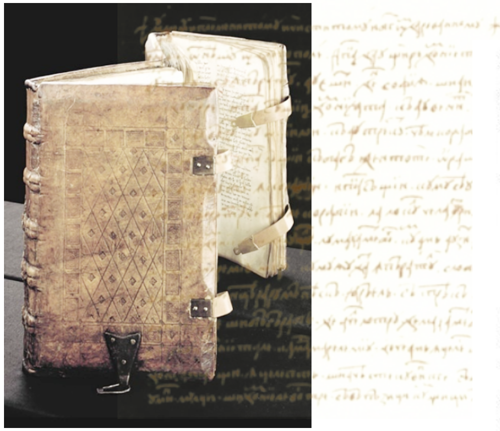
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Research Group
„Computerphilology“

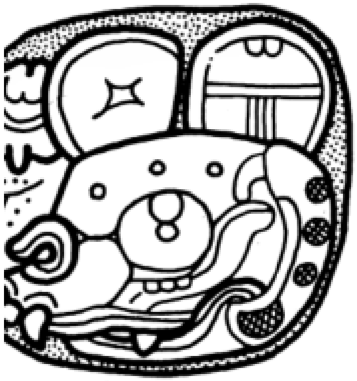
Generating Complex Big Data through Annotations in Digital Humanities

Cristina Vertan

cristina.Vertan@uni-hamburg.de



እኍሃ፡ለእሙ፡ወሰዓማ፡ያዕቆብ፡
ለራሔል፨ወጸርኅ፡በቃሉ፡ወበ
ከዩ፡ወዓድአ፡ለራሔል፡ከመ፡
ወልደ፡እኅቱ፡ለላባ፡ውእቱ፡ወ
ከመ፡ወልደ፡ርብቃ፡ውእቱ፨
ወሮጸት፡ራሔል፡ወአይድዓቶ
ለአቡሃ፡ዘንተ፡ነገረ፨ወሶበ፡
ሰምዓ፡ላባ፡ቃለ፡ስሙ፡ያዕቆብ፡ወ
ልደ፡ርብቃ፡እኅቱ፡ሮጸ፡ወተቀ
በሎ፡ወሐቀፎ፡ወሰዓሞ፡ወወሰ
ዶ፡ቤቶ፨ወነገሮ፡ለላባ፡ኵሎ፡
ዘንተ፡ነገረ፨ወይቤሎ፡ላባ፡ለ
ያዕቆብ፡እምነ፡ዐጽምየ፡ወእም
ነ፡ሥጋየ፡አንተ፨ወነበረ፡ምስሌ
ሁ፡ሠላሳ፡መዋዕለ፨፨
ወይቤሎ፡ላባ፡ለያዕቆብ፡እ
ስመ፡እኁየ፡አንተ፡ኢትትቀነይ፡
ሊተ፡በከ፡ንግረኒ፡ዐስበከ፡ም
ንት፡ውእቱ፨ወቦቱ፡ለላባ፡ክል
ኤ፡አዋልድ፡ስማ፡ለእንተ፡ትል

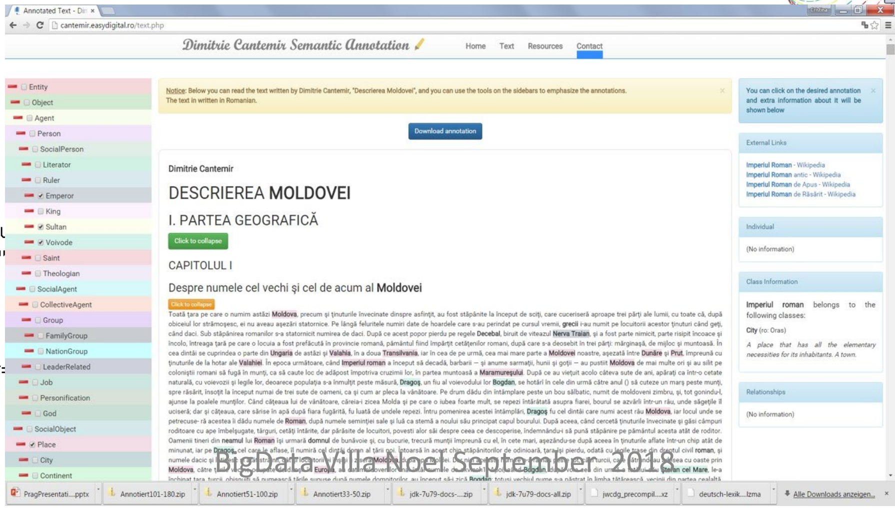


ſerpens, *Kynele*, canis &c. Im plurali ſetzen ſie hinten an die Wörter, die eine lebendige Sache bedeuten, den Artickel *ij*; als: *Caij*, *Oamenij*, equi, homines: lebloſe Kreaturen aber endigen ſich im Plurali auf *ele*, als *Scaunele*, *Vaſſele*, u. ſ. w. Auch haben die Moldauer zween Articulos fœminini 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, *Kivernisire*, κυβέρνησις, *Procopie*, προκοπὴ, *Blaſtem*, βλασφημῶ, *azyma*, ἄζυμον, *Drum*, δρόμος, *Pizma*, πῆζμα. Da wir nun alſo beyder Parthehen Meynungen vorgetragen haben, ſo getrauen wir uns nicht zu beſtimmen, welche von beyden der Wahrheit am nächſten ſey? aus Furcht,

```
<phrase>
     <ne> <person personT="f“ first=„Dragosch“
          relationT=„son“ refT= „w123“/> </ne>
   <phraseParts constituentT="NP" mult="false">
  <tok idT="" langT="ro" styleTranscr="de">
     <pos posT="N"/>
     <string>  Dragosch</string>
   </tok></phraseParts>
</phrase>
  , ein Sohn ihres
 <phrase>
     <ne> <person personT="tl"/></ne>
   <phraseParts constituentT="NP" mult="tru
     <tok idT="" langT="de" styleTranscr="de"
     <pos posT="N"/>
      <string> Königs  </string>
    </tok>
     <tok idT=“w123" langT="ro" styleTranscr=
     <pos posT="N"/>
     <string> Bogdan</string>
   </tok> </phraseParts>
</phrase>
```



Dimitrie Cantemir Semantic Annotation

- ☐ Entity
- ☐ Object
- ☐ Agent
- ☐ Person
- ☐ SocialPerson
- ☐ Literator
- ☐ Ruler
- ☑ Emperor
- ☐ King
- ☑ Sultan
- ☑ Voivode
- ☐ Saint
- ☐ Theologian
- ☐ SocialAgent
- ☐ CollectiveAgent
- ☐ Group
- ☐ FamilyGroup
- ☐ NationGroup
- ☐ LeaderRelated
- ☐ Job
- ☐ Personification
- ☐ God
- ☐ SocialObject
- ☑ Place
- ☐ City
- ☐ Continent

Notice: Below you can read the text written by Dimitrie Cantemir, "Descrierea Moldovei", and you can use the tools on the sidebars to emphasize the annotations. The text in written in Romanian.

Dimitrie Cantemir

DESCRIEREA **MOLDOVEI**

I. PARTEA GEOGRAFICĂ

CAPITOLUL I

Despre numele cel vechi şi cel de acum al **Moldovei**

Toată ţara pe care o numim astăzi **Moldova**, precum şi ţinuturile învecinate dinspre asfinţit, au fost stăpânite la început de sciţi, care cuceriseră aproape trei părţi ale lumii, cu toate că, după obiceiul lor strămoşesc, ei nu aveau aşezări statornice. Pe lângă feluritele numiri date de hoardele care s-au perindat pe cursul vremii, **grecii** i-au numit pe locuitorii acestor ţinuturi când geţi, când daci. Sub stăpânirea romanilor s-a statornicit numirea de daci. După ce acest popor pierdu pe regele **Decebal**, biruit de viteazul **Nerva Traian**, şi a fost parte nimicit, parte risipit încoace şi încolo, întreaga ţară pe care o locuia a fost prefăcută în provincie romană, pământul fiind împărţit cetăţenilor romani, după care s-a deosebit în trei părţi: mărginaşă, de mijloc şi muntoasă. În cea dintâi se cuprindea o parte din **Ungaria** de astăzi şi **Valahia**, în a doua **Transilvania**, iar în cea de pe urmă, cea mai mare parte a **Moldovei** noastre, aşezată între **Dunăre** şi **Prut**, împreună cu ţinuturile de la hotar ale **Valahiei**. În epoca următoare, când **Imperiul roman** a început să decadă, barbarii — şi anume sarmaţii, hunii şi goţii — au pustiit **Moldova** de mai multe ori şi au silit pe coloniştii romani să fugă în munţi, ca să caute loc de adăpost împotriva cruzimii lor, în partea muntoasă a **Maramureşului**. După ce au vieţuit acolo câteva sute de ani, apăraţi ca într-o cetate naturală, cu voievozii şi legile lor, deoarece populaţia s-a înmulţit peste măsură, **Dragoş**, un fiu al voievodului lor **Bogdan**, se hotărî în cele din urmă către anul () să cuteze un marş peste munţi, spre răsărit, însoţit la început numai de trei sute de oameni, ca şi cum ar pleca la vânătoare. Pe drum dădu din întâmplare peste un bou sălbatic, numit de moldoveni zimbru, şi, tot gonindu-l, ajunse la poalele munţilor. Când căţeaua lui de vânătoare, căreia-i zicea Molda şi pe care o iubea foarte mult, se repezi întărâtată asupra fiarei, bourul se azvârli într-un râu, unde săgeţile îl uciseră; dar şi căţeaua, care sărise în apă după fiara fugărită, fu luată de undele repezi. Întru pomenirea acestei întâmplări, **Dragoş** fu cel dintâi care numi acest râu **Moldova**, iar locul unde se petrecuse- ră acestea îi dădu numele de **Roman**, după numele seminţiei sale şi luă ca stemă a noului său principat capul bourului. După aceea, când cercetă ţinuturile învecinate şi găsi câmpuri roditoare cu ape îmbelşugate, târguri, cetăţi întărite, dar părăsite de locuitori, povesti alor săi despre ceea ce descoperise, îndemnându-i să pună stăpânire pe pământul acesta atât de roditor. Oamenii tineri din **neamul** lui **Roman** îşi urmară **domnul** de bunăvoie şi, cu bucurie, trecură munţii împreună cu el, în cete mari, aşezându-se după aceea în ţinuturile aflate într-un chip atât de minunat, iar pe **Dragoş**, cel care le aflase, îl numiră cel dintâi domn al ţării noi. Întoarsă în acest chip stăpânitorilor de odinioară, ţara îşi pierdu, odată cu legile trase din dreptul civil **roman**, şi numele dacic şi [illegible] străini, [illegible] locuitorii ei înşişi [illegible] zisera **Moldova** după apa Moldei. [illegible] turcii, ca [illegible] cu oaste prin **Moldova**, către [illegible] **Europa**, [illegible] **Bogdan**, [illegible] **Ştefan cel Mare**, le-a închinat ţara, turcii, obişnuiţi să numească ţările supuse după numele domnitorilor, au început să-i zică **Bogdan**; totuşi vechiul nume s-a păstrat în limba tătărească, vecinii din partea cealaltă

You can click on the desired annotation and extra information about it will be shown below

External Links

Imperiul Roman - Wikipedia
Imperiul Roman antic - Wikipedia
Imperiul Roman de Apus - Wikipedia
Imperiul Roman de Răsărit - Wikipedia

Individual

(No information)

Class Information

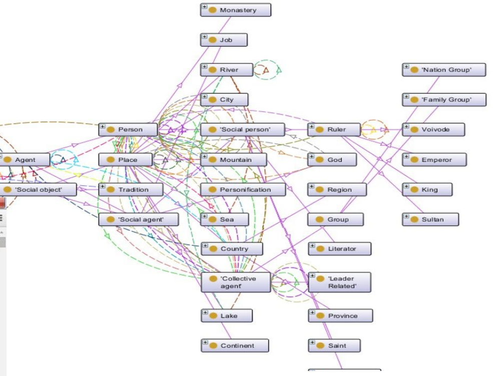
Imperiul roman belongs to the following classes:

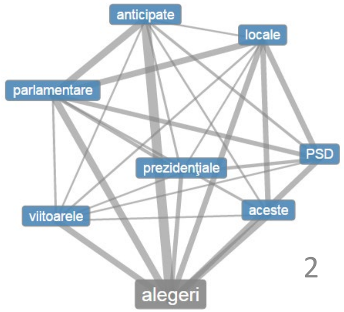
City (ro: Oras)

A place that has all the elementary necessities for its inhabitants. A town.

Relationships

(No information)





Raw vs. Annotated Data

- Raw – Data does not provide any explicit information to computer. ML Algorithms try to infer relations between data through statistical methods. If these relations really exist, or it is just a correlation without a really scientific motivation behind, is often not clear.
- Raw Data is cheap, does not need preprocessing.
- Annotated Data tries to supply Computer with domain knowledge so that inferences are grounded.
- Automatic Annotation introduces errors
- Manual Annotation is expensive and impossible for large amount of data

Big Data or Big Generated Data?

- Historical Texts (until 18th century) do not constitute massive big raw data as often:
  - Language changed so much over century that we cannot analyse together a text from 13th and 18th century
  - Even if we try ML algorithms will fail (e.g. Romanian Cyrillic vs. modern transcriptions, German complete change of language)
  - Many languages have overall few testimonies
- As no statistical correlations can be inferred annotations are required and
  - Annotations generate new data
- Usually applications on such texts are demanded by specialists who need a lot of annotations in order to validate (generate new) scientific hypothesis
- Annotation ≠ Linguistic Annotation (PoS) but also: domain specific, annotations of further copists, translators , editors or even author's
- Annotation ≠ Word Annotation but also: Sentence, Discourse-Entity, particular text units

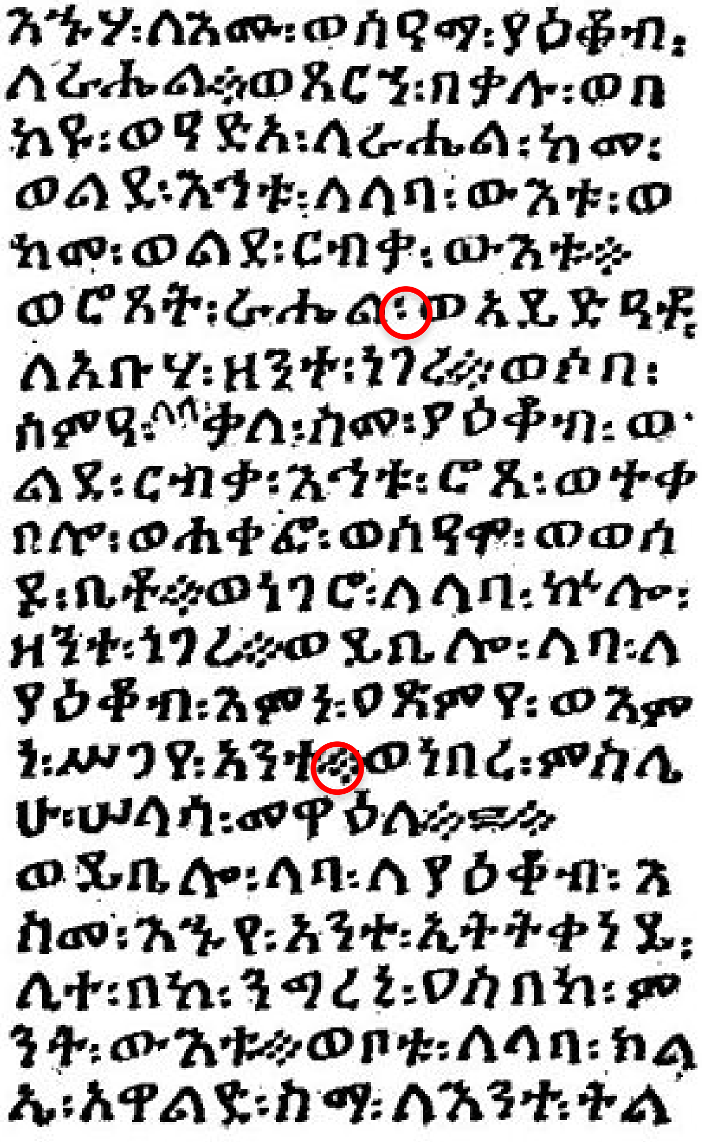
Particularities of the Annotation Process for historical texts

- Often several Layers of Annotation (e.g. linguistic, editorial, text structure, domain specific). Annotation layers are sometimes interconnected
- Sometimes synchronisation between different text variants (e.g. original, transliteration, translation)
- Non-continuous annotation segments
- Changes on the base text during the annotation required
- Often more linguistic categories as for modern data
- Need of user-friendly annotation interfaces
- Modular Architecture flexible at changes (new layers, new annotation categories)
- Often need of manual annotation

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

- ERC Advanced Grant 2014–2019
- **aim**: reliable and extensive linguistic **data** based on annotated texts for a **diachronic analysis** of classical Ethiopic (Gəʿz) (lexicography, morphology and style)
- **corpus**: several texts belonging to **different periods and genres** of Ethiopic literature (**text-critical editions)**

Initial Idea: Linguistic annotation similar with British National Corpus (each token=string sepparated by spaces, receives a PoS)



እኁሃ፡ለእሙ፡ወሰዐማ፡ያዕቆብ፡
ለራሔል፨ወጸርኀ፡በቃሉ፡ወበ
ከየ፡ወዋድአ፡ለራሔል፡ከመ፡
ወልደ፡እኅቱ፡ለላባ፡ውእቱ፡ወ
ከመ፡ወልደ፡ርብቃ፡ውእቱ፨
ወሮጸት፡ራሔል፡ወአይድዓቶ
ለአቡሃ፡ዘንተ፡ነገረ፨ወሶበ፡
ሰምዓ፡ላባ ቃለ፡ስሙ፡ያዕቆብ፡ወ
ልደ፡ርብቃ፡እኅቱ፡ሮጸ፡ወተቀ
በሎ፡ወሐቀፎ፡ወሰዐሞ፡ወወሰ
ዶ፡ቤቶ፨ወነገሮ፡ለላባ፡ኵሎ፡
ዘንተ፡ነገረ፨ወይቤሎ፡ላባ፡ለ
ያዕቆብ፡እምኔ፡ዐጽምየ፡ወእም
ኔ፡ሥጋየ፡አንተ፨ወነበረ፡ምስሌ
ሁ፡ሠላሳ፡መዋዕለ፨፨
ወይቤሎ፡ላባ፡ለያዕቆብ፡እ
ስመ፡እኁየ፡አንተ፡ኢትትቀነይ፡
ሊተ፡በከ፡ንግረኒ፡ዐስበከ፡ም
ንት፡ውእቱ፨ወቦቱ፡ለላባ፡ክል
ኤ፡አዋልደ፡ስማ፡ለእንቲ፡ትል

Language particularities

- Vowels can be independent part-of-speech

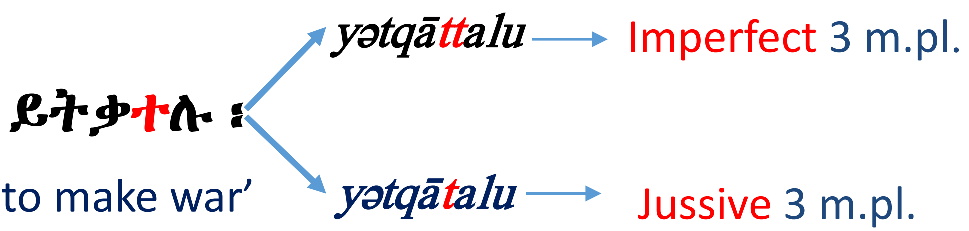
and	House	his
ወ	ቤ	ቱ
wa	be	tu
wa Conj	bet N	u Pr

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

before		the days		
እ	መ	ዋ	ዕ	ል
ʾə	ma	wā	ʿə	l
ʾəm Prep		mawāʿəl N		

Transcription vs. Transliteration

Gemination of a consonant



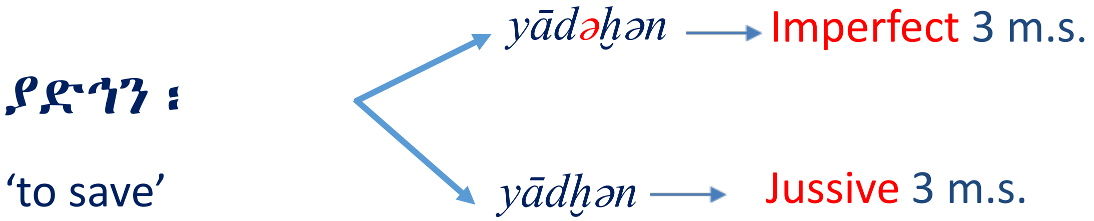
ይትቃተሉ ፡

‘to make war’

yətqāttalu → Imperfect 3 m.pl.

yətqātalu → Jussive 3 m.pl.

Disambiguation of the vowel -*ə*



ያድኅን ፡

‘to save’

yādəḫən → Imperfect 3 m.s.

yādḫən → Jussive 3 m.s.

ገብሩ ፡ same Gəʿəz forms with

different meanings different no. tokens

gabru ‘they did’ or *gabr*-*u* ‘his servant’

1 TOKEN = VERB

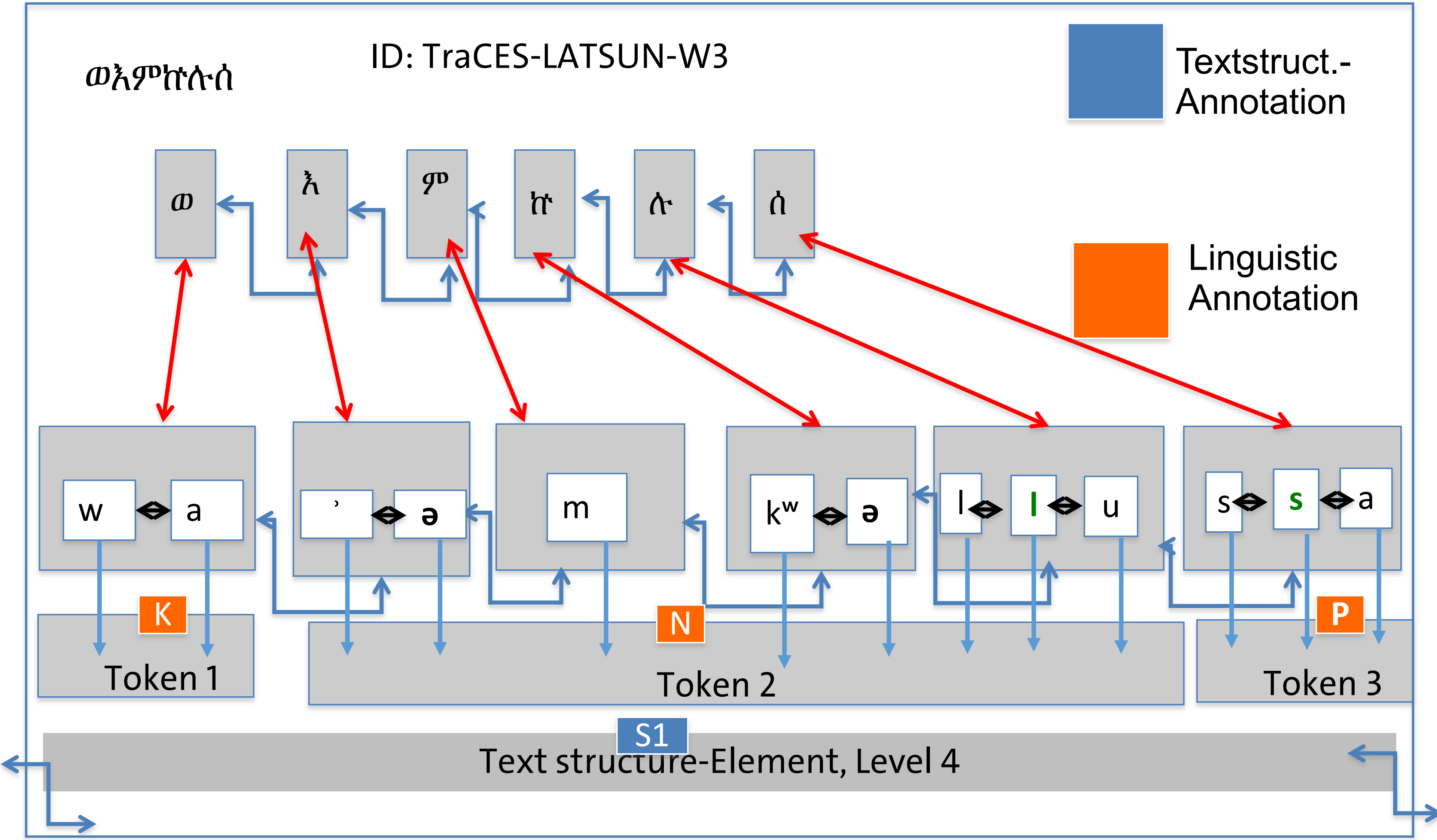
2 TOKENS = Ncom + PSuff

- Annotation MUST be done on Transcription
- Transliteration is a scientific process. For ML one needs first a large annotated corpus
- One need fine-grained morphological information in order to make the correct transliteration and tokenisation

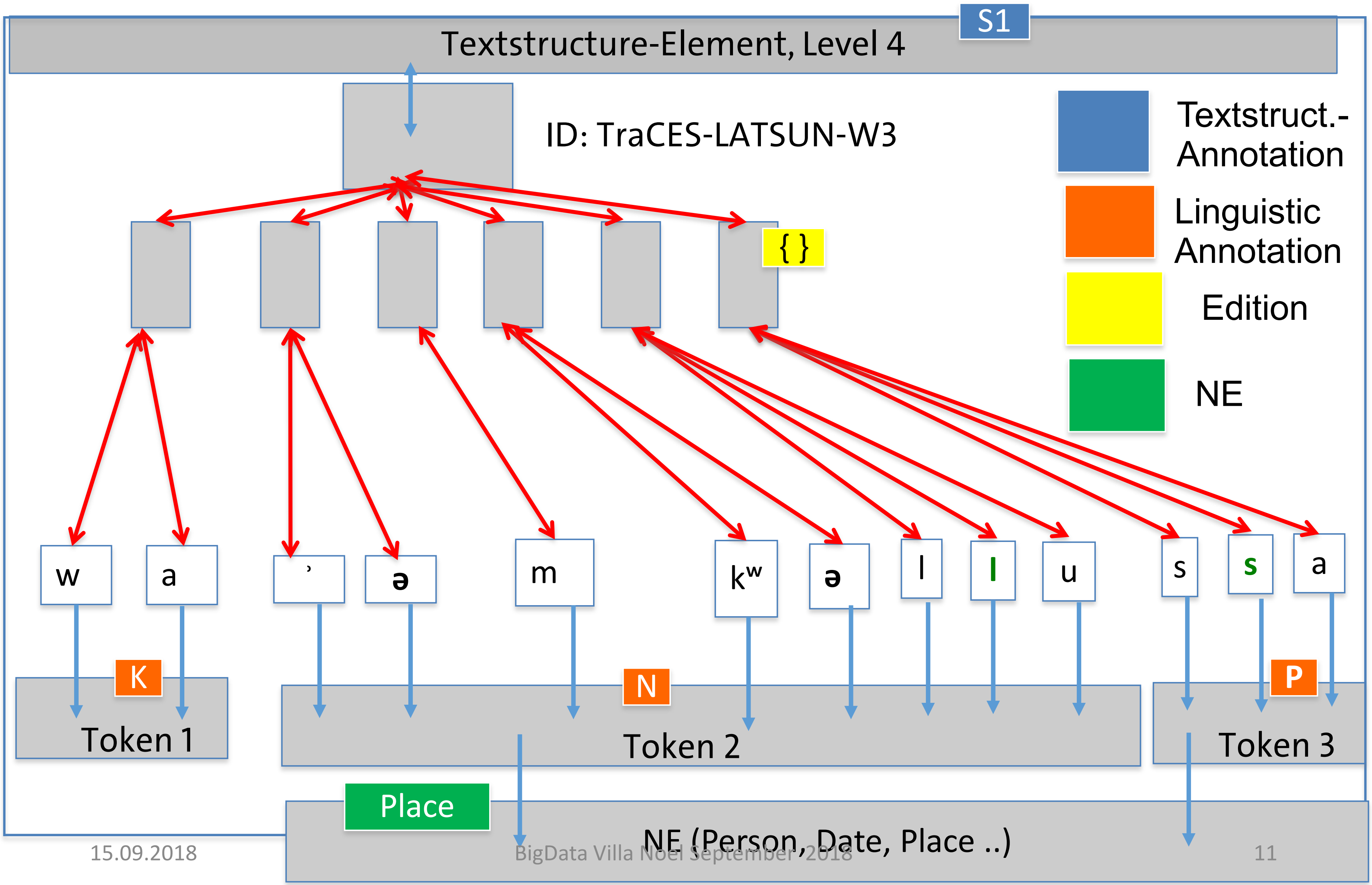
User requests and Challenges for the annotation

- Automatic transcription
- Synchronisation between original and transcription during the annotation
- Correction of the text during the annotation (while kkeping the annotation)
- Controlled automatic Annotation:
  - Tokenising
  - Change of the text
  - Linguistic Annotation
  - Sentence Annotation
  - NE-Annotation
- Possibility of very flexible text divisions (not necessary hierarchical)
- Multilevel annotation (flexible change of annotation level)
- Approx. 30 linguistic categories (PoS)(e.g. Number following 3 categories : Nature, Pattern and Syntax)
- User-friendly GUI
- Possibility of adapting the system for other scripts and transcriptions

TraCES-Annotation: Data-Model -1



- # TraCES-Annotation: Data-Model -2



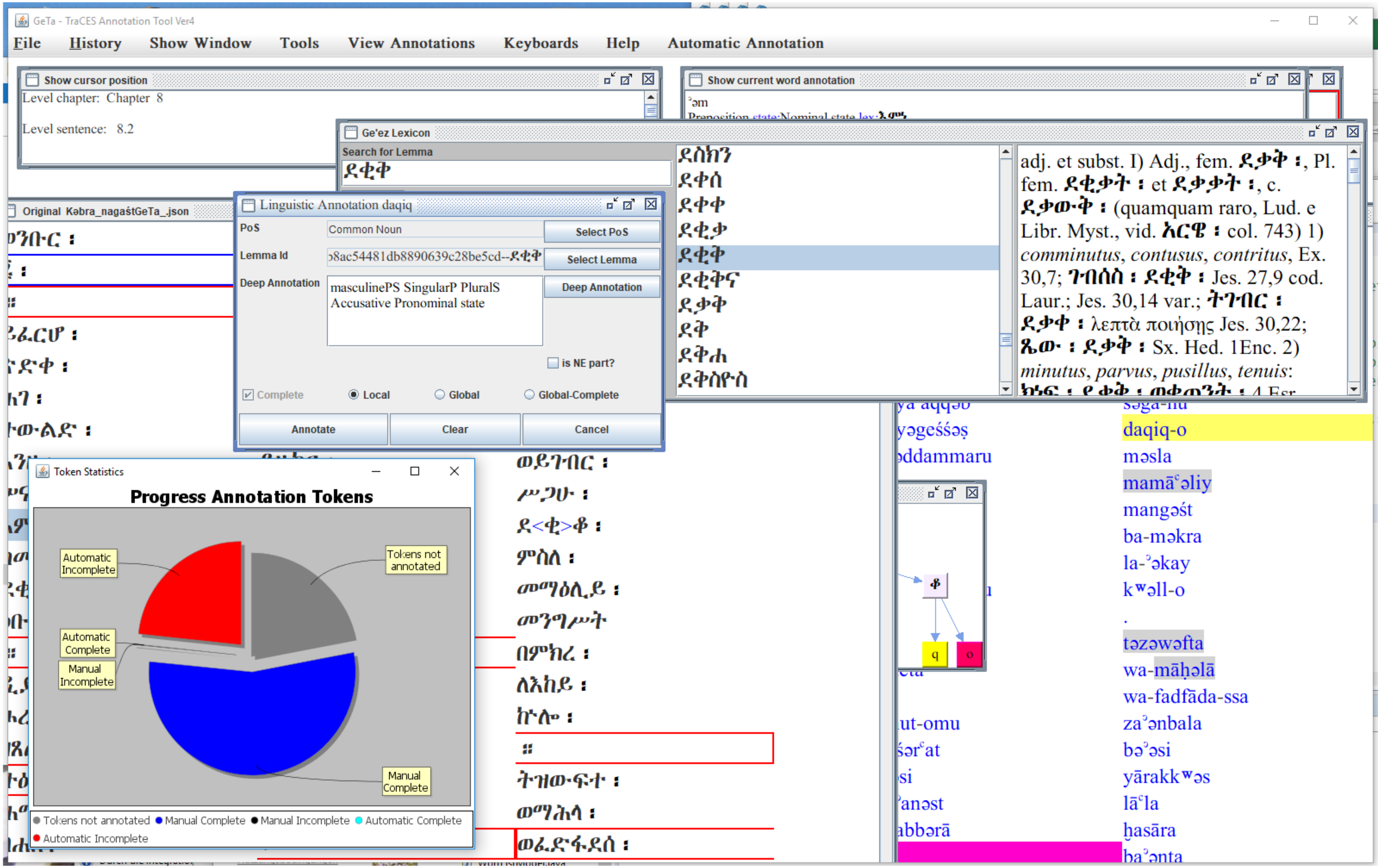
GeTa AnnotationTool

- **Features**
- Easy to use GUI
- Automatic initial transcription (vocalized or unvocalized)
- Synchronisation between original and transcription
- Controlled changes on text while annotating
- possible
- Controlled semi-automatic:
  - tokenization,
  - change of the transcripts text,
  - deep linguistic annotation + link to lexicon
  - Name Entity annotation linked with the authority DB.
- Automatic „sentence“ recognition
- Visualisation of data model
- Visualisation of annotation progress
- Can read additionally Classical Ethiopic inscriptions written with South Arabic script

Software development

- Client-Application
- Open source; Java
- Data-encoding:JSON

GeTa



Level chapter: Chapter 8

Level sentence: 8.2

ʾəm
Preposition state:Nominal state lex:እምነ

Search for Lemma: ደቂቅ

ደስክን
ደቀሰ
ደቀቀ
ደቂቃ
ደቂቅ
ደቂቅና
ደቃቅ
ደቅ
ደቅሐ
ደቅስዮስ

adj. et subst. I) Adj., fem. **ደቃቅ ፡**, Pl. fem. **ደቂቃት ፡** et **ደቃቃት ፡**, c. **ደቃውቅ ፡** (quamquam raro, Lud. e Libr. Myst., vid. **አርዌ ፡** col. 743) 1) *comminutus*, *contusus*, *contritus*, Ex. 30,7; **ገበሰስ ፡ ደቂቅ ፡** Jes. 27,9 cod. Laur.; Jes. 30,14 var.; **ትገብር ፡ ደቃቀ ፡** λεπτὰ ποιήσῃς Jes. 30,22; **ጼው ፡ ደቃቅ ፡** Sx. Hed. 1Enc. 2) *minutus*, *parvus*, *pusillus*, *tenuis*:

PoS: Common Noun

Lemma Id: ɔ8ac54481db8890639c28be5cd--ደቂቅ

Deep Annotation: masculinePS SingularP PluralS Accusative Pronominal state

☐ is NE part?

☑ Complete ● Local ○ Global ○ Global-Complete

Progress Annotation Tokens

● Tokens not annotated ● Manual Complete ● Manual Incomplete ● Automatic Complete ● Automatic Incomplete

ወይገብር ፡
ሥጋሁ ፡
ደ<ቂ>ቆ ፡
ምስለ ፡
መማዕሊይ ፡
መንግሥት
በምክረ ፡
ለእከዩ ፡
ኵሎ ፡
።
ትዝውፍት ፡
ወማሕላ ፡
ወፊድፋደሰ ፡

yəgeśśəṣ
ddammaru
ut-omu
śərʿat
si
ʾanəst
abbərā

daqiq-o
məsla
mamāʿəliy
mangəśt
ba-məkra
la-ʾəkay
kʷəll-o
.
təzəwəfta
wa-māḥəlā
wa-fadfāda-ssa
zaʾənbala
bəʾəsi
yārakkʷəs
lāʿla
ḫasāra
baʾənta

Generated Big Data

From one text File with 534 Kb Size

- 37764 „graphical units“=strings in classical Ethiopic text
- 56413 transliterated tokens
- 260433 annotated objects (single letters) + 37764 graphical units objects + 220215 ethiopic letters objects in the Data Structure file

7 Files for the annotation (3 annotation layers , 1 with the structure, 1 Metadata , 2 indexes)

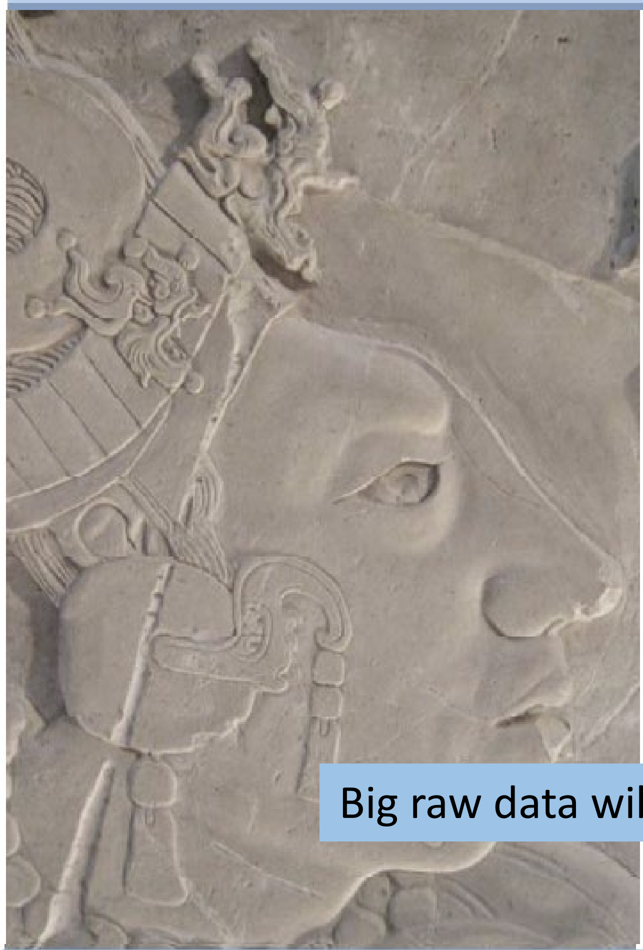
- 30,5 MB File containing the data structure
- 13,7 MB File containin´g linguistic annotations

Consequences

- Controlled automated annotation does not allow splitting the processed file
- Annotation tool must be able to handle this size of the data with implications in:
  - Reading
  - Searching
  - Global annotation
  - Global edit operations (delete, replace, modify transliteration)

Textdatenbank und Wörterbuch des Klassischen Maya

AKADEMIE WISSENSCHAFTEN UND KÜNSTE NORDRHEIN-WESTFALEN



Prof. Dr. Nikolai Grube[1]

Dr. Christian Prager[1]

Dr. Sven Gronemeyer[1,2]

Elisabeth Wagner[1]

Katja Diederichs[1]

Franziska Diehr[3]

Maximilian Brodhun[3]

[1] Rheinische Friedrich-Wilhelms-Universität, Bonn

[2] La Trobe University, Melbourne

[3] Niedersächsische Staats- und Universitätsbibliothek, Göttingen

Big raw data will never be an issue BUT :

UNION DER DEUTSCHEN AKADEMIEN DER WISSENSCHAFTEN

UNIVERSITÄT BONN

Abteilung für Altamerikanistik

Universitäts- und Landesbibliothek Bonn

SUB Niedersächsische Staats- und Universitätsbibliothek Göttingen

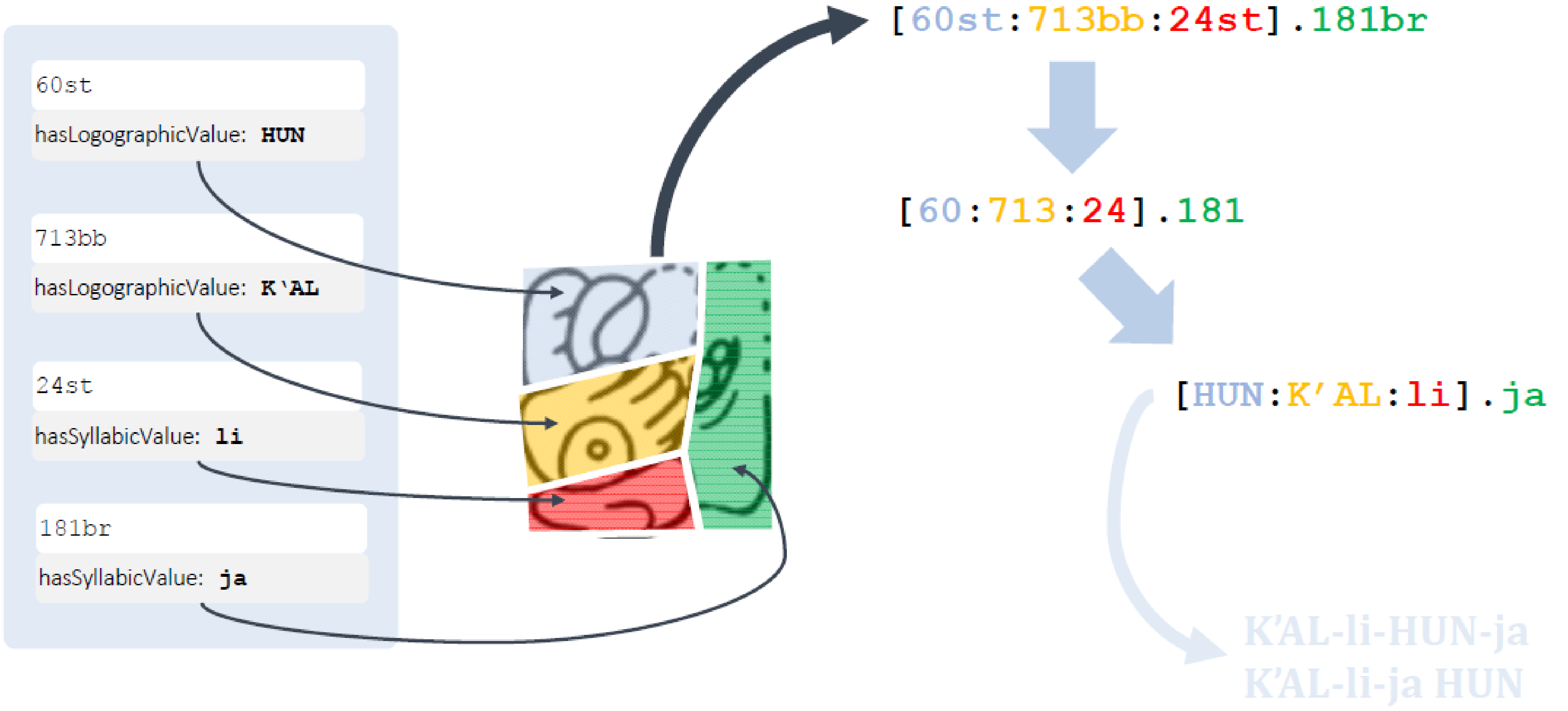
DARIAH-DE

TextGrid Virtuelle Forschungsumgebung für die Geisteswissenschaften

From Image to Text

Sign Catalogue → Text-Image-Link-Editor (Documentation Layer) → Text Markup (Text Layer) → Text Annotation (Analysis Layer)

Bsp.: PAL: PMI1, A3



60st
hasLogographicValue: **HUN**

713bb
hasLogographicValue: **K'AL**

24st
hasSyllabicValue: **li**

181br
hasSyllabicValue: **ja**

`[60st:713bb:24st].181br`

`[60:713:24].181`

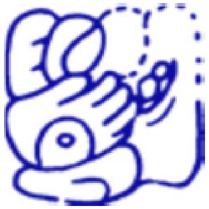
`[HUN:K'AL:li].ja`

K'AL-li-HUN-ja
K'AL-li-ja HUN

©Prager, Grönemayer et. al.

9 -Annotations Layers

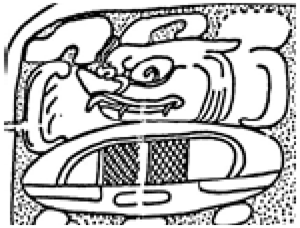
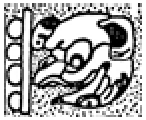
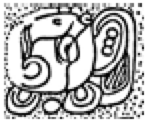
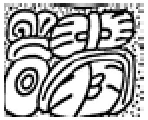
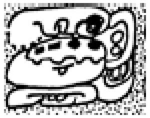
Text Markup (Semiautomatic) → **Text Annotation (Manual)** → **Dictionary**



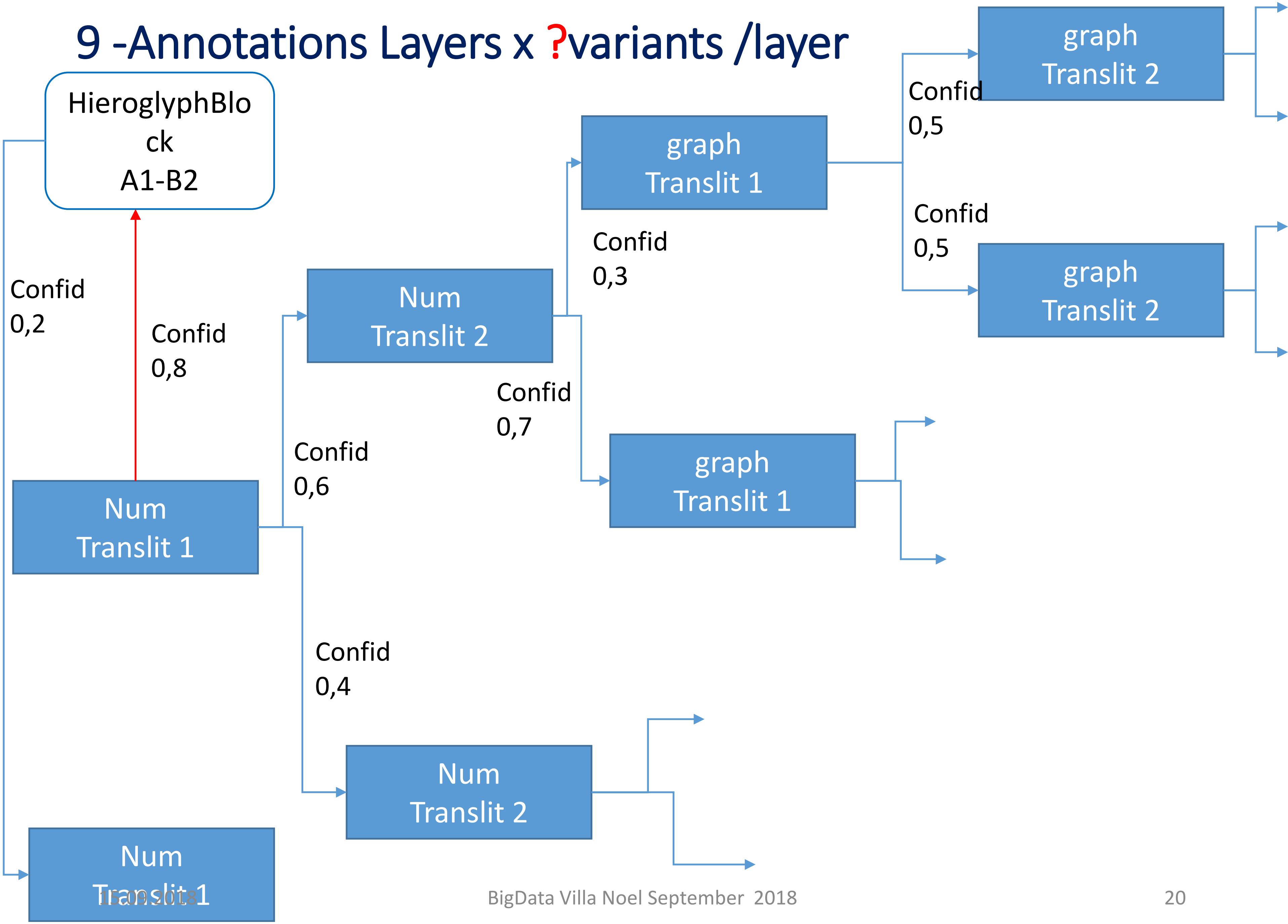
Alphanumeric Transliteration
[60st:713bb:24st].181br

Numeric Transliteration
[60:713:24].181

Graphematic Transliteration	**[HUN:K’AL:li].ja**
Graphemic Transliteration	**K’AL-li-HUN-ja**
Phonemic Transliteration	**k’al^{li}=hu’un=ja**
Morphological Transcription	*k’al-Ø+hu’un-[a]j-Ø*
Morphophonemic Transcription	*k’al-Ø+hu’un-aj-Ø*
Morphosyntactic Glossing	V.TR:hold-NMLZ+N:paper-INCH-3s.ABS
Final Transcription	*k’alhu’unaj*
Translation	he was presented the crown

Textbeispiel						**Anmerkungen**
						Zeichenkatalog abfragen und muss Relation Graph → Zeichen heraussuchen
Alternative 2	124:[28°738]:548:74	–	–	–	[5:201:23].683	
Alternative 3	124:[28°738]:548:74	–	–	–		
Analyse-ebene	**graphemische Transliteration 1**					semi-automatische Einsetzung der Lautwerte entsprechend der Werte im Zeichenkatalog, Auswahl entsprechend der hinterlegten Zeichenfunktionen, sofern größer 1
Alternative 1	**tzi:ka°XOK:HAB:ma**	**9.PIK**	**SIH:ya.ja**	**SAK.ma:su**	**5:TZ'AM:na.ja**	- mit Information zur Sicherheit (Konfidenz-Angabe)
Alternative 2	**tzi:ka°XOK:HAB:ma**	-	-	-	-	
Alternative 3	**tzi:ka°XOK:HAB:ma**	-	-	-	-	
Analyse-ebene	**graphemische Transliteration 2**					Traditionelle Transliteration in der Fachwissenschaft und korrekte Lesereihenfolge.
Alternative 1	**tzi-ka XOK HAB-ma**	**9 PIK**	**SIH-ya-ja**	**SAK ma-su**	**5-TZ'AM-na-ja**	Transliterationen werden in Fettschrift angegeben. (Diese Analyseebene ist relevant für Visual Library. Export in METS/MODS ist nicht nötig. Muss per OAI PMH ausgeliefert werden können.)
Alternative 2	**TZIK XOK HAB**	–	–	–	**5-201-na-ja**	TZIK ist eine angenommene Lesung, weil wir davon ausgehen, dass 124 und 28 ein komplexes

9 -Annotations Layers x ?variants /layer



HieroglyphBlock A1-B2

Confid 0,2

Confid 0,8

Num Translit 1

Confid 0,6

Num Translit 2

Confid 0,3

graph Translit 1

Confid 0,5

graph Translit 2

Confid 0,5

graph Translit 2

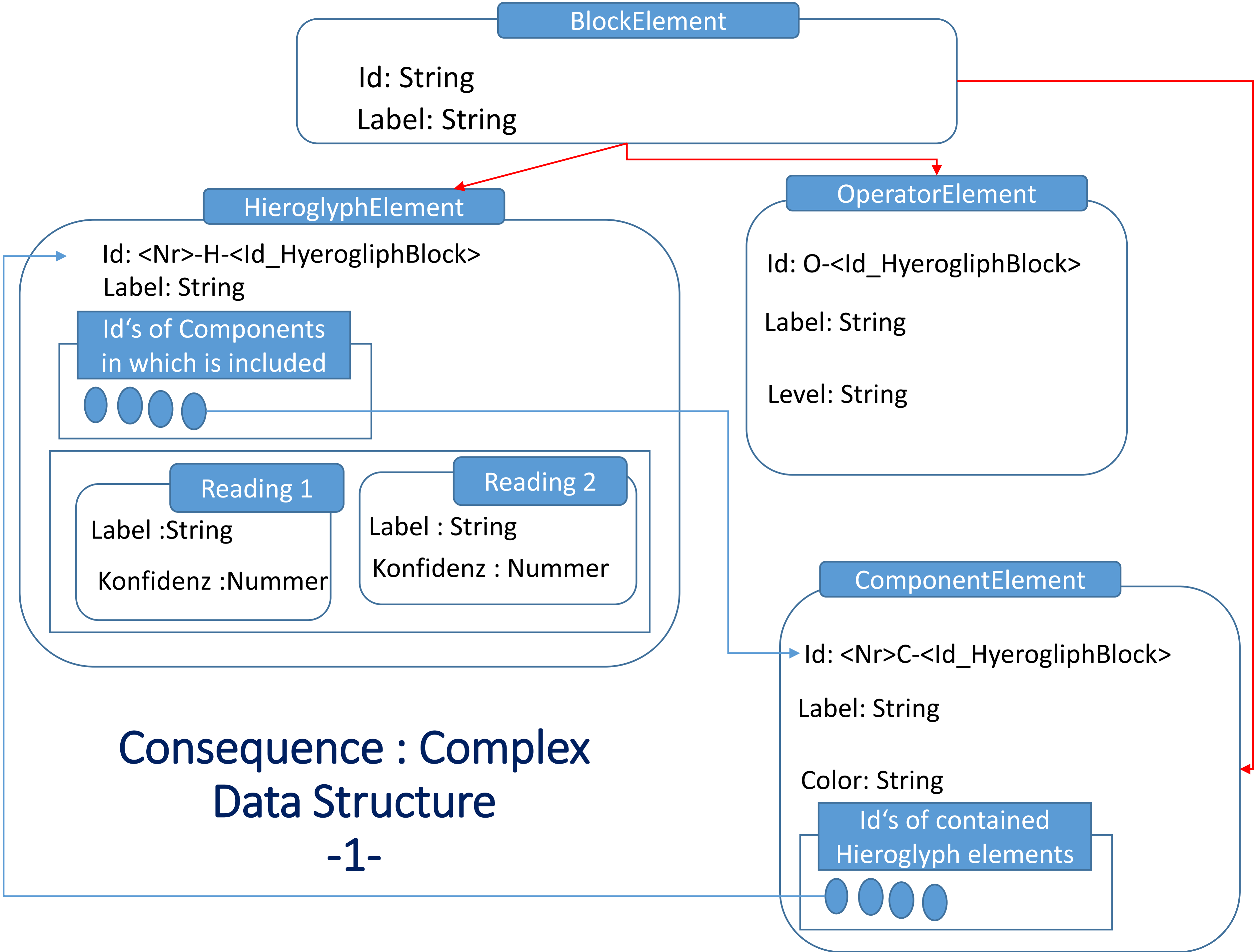
Confid 0,7

graph Translit 1

Confid 0,4

Num Translit 2

Num Translit 1



BlockElement

Id: String

Label: String

HieroglyphElement

Id: <Nr>-H-<Id_HyerogliphBlock>

Label: String

Id‘s of Components in which is included

Reading 1

Label :String

Konfidenz :Nummer

Reading 2

Label : String

Konfidenz : Nummer

OperatorElement

Id: O-<Id_HyerogliphBlock>

Label: String

Level: String

ComponentElement

Id: <Nr>C-<Id_HyerogliphBlock>

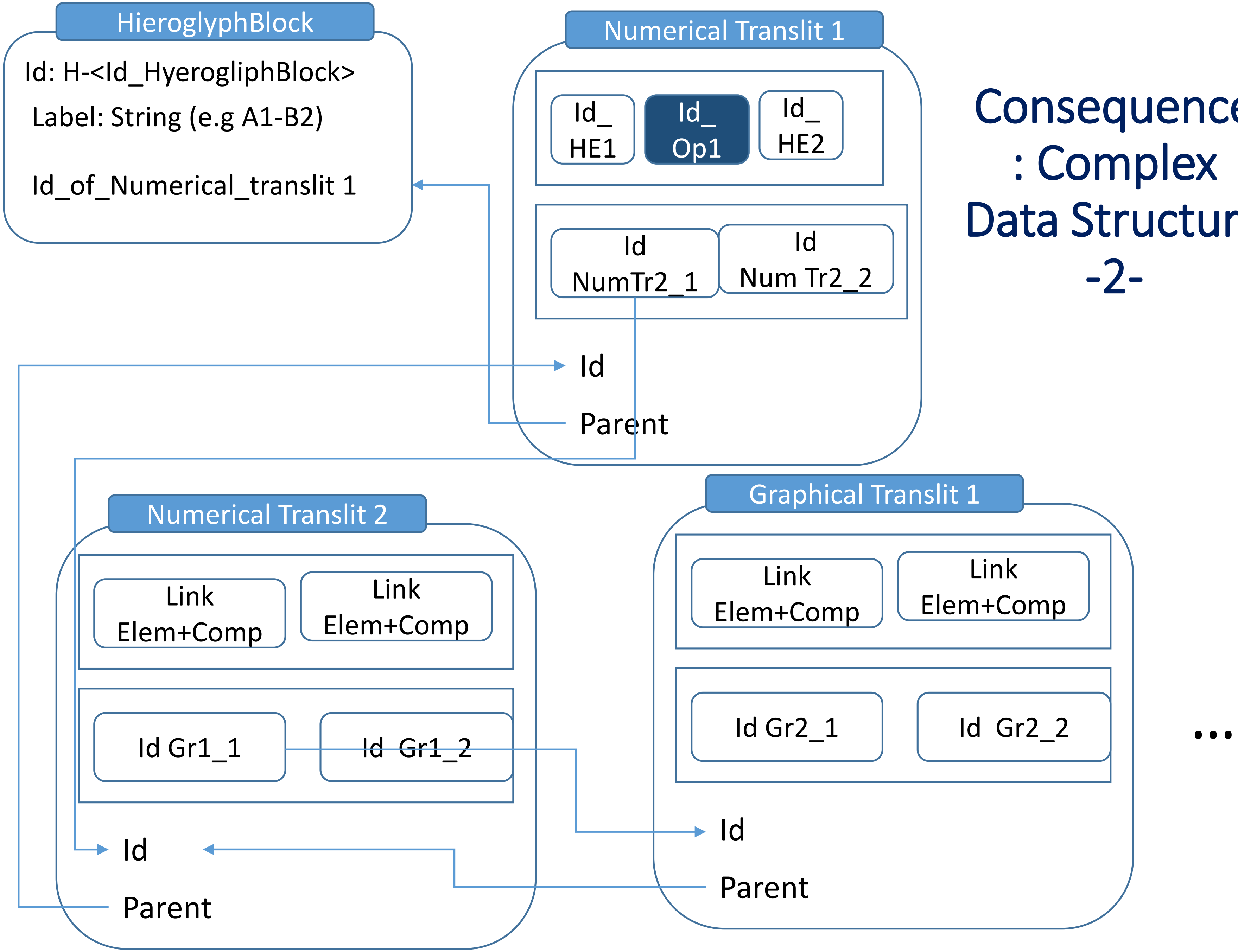
Label: String

Color: String

Id‘s of contained Hieroglyph elements

Consequence : Complex Data Structure -1-

Consequence : Complex Data Structure -2-



HieroglyphBlock

Id: H-<Id_HyerogliphBlock>

Label: String (e.g A1-B2)

Id_of_Numerical_translit 1

Numerical Translit 1

Id_ HE1

Id_ Op1

Id_ HE2

Id NumTr2_1

Id Num Tr2_2

Id

Parent

Numerical Translit 2

Link Elem+Comp

Link Elem+Comp

Id Gr1_1

Id Gr1_2

Id

Parent

Graphical Translit 1

Link Elem+Comp

Link Elem+Comp

Id Gr2_1

Id Gr2_2

Id

Parent

…

Link numerical Translit 2

Id_Element $ LabelElement & Id_Component * Color_Of_Component

1H-HB20 $ 123 & 1C-HB20 * Black
O-HB20 $: *

Link numerical Graphical Translit 1

Id_Element $ LabelElement @ Confidence & Id_Component * Color_Of_Component

1H-HB20 $ tzi @ 080 & 1C-HB20 * Black

O-HB20 $: *

Consequence : need for strategies for information compression

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

- Illustrated through two main works of Dimitrie Cantemir-

UHH 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

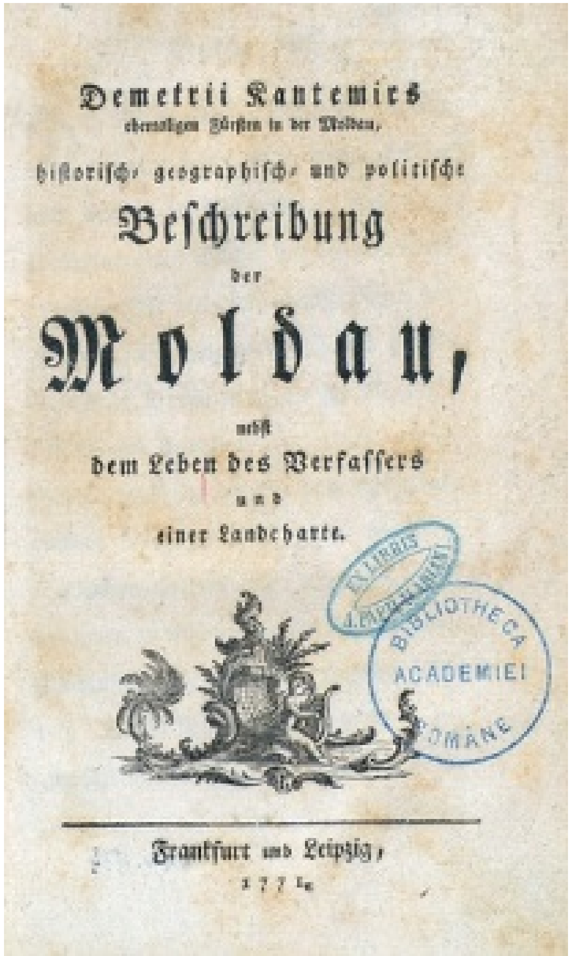
Progress work in automatic recognition of vague expressions

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.



Demetrii Kantemirs
ehemaligen Fürſten in der Moldau,
hiſtoriſch- geographiſch- und politiſche
Beſchreibung
der
Moldau,
nebſt
dem Leben des Verfaſſers
und
einer Landcharte.

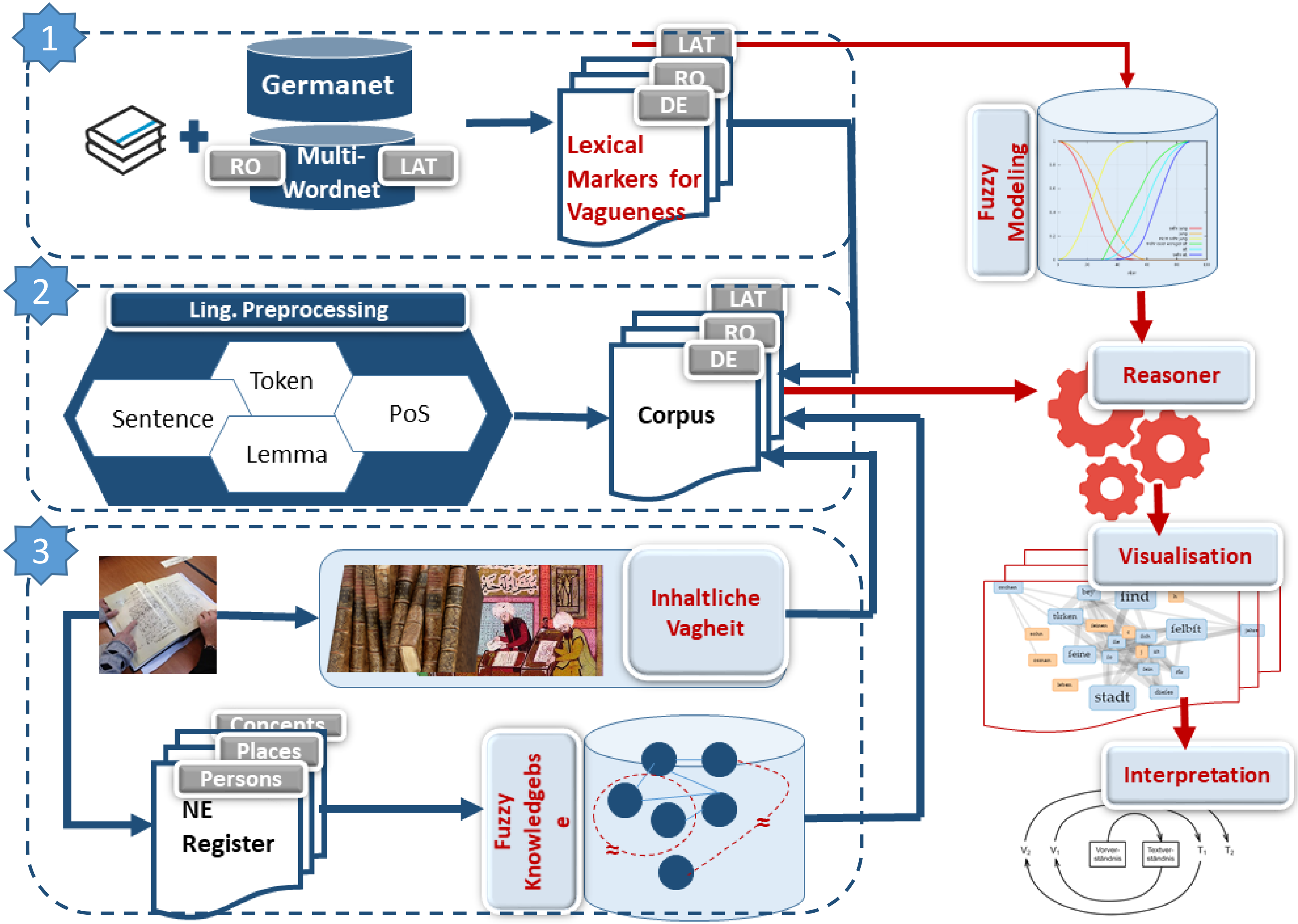
Frankfurt und Leipzig,
1771.

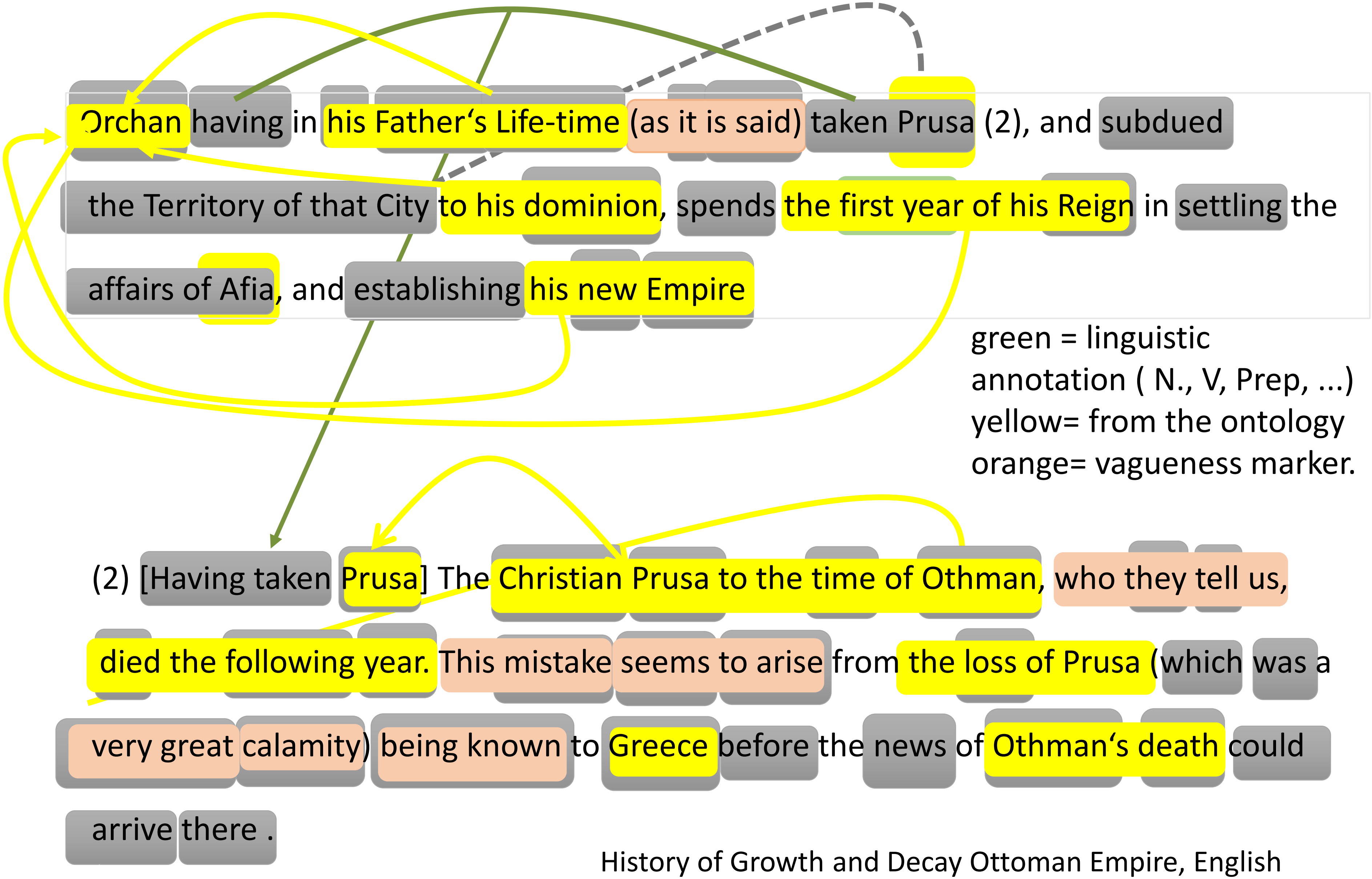
„Nu îndrăznim să spunem ce e adevărat şi ce e fals într-o asemenea întunecime a istoriei. “

(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“

System Architecture



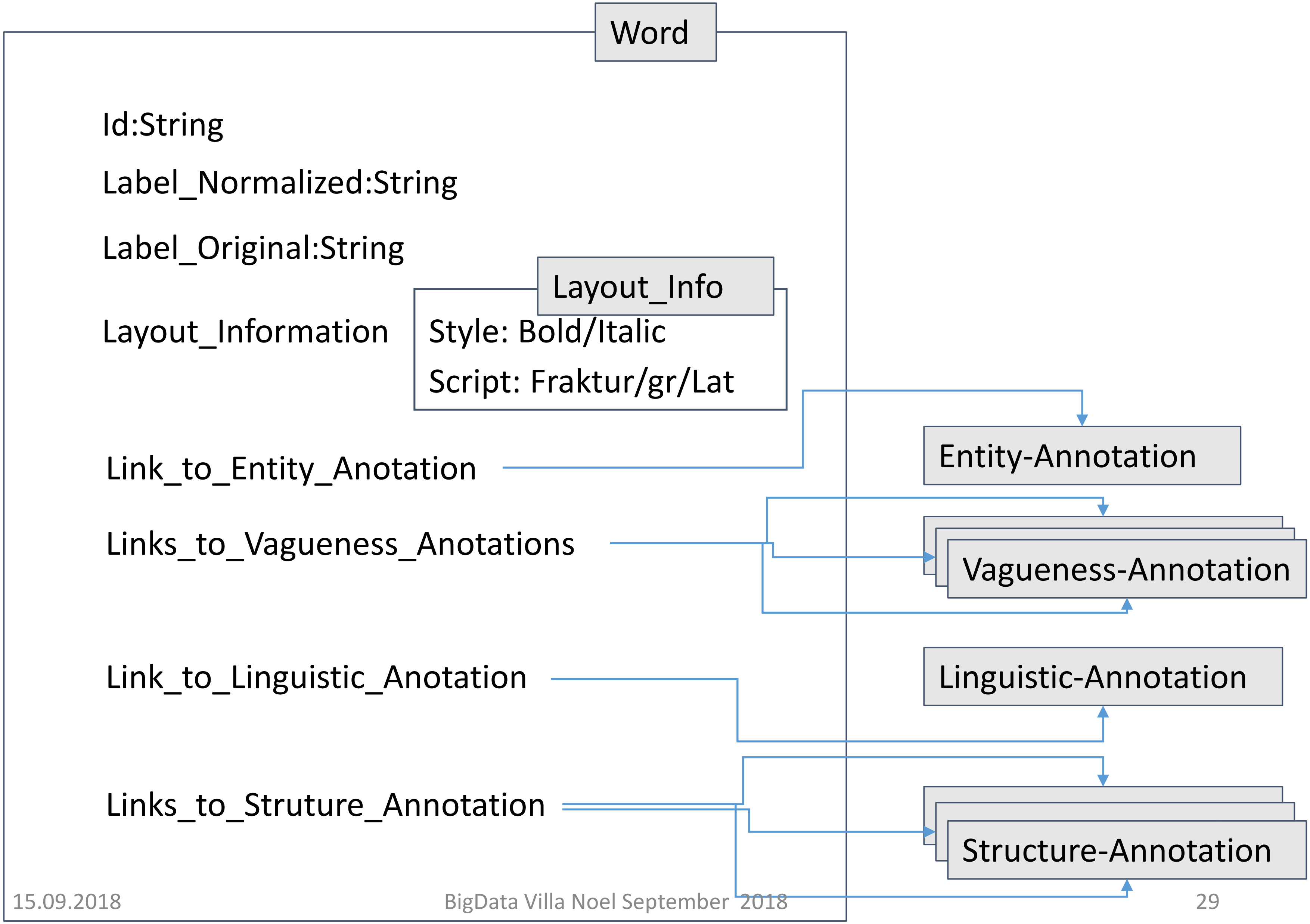


Orchan having in his Father‘s Life-time (as it is said) taken Prusa (2), and subdued the Territory of that City to his dominion, spends the first year of his Reign in settling the affairs of Afia, and establishing his new Empire

green = linguistic annotation (N., V, Prep, ...)
yellow= from the ontology
orange= vagueness marker.

(2) [Having taken Prusa] The Christian Prusa to the time of Othman, who they tell us, died the following year. This mistake seems to arise from the loss of Prusa (which was a very great calamity) being known to Greece before the news of Othman‘s death could arrive there .

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



Word

Id:String

Label_Normalized:String

Label_Original:String

Layout_Information

Layout_Info

Style: Bold/Italic

Script: Fraktur/gr/Lat

Link_to_Entity_Anotation

Links_to_Vagueness_Anotations

Link_to_Linguistic_Anotation

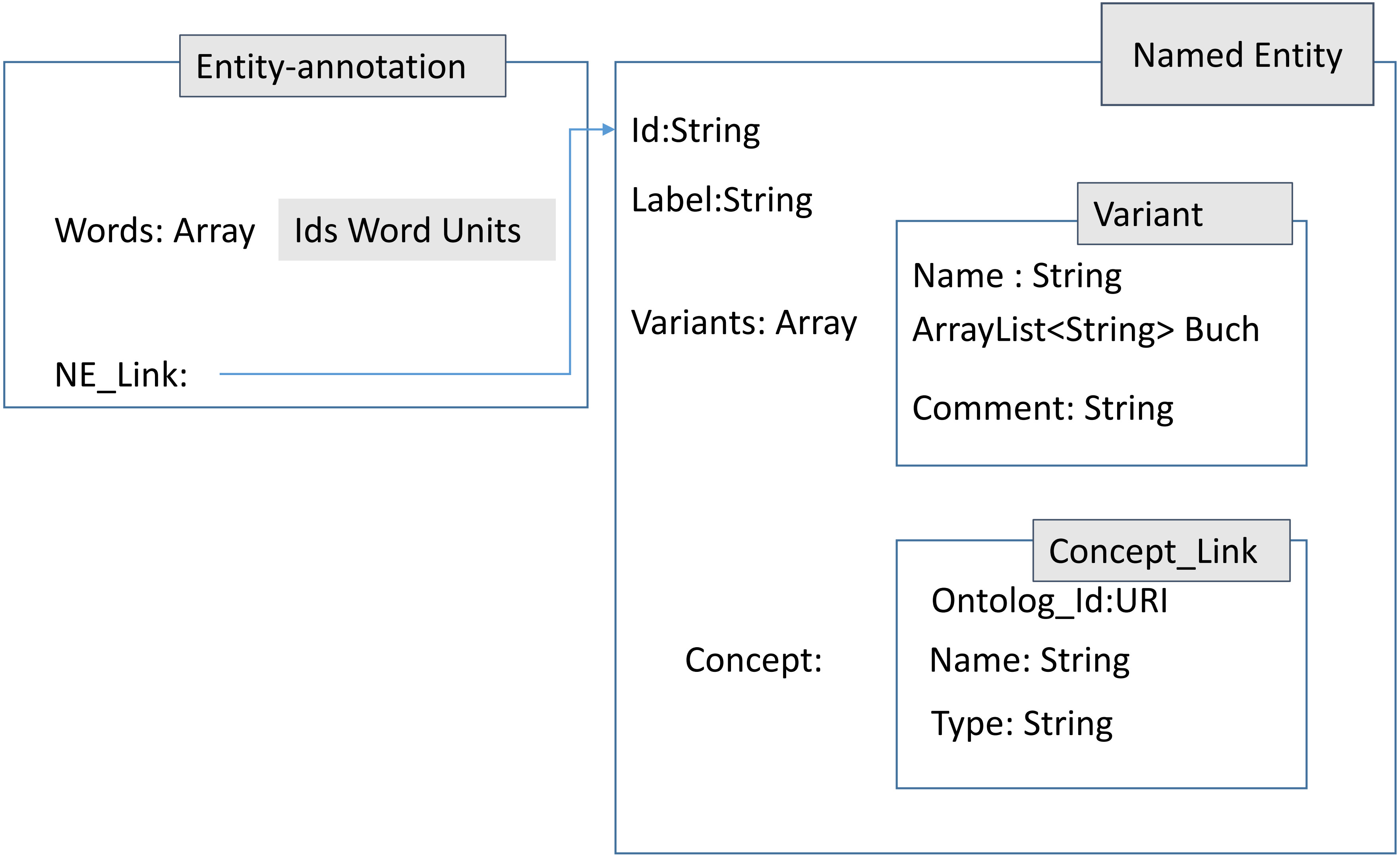
Links_to_Struture_Annotation

Entity-Annotation

Vagueness-Annotation

Linguistic-Annotation

Structure-Annotation



Entity-annotation

Words: Array

Ids Word Units

NE_Link:

Named Entity

Id:String

Label:String

Variants: Array

Variant

Name : String

ArrayList<String> Buch

Comment: String

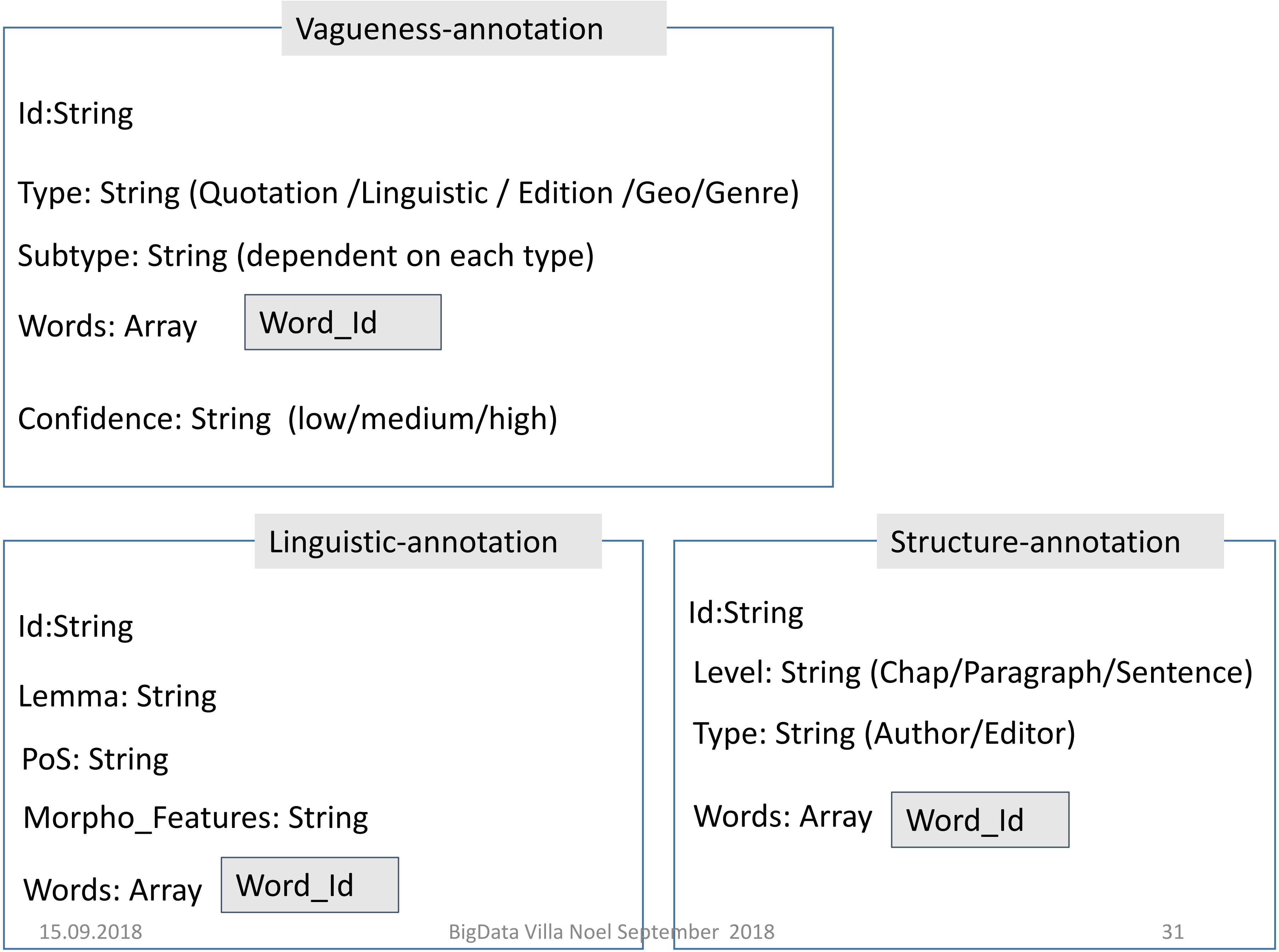
Concept:

Concept_Link

Ontolog_Id:URI

Name: String

Type: String



Vagueness-annotation

Id:String

Type: String (Quotation /Linguistic / Edition /Geo/Genre)

Subtype: String (dependent on each type)

Words: Array Word_Id

Confidence: String (low/medium/high)

Linguistic-annotation

Id:String

Lemma: String

PoS: String

Morpho_Features: String

Words: Array Word_Id

Structure-annotation

Id:String

Level: String (Chap/Paragraph/Sentence)

Type: String (Author/Editor)

Words: Array Word_Id

Big Data ?

- Initially:
  - Approx. 1000 pages / volume x 3 languages
- Annotation will be done mostly at word level BUT

- Each “Word-Object” has a very complex structure AND
- A proper annotation must have in background a Knowledge Base containing only as individuals:
  - Over 300 Persons
  - Approx. 500 geographical names
  - Over 300 domain specific concepts
- Approx. 200 vagueness indicators /language will be annotates

Conclusion

- Raw “small” data may lead to “big” annotated data.
- Raw “small” data need (manual) annotation as no statistical algorithm may work -> user has control on the knowledge fed into the computer
- Big raw data cannot afford manual annotation
- Automatic Annotation introduce a degree of errors.
- Is it a trade-off between using no additional information (raw data) and possibly annotated data with some errors.
- How can automatic annotations on big data being improved (manual annotations -> evaluation test set)

Merci pour votre attention!