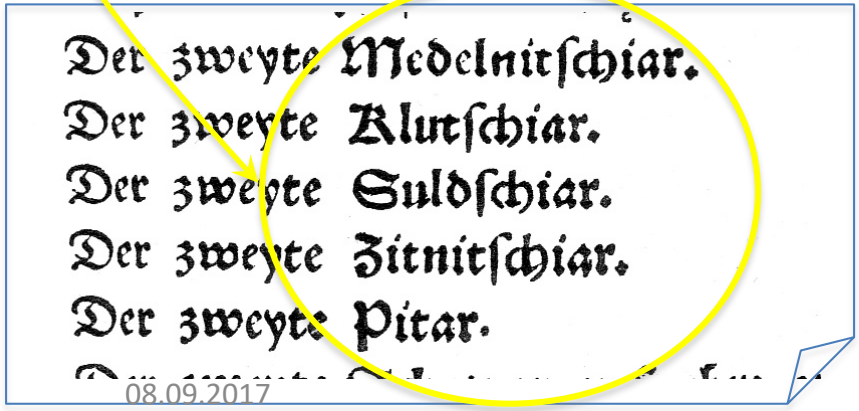
- It is multilingual -> involves a big number of researchers to deal with
- It is frequently geographically distributed.
- Parts of data may be replicated
- Often data models have to go character level, then the size increase fast

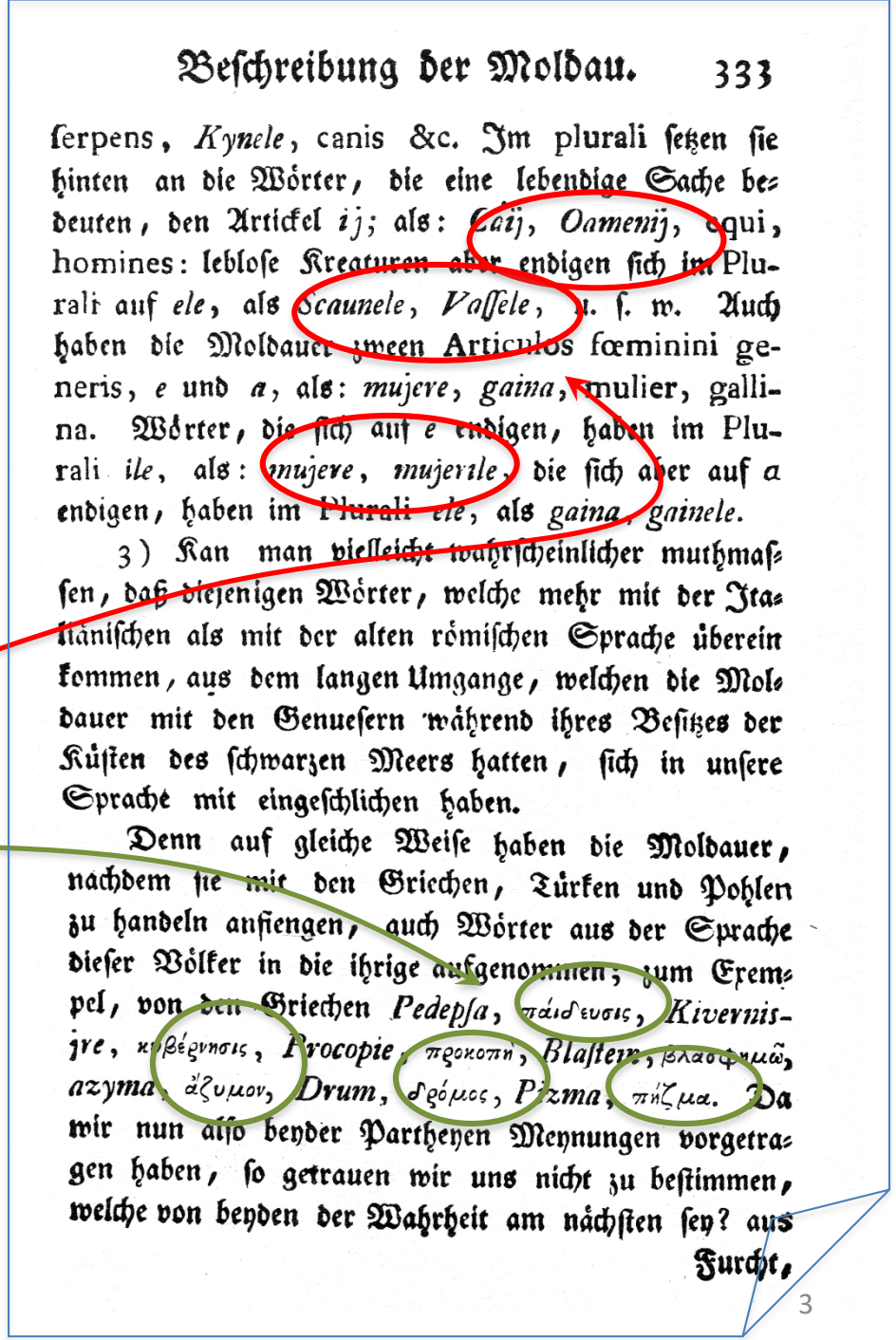
Quantitative measurements may lead to false conclusions as

- data is scattered
- Language is not normalized

Example of data (D.Cantemir- Descriptio Moldaviae)

- Surface form – level
  - German texts are in black-letter typeface
  - Part of the foreign words are transcribed after German rules and represented in black-letter typeface
  - Mixed typefaces
  - Mixed scripts


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* endigen, haben im Plurali *ele*, als *gaina*, *gainele*.

3) Kan man vielleicht wahrſcheinlicher muthmaſſen, daß diejenigen Wörter, welche mehr mit der Italiäniſchen als mit der alten römiſchen Sprache überein kommen, aus dem langen Umgange, welchen die Moldauer 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, *Kivernisjre*, 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,

Processing challenges for historical texts

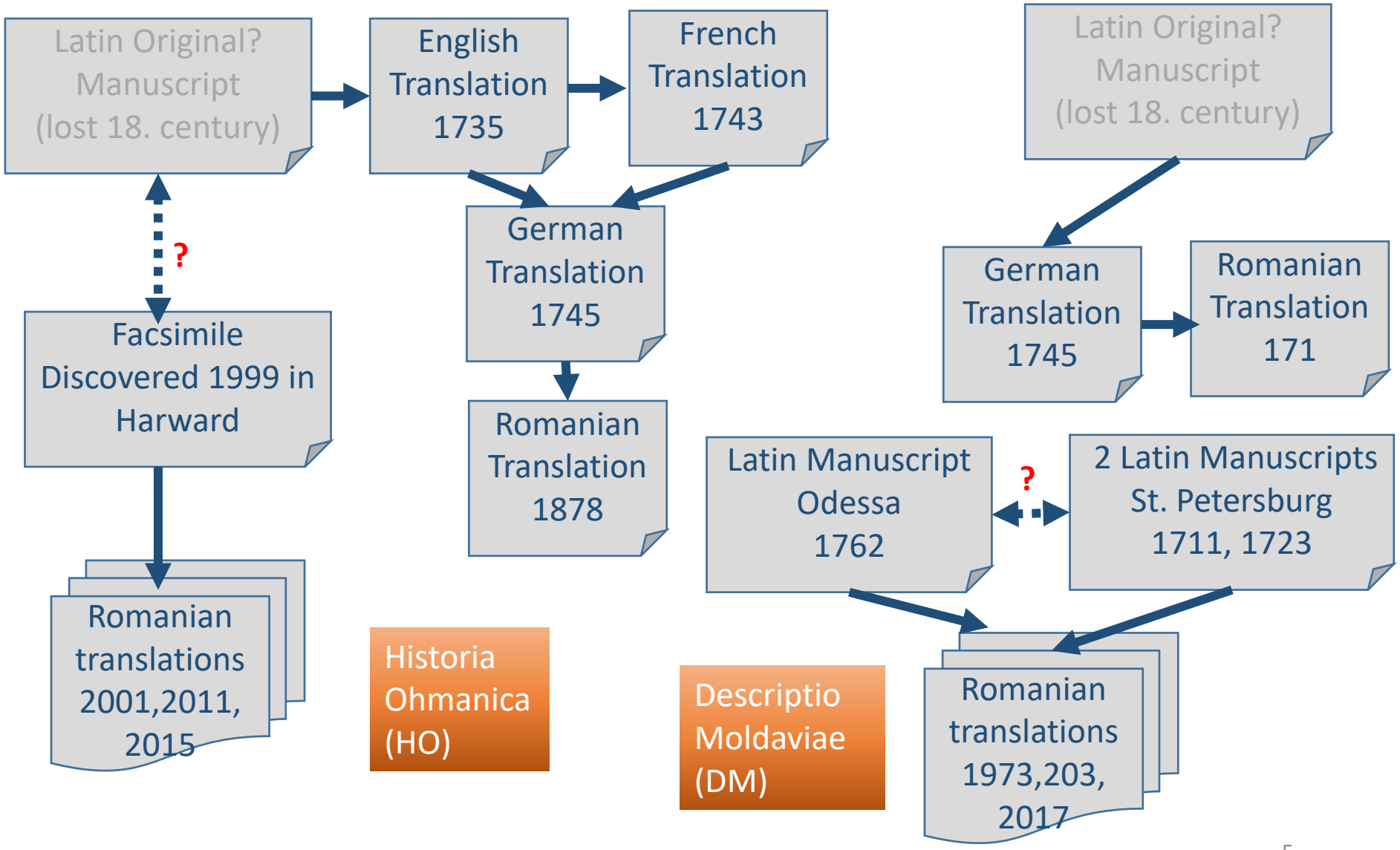
modern Texts

- **Representation**
  - Unicode
- **Language Identification**
  - Data-driven algorithms
- **Algorithms**
  - Highly dependent either on language specific rules or language and style specific data (corpora)
- **NEs**
  - Easier to identify automaticalla
  - Can be annot with geo datata
- **Conceptual**
  - Ontologies available for many domains
  - Automatic extraction of ontologies makes progress
  - Semantic lexical resources available

Historical texts

- **Representation**
  - **Still languages with non covered characters. (use of private zone)**
- **Language Identification**
  - **Problem with short passages (1-2 words)**
- **Algorithms**
  - **Need for new algorithms dealing with multilinguality within one document**
- NEs
  - **Great variance in spelling and names**
  - **Ge locations difficult to be traces**
- **Conceptual level**
  - **Few ontologies,**
  - **Mostly related to metadata information**
  - **Conceptual space different across centuries and regions**
  - **Few semantic resources (Latin Wordnet)**

Manuscripts, editions, translations- example of data replication



Usage of NLP tools

- For diachrone analysis of the language -> analysis of the influence of group interractions, cultural changes etc.
  - Language cannot be normalized to modern versions
  - Deep annotaton is needed: usually proeminent features do not change, while chanages in morphlogical features and syntactical rules may be found
- As preprocessing step, before further analysis is performed.

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

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

CS

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

Progress work in automatic recognition of vague expressions

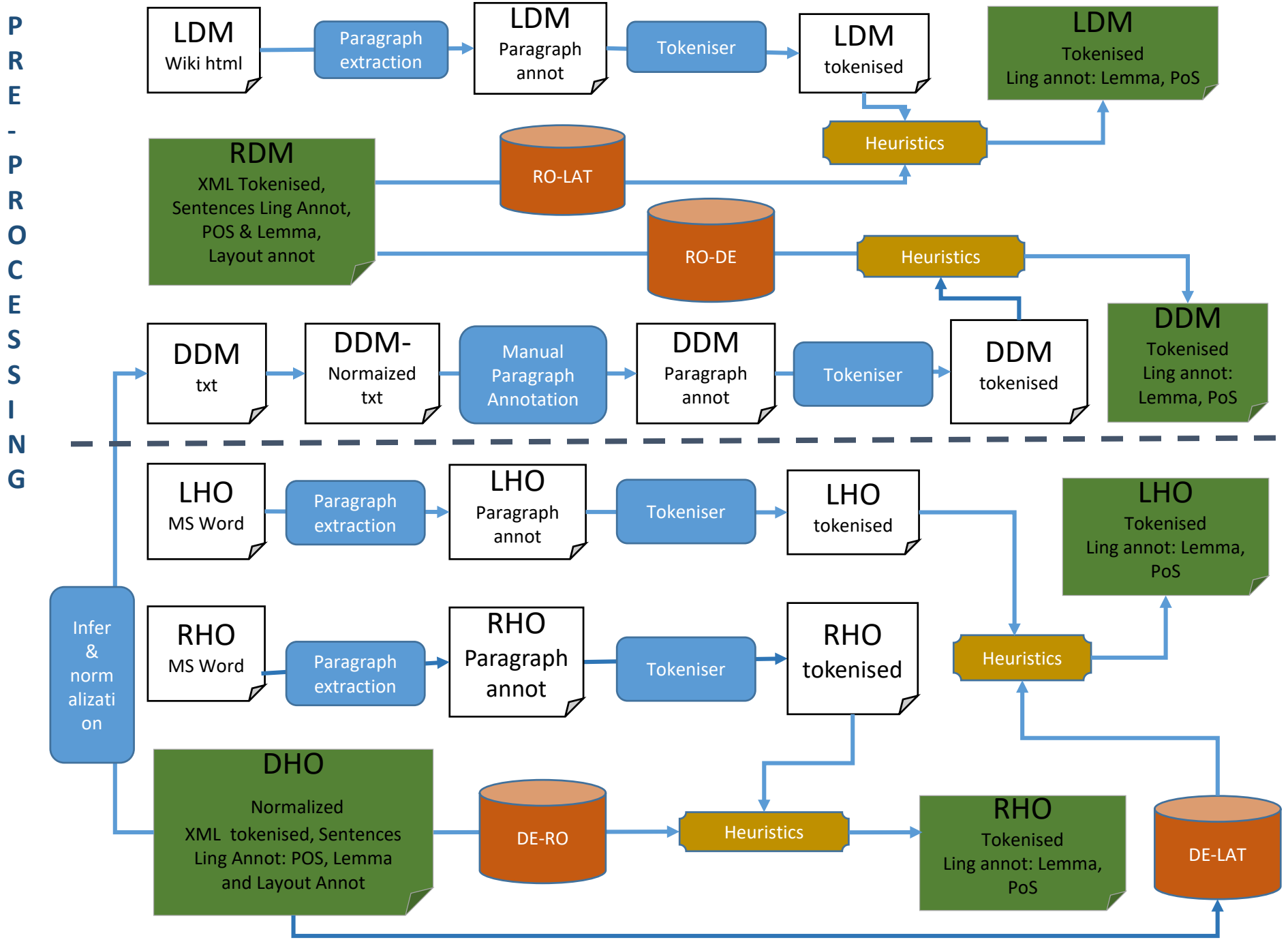
H

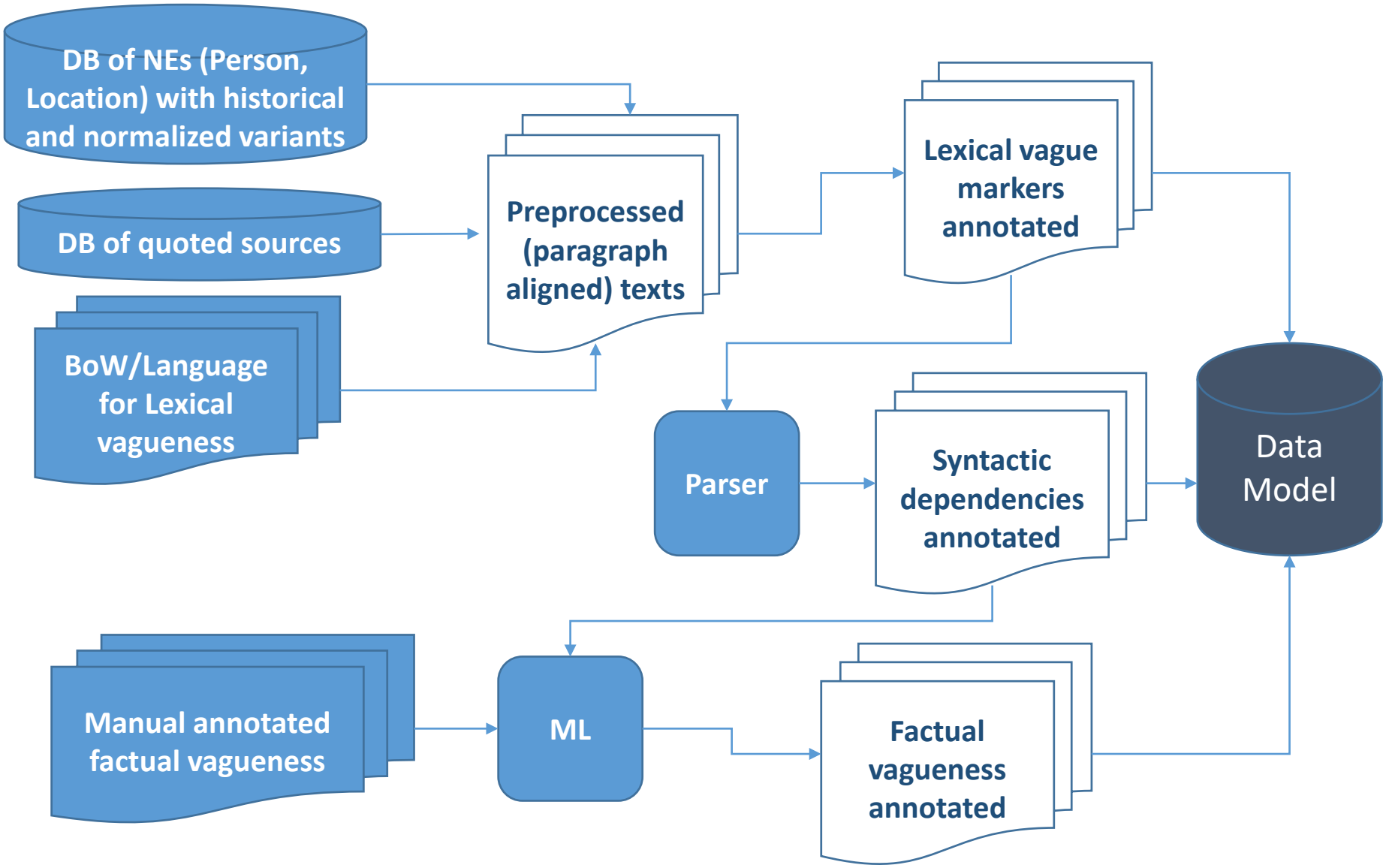
Compare for the first time original with translations done in the 18th,- 19th century

(In)Validate assumptions about source quotations in original text

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


PRE-PROCESSING
LDM
Wiki html
Paragraph extraction
LDM
Paragraph annot
Tokeniser
LDM
tokenised
LDM
Tokenised
Ling annot: Lemma, PoS
Heuristics
RDM
XML Tokenised, Sentences Ling Annot, POS & Lemma, Layout annot
RO-LAT
RO-DE
Heuristics
DDM
txt
DDM-
Normaized txt
Manual Paragraph Annotation
DDM
Paragraph annot
Tokeniser
DDM
tokenised
DDM
Tokenised
Ling annot: Lemma, PoS
LHO
MS Word
Paragraph extraction
LHO
Paragraph annot
Tokeniser
LHO
tokenised
LHO
Tokenised
Ling annot: Lemma, PoS
Infer & norm alizati on
RHO
MS Word
Paragraph extraction
RHO
Paragraph annot
Tokeniser
RHO
tokenised
Heuristics
DHO
Normalized
XML tokenised, Sentences
Ling Annot: POS, Lemma and Layout Annot
DE-RO
Heuristics
RHO
Tokenised
Ling annot: Lemma, PoS
DE-LAT

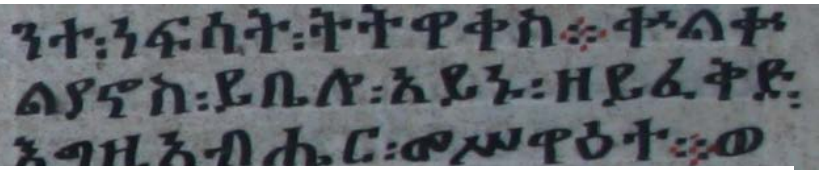

DB of NEs (Person, Location) with historical and normalized variants
DB of quoted sources
BoW/Language for Lexical vagueness
Preprocessed (paragraph aligned) texts
Lexical vague markers annotated
Parser
Syntactic dependencies annotated
Data Model
Manual annotated factual vagueness
ML
Factual vagueness annotated

TraCES - From Translation to Creation: Changes in Ethiopic Style and Lexicon from Late Antiquity to the Middle Ages

- ERC Advanced Grant 2014–2019
- Goal: to perform the first diachronic Analysis of the Classical Ethiopic Language related to Lexicon, Morphology and Style
- Interdisciplinary Approach: apply methods from the Linguistics, Philology and DH.
- Side effect: developpment of first digital Resources for the Classical Ethiopic:
  - A Corpus, annotated at several layers
  - A Lexicon (first digital Lexicon for this language)
  - Tools Corpus-Annotation, -Analysis and
  - Visualisation as well as search in the lexicon.

Challenges of the System

- Computer-linguistic tools for classical semitical languages are rare.
- The few tools which are available, are oriented towards quantitative analysis
- The particularities of Gəʿəz (especially the Script) make impossible any adaptation of the few existent tools (mainly for classical Arabic).
- In contrast with many other aplications which rely on an existent digital lexicon, here the development of lexicon goes in parallel with the construction of the corpus.



Schrifttabelle

	a	u	i	ā	e	ə/ø	o
h	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ
l	ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ
ḥ	ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ
m	መ	ሙ	ሚ	ማ	ሜ	ም	ሞ
ś	ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ
r	ረ	ሩ	ሪ	ራ	ሬ	ር	ሮ
s	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ
q	ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ
b	በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ
t	ተ	ቱ	ቲ	ታ	ቴ	ት	ቶ
ẖ	ኀ	ኁ	ኂ	ኃ	ኄ	ኅ	ኆ
n	ነ	ኑ	ኒ	ና	ኔ	ን	ኖ
ʾ	አ	ኡ	ኢ	ኣ	ኤ	እ	ኦ
k	ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ
w	ወ	ዉ	ዊ	ዋ	ዌ	ው	ዎ
c	ዐ	ዑ	ዒ	ዓ	ዔ	ዕ	ዖ
z	ዘ	ዙ	ዚ	ዛ	ዜ	ዝ	ዞ
y	የ	ዩ	ዪ	ያ	ዬ	ይ	ዮ
d	ደ	ዱ	ዲ	ዳ	ዴ	ድ	ዶ
g	ገ	ጉ	ጊ	ጋ	ጌ	ግ	ጎ
ṭ	ጠ	ጡ	ጢ	ጣ	ጤ	ጥ	ጦ
p	ጰ	ጱ	ጲ	ጳ	ጴ	ጵ	ጶ
ṣ	ጸ	ጹ	ጺ	ጻ	ጼ	ጽ	ጾ
ḍ	ፀ	ፁ	ፂ	ፃ	ፄ	ፅ	ፆ
f	ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ
p	ፐ	ፑ	ፒ	ፓ	ፔ	ፕ	ፖ
q^w	ቈ		ቊ	ቋ	ቌ	ቍ	
$ẖ^w$	ኈ		ኊ	ኋ	ኌ	ኍ	
k^w	ኰ		ኲ	ኳ	ኴ	ኵ	
g^w	ጐ		ጒ	ጓ	ጔ	ጕ	

The Classical Ethiopic (Gəʿəz)-1

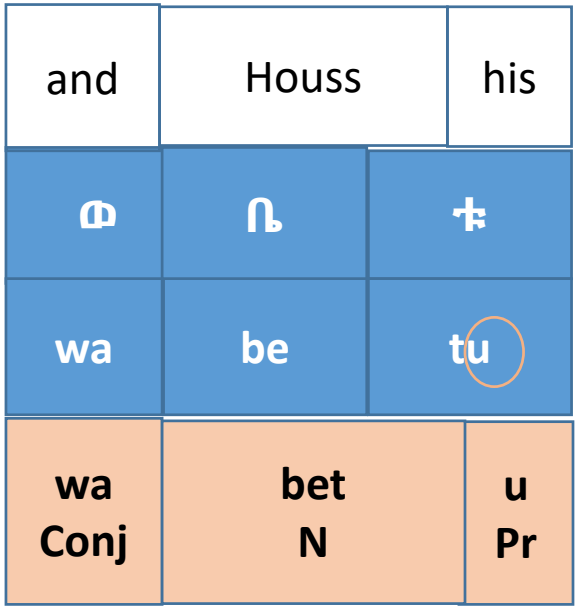
- Each symbol represent a syllable (Fidäl)
- Wowels are represented.
- Rechtsläufig
- Dedicated symbols for Digits
  - For a number representation and value fora number are different
  - ፩ወ፭ Transcription 1 wa 5
    Value 6 (1+5)
- Dedicated interpunction፡ (Space) ። (. or ,) ፣ (,) ፤ (:) ፥ (;) ፧ (?) ፨ (Paragraph)
- Laryngals and Sibilants are changeblae from the phonemisc and graphemic point of view.
  - ሃሌሉያ (hāleluyā)
  - Uሌሉያ ሐሌሉያ ኀሌሉያ ሓሌሉያ
  - ኃሌሉያ
- Nonconcatenative Morphology

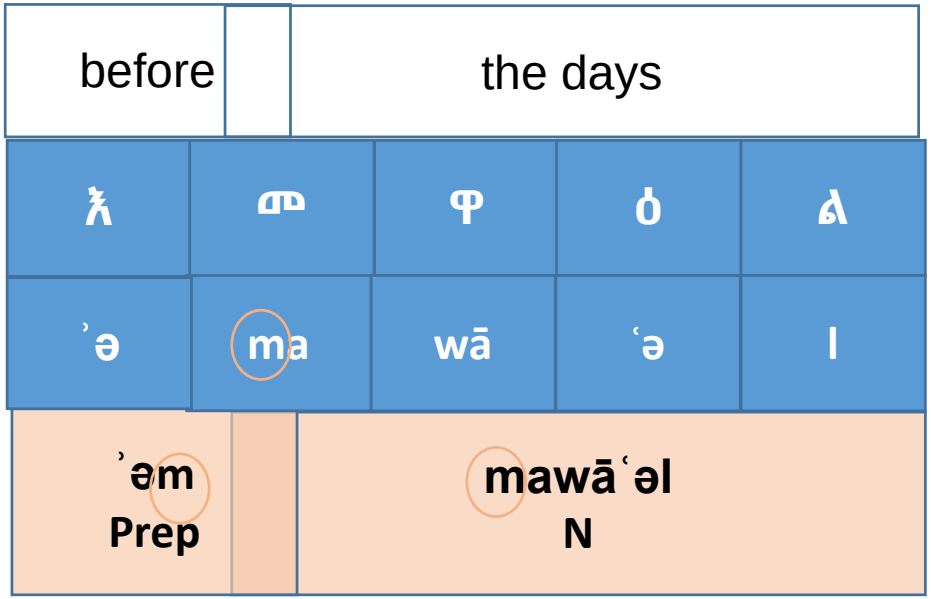
√qds × ʾasta1a22a3a → ʾastaqaddasa

The Classical Ethiopic(Gəʿəz) – 2

- Vowels can be themelves part-of-speech

Letter compression in Originalscript (Fidäl), but not in the transcription

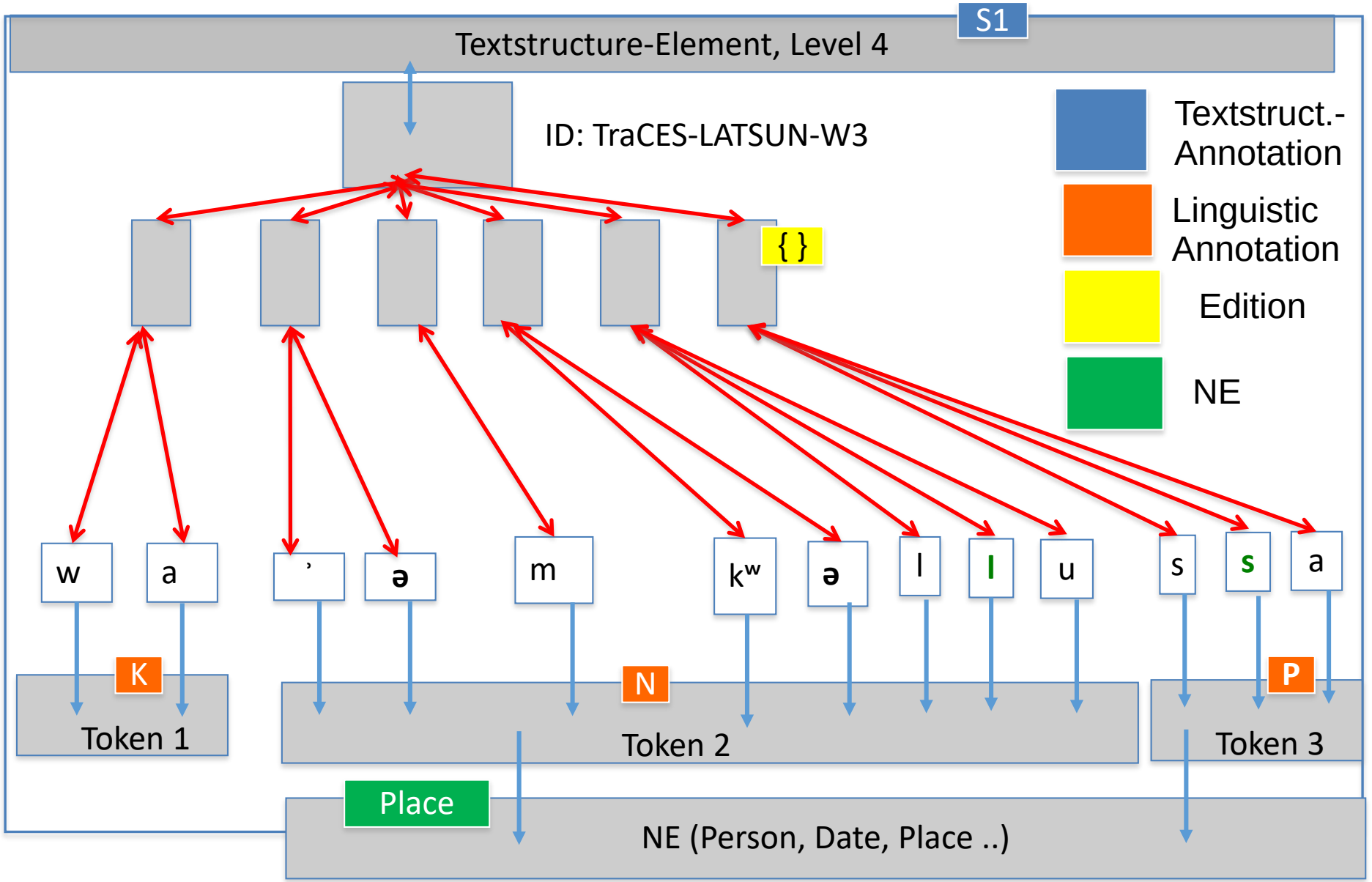




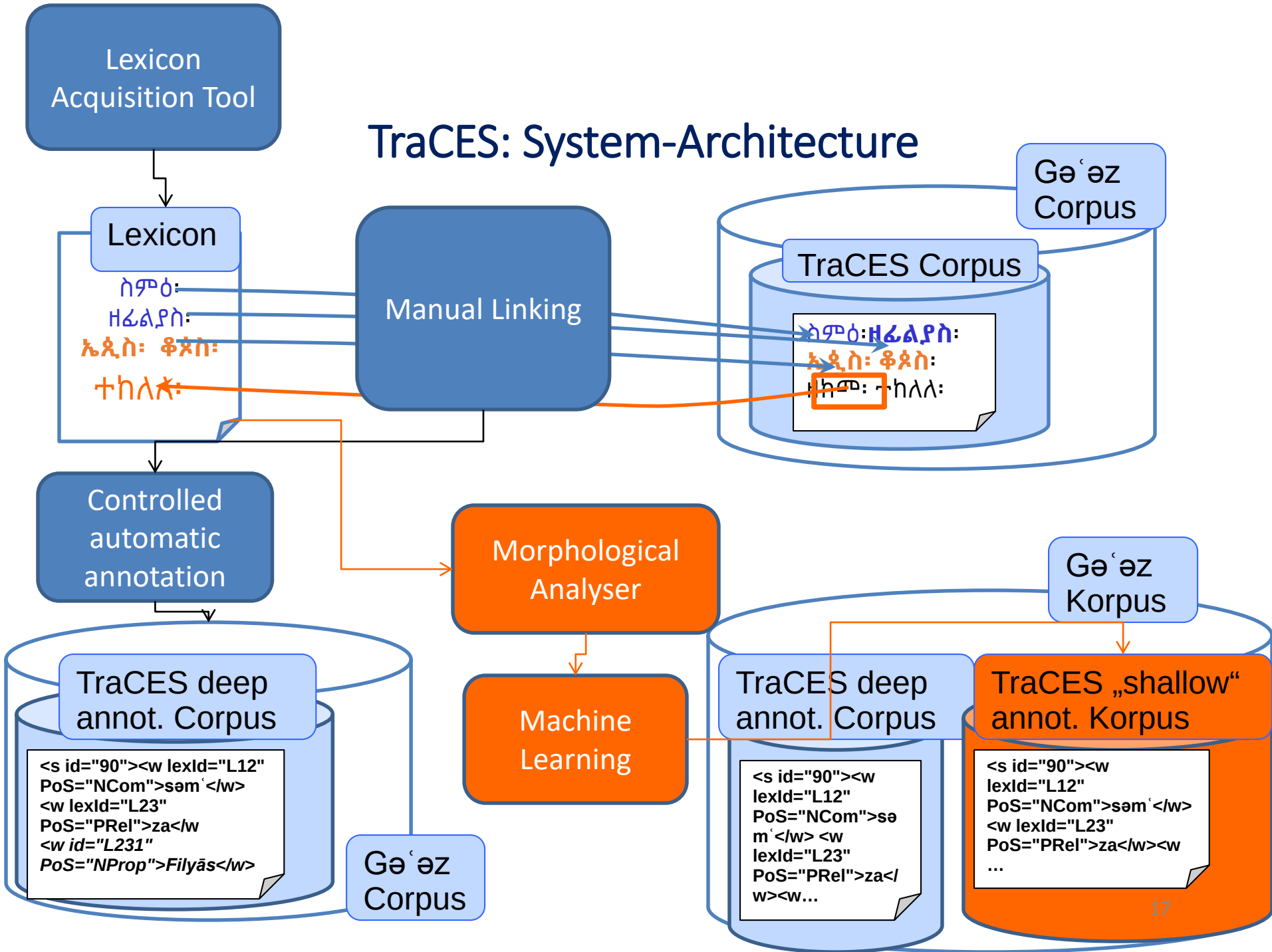
Challenges for the Annotation

- The annotation cannot be done on the *Fidäl*-Script, as one syllable can contain several PoS
- Necessary steps:
  - Automatic Transliteration
  - Error correction in der Transliteration DURING the (linguistic) Annotation
  - Transliteration and *Fidäl*-Text must be synchronised
  - Controlled semi-automatic Annotation for:
    - Speed-up of the annotation process
    - Consistency
- Good grammar description: Dillmann (1899), Troper (2002), Weniger (2001)
- There is no formal representation of Gəʿəz Grammatik → Development of aTag-set for mophological annotation is necessary
- The annotation on several leyers should be possible amy time.

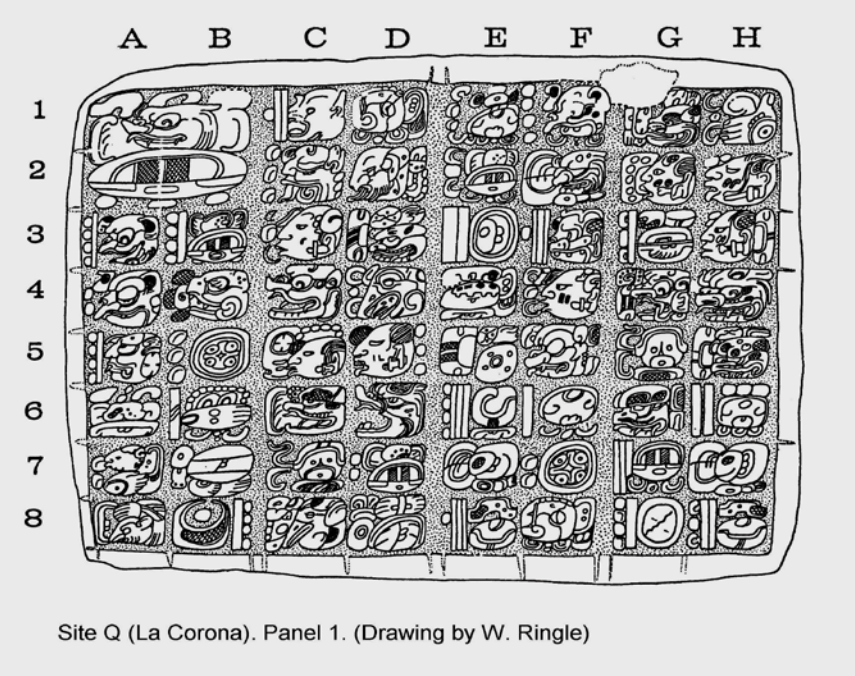
• TraCES-Annotation: Data-Model



TraCES: System-Architecture



One Document several interpretations



Site Q (La Corona). Panel 1. (Drawing by W. Ringle)

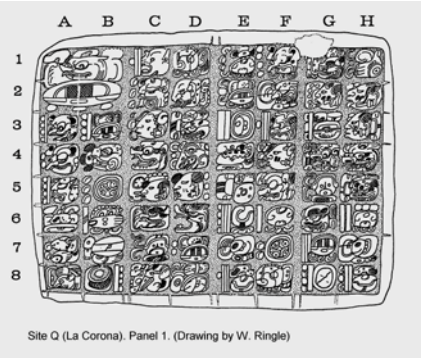
http://mayawoerterbuch.de/

Text Datenbank und Wörterbuch des klassischen Maya

Each block may have several interpretations

Each interpretation generates other annotations

Document



Id: D1 **Dokument**

Koord -Label A1-B2 A3 D1 D8 E4

Num -Label
124vt:[28bh°738st]:548bt:74bb
5009st.1033st
[740st:126st].683br
58st.[74tt:564st]
[5st:201st:23st].683br

Id: D1-G1 **GU**

Koord -Label A1-B2

Num -Label
124vt:[28bh°738st]:548bt:74bb

Num -Label _Alt1
124:[28°738]:548:74 ...

Graph-Label1 **Graph-Label1_Alt 1** ...

...

Id: D1-G2 **GU**

Koord –Label A3

Num -Label
[740st:126st].683br

Num –Label_Alt1
[740:126].683b ...

Graph-Label1 ...

Id: D1-G3 **GU**

Koord -Label D1

Num -Label
[740st:126st].683br

Num –Label_Alt1
[740:126].683 ...

Graph-Label1 ...

Id: D1-G4 **GU**

Koord -Label D8

Num -Label
[58st.[74tt:564st]

Num -Label
[58.[74:564] ...

Graph-Label1 ...

Id: D1-G5 **GU**

Koord -Label E4

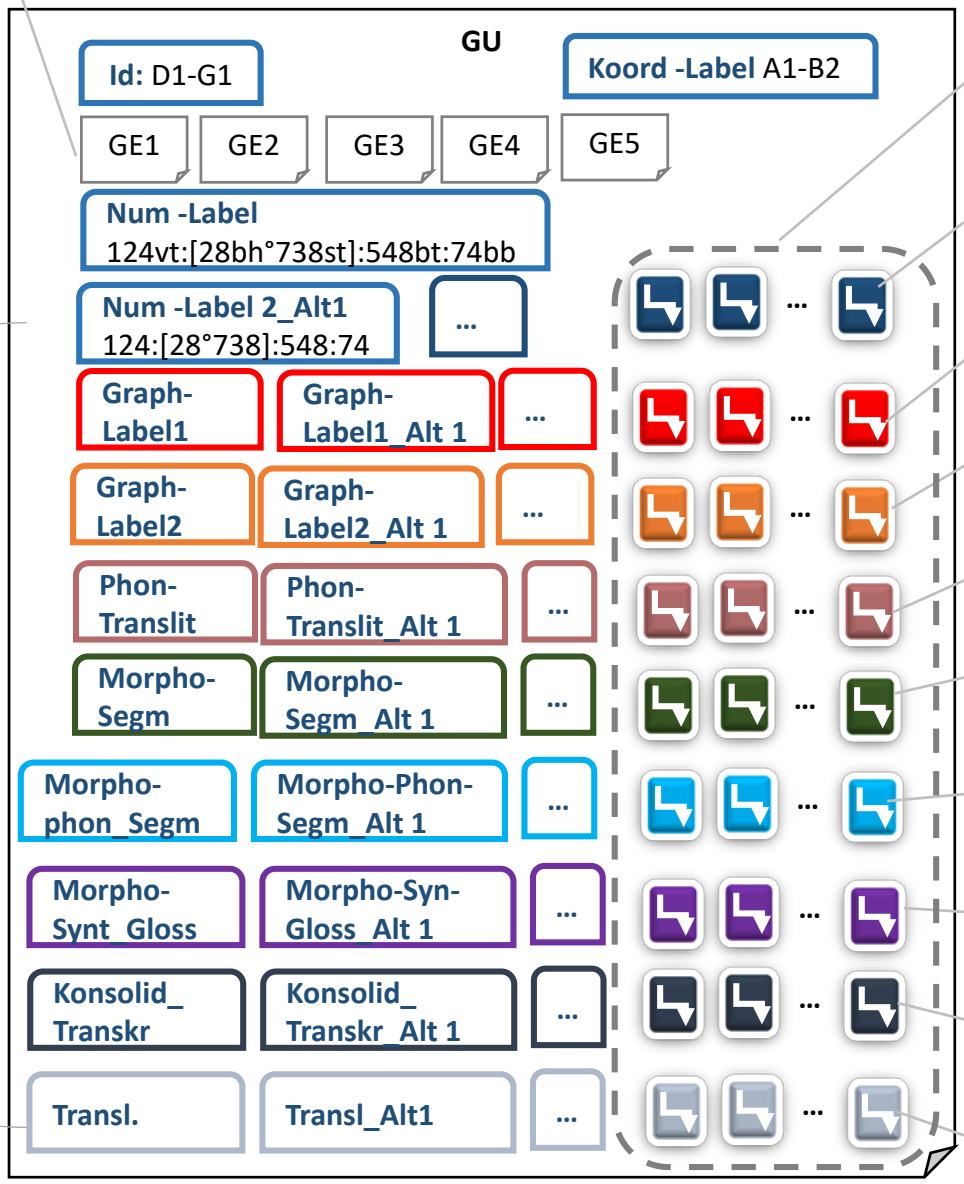
Num -Label
[5st:201st:23st].683br

Num -Label
[5:201:23].683 ...

Graph-Label1 ...

Graphical Unit

GEs automatisch aus Num-Label generiert (s.für detaliierte Struktur nächste Folie)



- Pointers sind am Anfang NULL und werden nach einer Annotation zugefügt
- Pointers zu einzelnen Num_Translit2 Annotation Span
- Pointers zu einzelnen Graphemischen Translit 1 Annotation Span
- Pointers zu einzelnen Graphemischen Translit 2 Annotation Span
- Pointers zu einzelnen Phonemische Translit Annotation Span
- Pointers zu einzelnen Morphologische segmentierte Transkription Annotation Span
- Pointers zu einzelnen morphonemisch Annotation Span
- Pointers zu einzelnen morphosynt Glossierung Annotation Span
- Pointers zu einzelnenkonsolidierte Transkription Annotation Span
- Pointers zu einzelnen Übersetzung Annotation Span

Labels werden nicht gespeichert sondern von den einzelnen Annotationen beim Bedarf zusammengesetzt

Conclusions

- Data modelling is an essential step and should be the first one taken.
- Models shoul reflect the data particularities and fulfill user requirements
- If no available tool fulfills the requirments, new development should not scare
- „Traditional“ Text minning algorithms do not perform the same way on historical data

As long as one uses black-boxes there is no way to control quality of the output (golden standard is usually missing)

TraCES Team

Prof. Dr. Alessandro Bausi
(Principal investigator)
- Philologiy and Linguistics
- Critical Text Edition
- Manuscript cultures

Susanne Hummel
- Philology
- Linguistics Analysis
- Korpus-Annotation

Wolfgang Dickhut
- Philology
- Linguistic Analysis
- Corpus-Annotation

Daria Elagina
- Ph. D. student
- OCR
- Corpus-Annotation

Hiruie Ermias
- Ph. D. Student
- corpus-Annotation

Dr.des Andreas Ellwardt
- Classical Ethiopic Lexikography
- Linguistic Analysis
- Translationof classical Ethiopic from Arabic

Dr. Vitagrazia Pisani
- Philology
- Linguistic Analysis
- Corpus-Annotation

Dr. Cristina Vertan
- Digital Humanities
- Computerlinguistics
- Computer Science

Eugenia Sokolinski
- Project-
- Coordination
- Web-Design
- Metadata & Indexing

HerCoRe Team

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