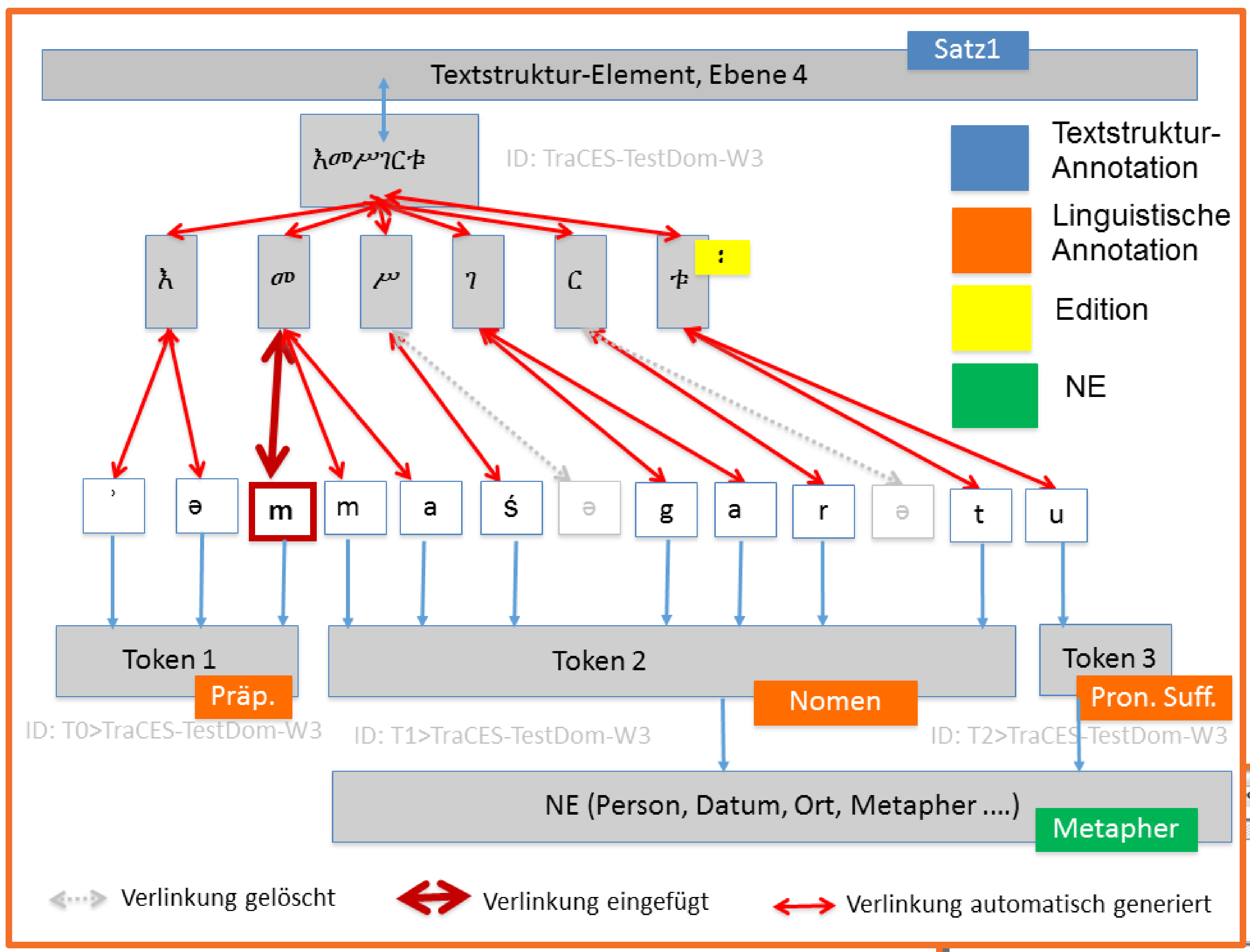
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

IDEAS
European Research Council
TraCES
VolkswagenStiftung
HerCoRe
Textdatenbank und Wörterbuch des Klassischen Maya
Akademie der Wissenschaften und der Künste Nordrhein-Westfalen

Data Modelling for Historical Corpus Annotation

Cristina Vertan, University of Hamburg



Gəʿəz (a south semitic language) is mentioned for the first time as lnguage of the Aksum Kingdom (approx. 1st-7th century. B.C.). The Aksum Kingdom covered the north alpine part of today Ethiopia (Təgray) as well as Eritra until the Read Sea.the following centuries, until nowadays Gəʿəz was preserved as main language of the Christian Ethiopic Church.

The very rich early Christian literature contains a large number of translations from ancient Geek and Arabic (after the 13th century). **Gramatical interference phenomena** show these influences

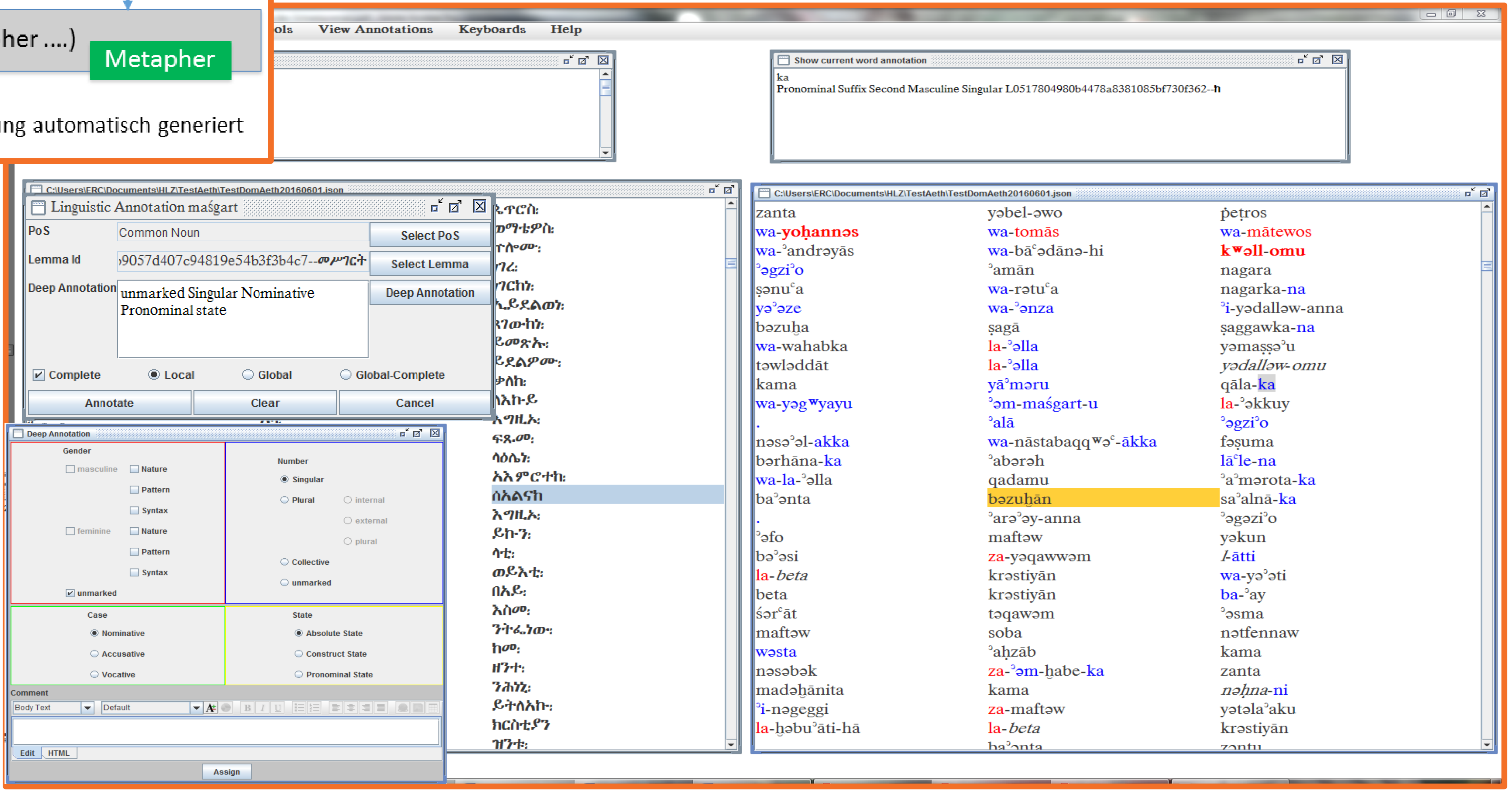
Own Alphabet with syllabic rules

- Written from left to right
- Complete representation of the vowels (in contrast with Arabic and Hebrew)
- **Laryngals and Sibilanta are phonemich and graphemic beliebig interchangeable.**

Noconcatenative Morphologi: a lexeme is described as a combination of two Elements: the Root and the Pattern (Scheme) .

√**qds** × ʾasta1a22a3a → **ʾastaqaddasa**

Classical Ethiopic



Show current word annotation

ka
Pronominal Suffix Second Masculine Singular L0517804980b4478a8381085bf730f362--ከ

C:\Users\ERC\Documents\HLZ\TestAeth\TestDomAeth20160601.json

Linguistic Annotation maśgart

PoS: Common Noun — Select PoS

Lemma Id: ›9057d407c94819e54b3f3b4c7--መሥገርት — Select Lemma

Deep Annotation: unmarked Singular Nominative Pronominal state — Deep Annotation

☑ Complete ● Local ○ Global ○ Global-Complete

Annotate Clear Cancel

Deep Annotation

Gender: □ masculine □ Nature □ Pattern □ Syntax; □ feminine □ Nature □ Pattern □ Syntax; ☑ unmarked

Number: ● Singular; ○ Plural ○ internal ○ external ○ plural; ○ Collective; ○ unmarked

Case: ● Nominative ○ Accusative ○ Vocative

State: ● Absolute State ○ Construct State ○ Pronominal State

Comment

Body Text — Default

Edit HTML

Assign

ጲጥሮስ:
ማቴዎስ:
ቶሎሙ:
ነገረ:
ገርክኒ:
ኢይደልወኒ:
እገሎኪኒ:
ይመጽእ:
ይደልዎሙ:
ቃለከ:
ለእኩይ:
እግዚኡ:
ፍጺመ:
ላዕሌነ:
አእምሮተከ:
ሰአልናከ
እግዚኡ:
ይኩን:
ላቲ:
ወይእቲ:
በአይ:
እስመ:
ንትፌነው:
ከመ:
ዘንተ:
ንሕነኒ:
ይትለአኩ:
ክርስቲያን
ዝንቱ:

C:\Users\ERC\Documents\HLZ\TestAeth\TestDomAeth20160601.json

zanta
wa-yoḥannəs
wa-ʾandrəyās
ʾəgziʾo
ṣənuʿa
yəʾəze
bəzuḫa
wa-wahabka
təwlədḍāt
kama
wa-yəgʷyayu
.
nəsəʾəl-akka
bərhāna-ka
wa-la-ʾəlla
baʾənta
.
ʾəfo
bəʾəsi
la-beta
beta
śərʿāt
maftəw
wəsta
nəsəbək
madəḫānita
ʾi-nəgeggi
la-ḫəbuʾāti-hā

yəbel-əwo
wa-tomās
wa-bāʿədānə-hi
ʾamān
wa-rətuʿa
wa-ʾənza
śagā
la-ʾəlla
la-ʾəlla
yāʾməru
ʾəm-maśgart-u
ʾalā
wa-nāstabaqqʷəʿ-ākka
ʾabərəh
qadamu
bəzuḫān
ʾarəʾəy-anna
maftəw
za-yəqawwəm
krəstiyān
krəstiyān
təqawəm
soba
ʾaḥzāb
za-ʾəm-ḫabe-ka
kama
za-maftəw
la-beta
baʾənta

ṗetros
wa-mātewos
kʷəll-omu
nagara
nagarka-na
ʾi-yədalləw-anna
ṣaggawka-na
yəmaṣṣəʾu
yədalləw-omu
qāla-ka
la-ʾəkkuy
ʾəgziʾo
fəṣuma
lāʿle-na
ʾaʾmərota-ka
saʾalnā-ka
ʾəgəziʾo
yəkun
lātti
wa-yəʾəti
ba-ʾay
ʾəsma
nətfennaw
kama
zanta
nəḥna-ni
yətəlaʾaku
krəstiyān
zəntu

Features

- Semi-automatic tokenisation as well as annotation of linguistic features and Nes
- Automatic Sentence recognition
- Visualisation of results of automatic operations
- Post-Annotation correction of the transliterated base Text without lost of annotation information
- Multi-levelAnnotation
- Synchronisation between original and transcribed/transliterated Text

Implementation

- Java -based client application
- Output-Format JSON
- Export TEI, ANNIS, PAULA-XML
- Big-data management

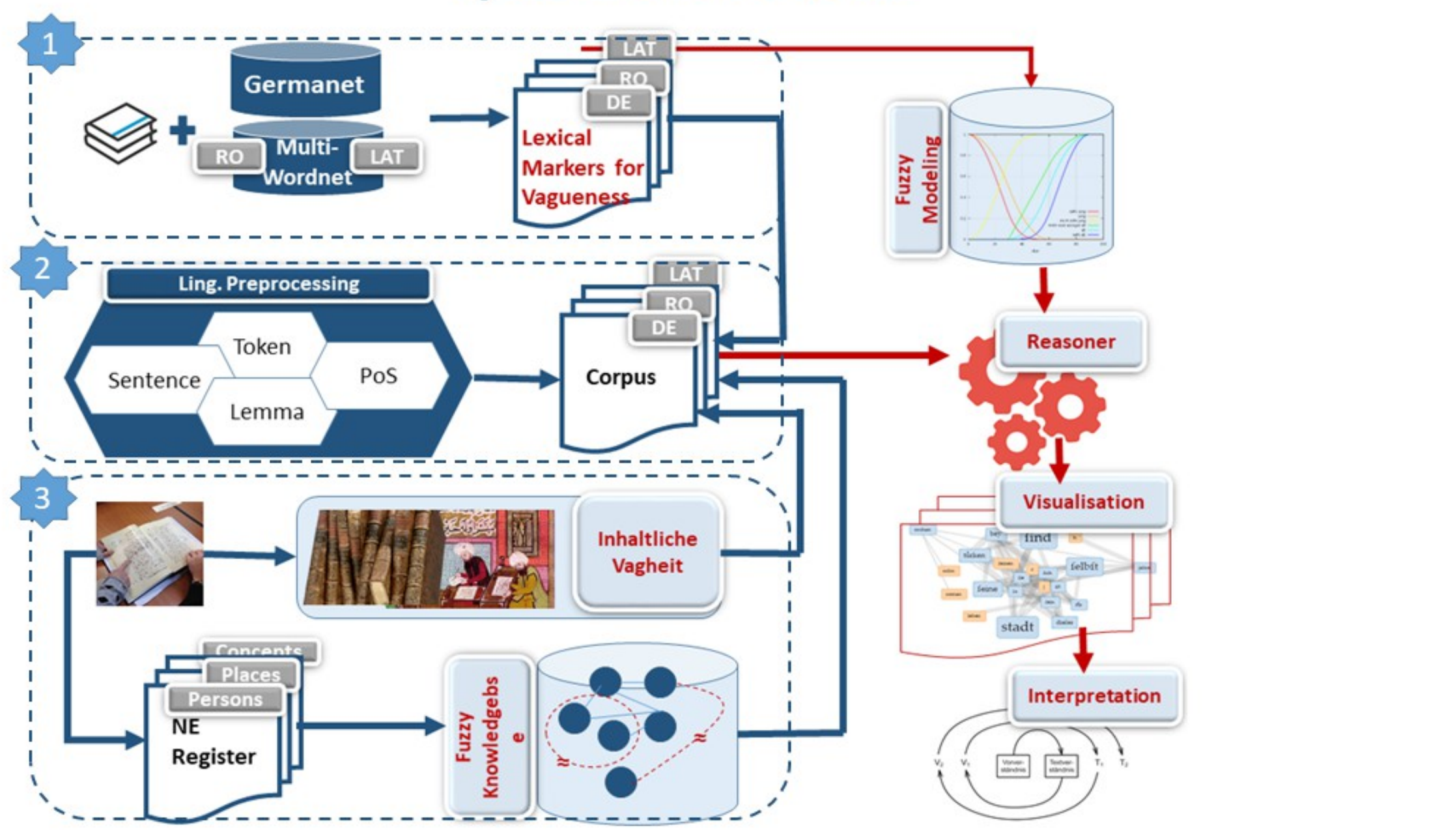
HerCoRe

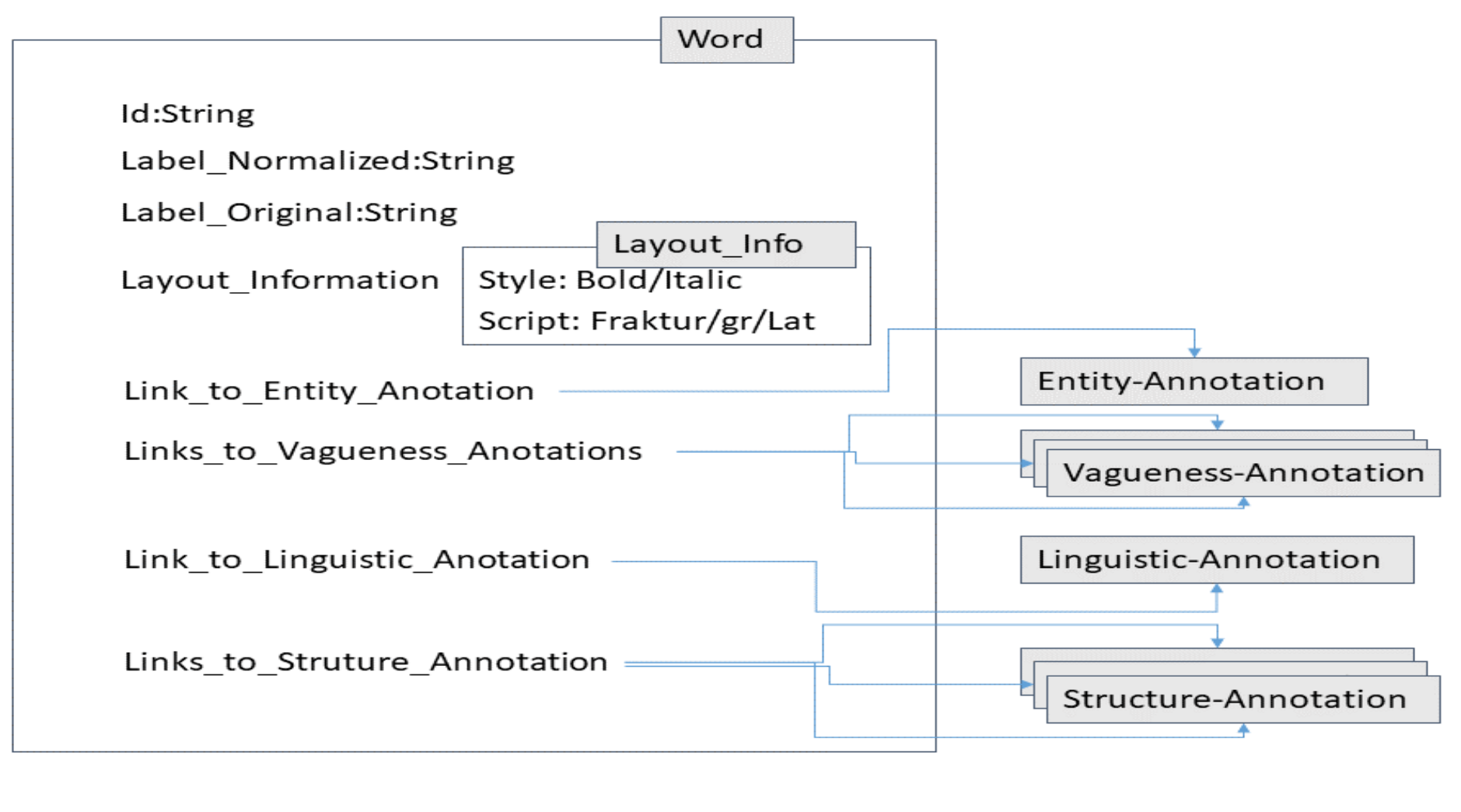
Dimitrie Cantemir (1673 -1723)

- 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.

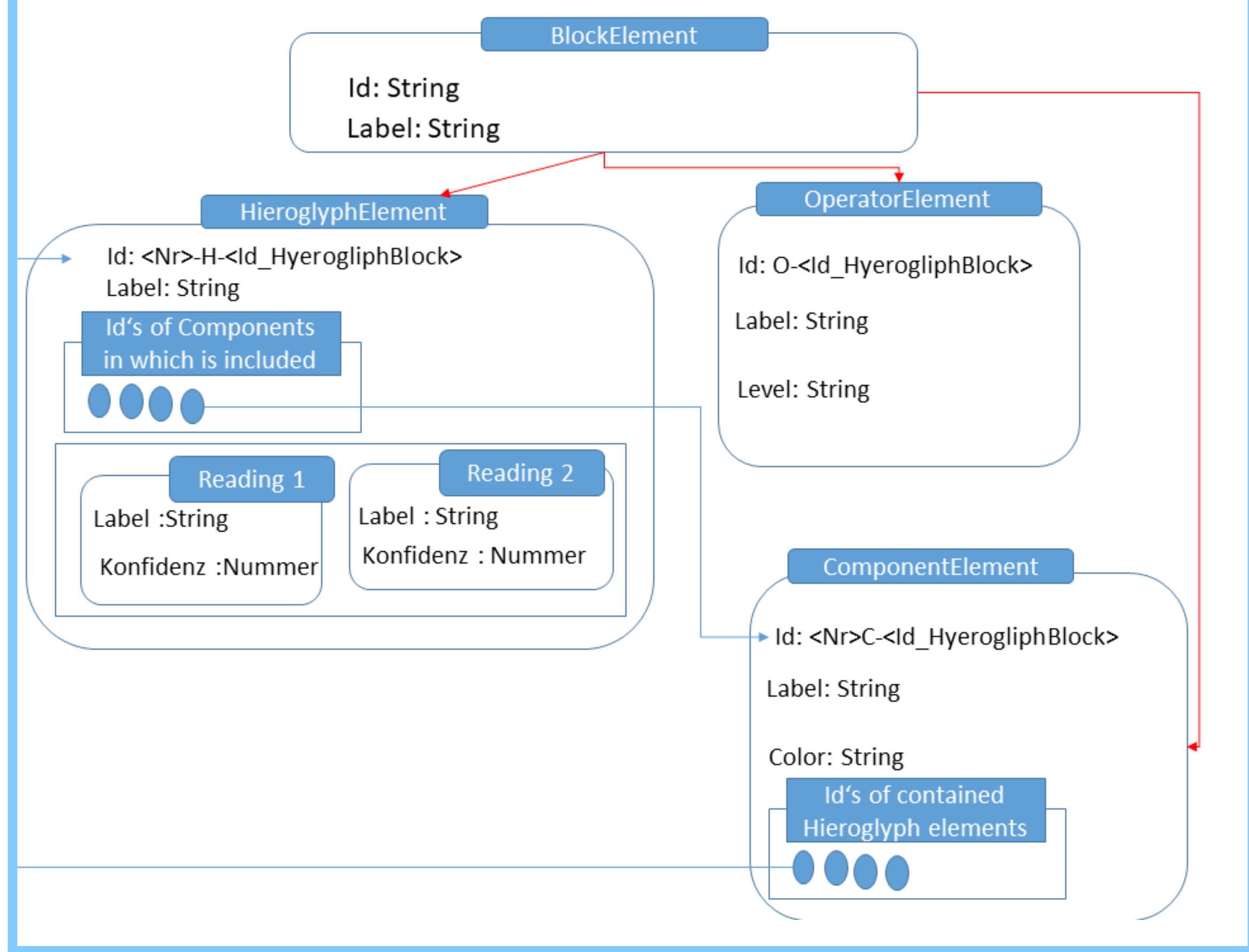
System Architecture



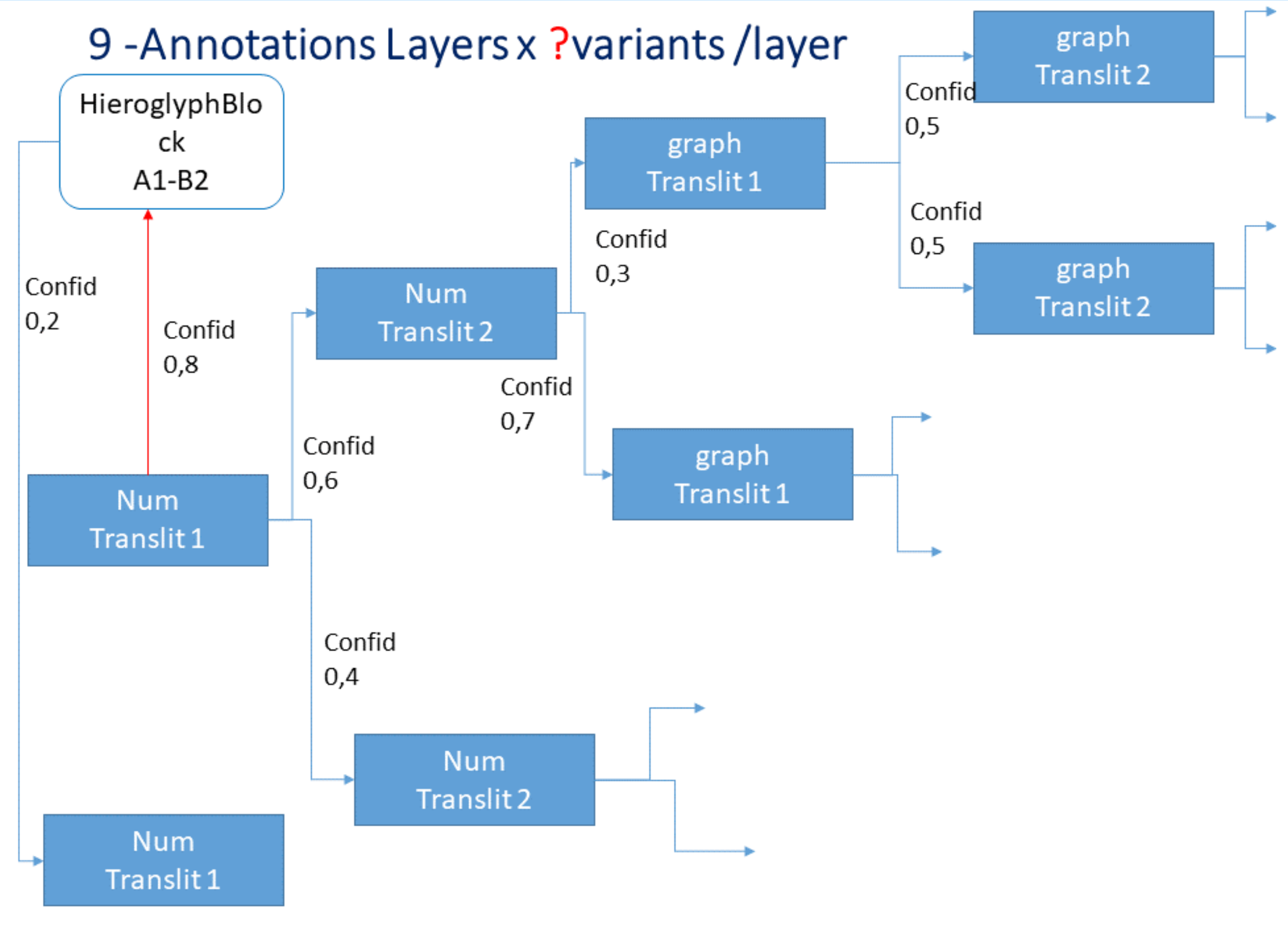


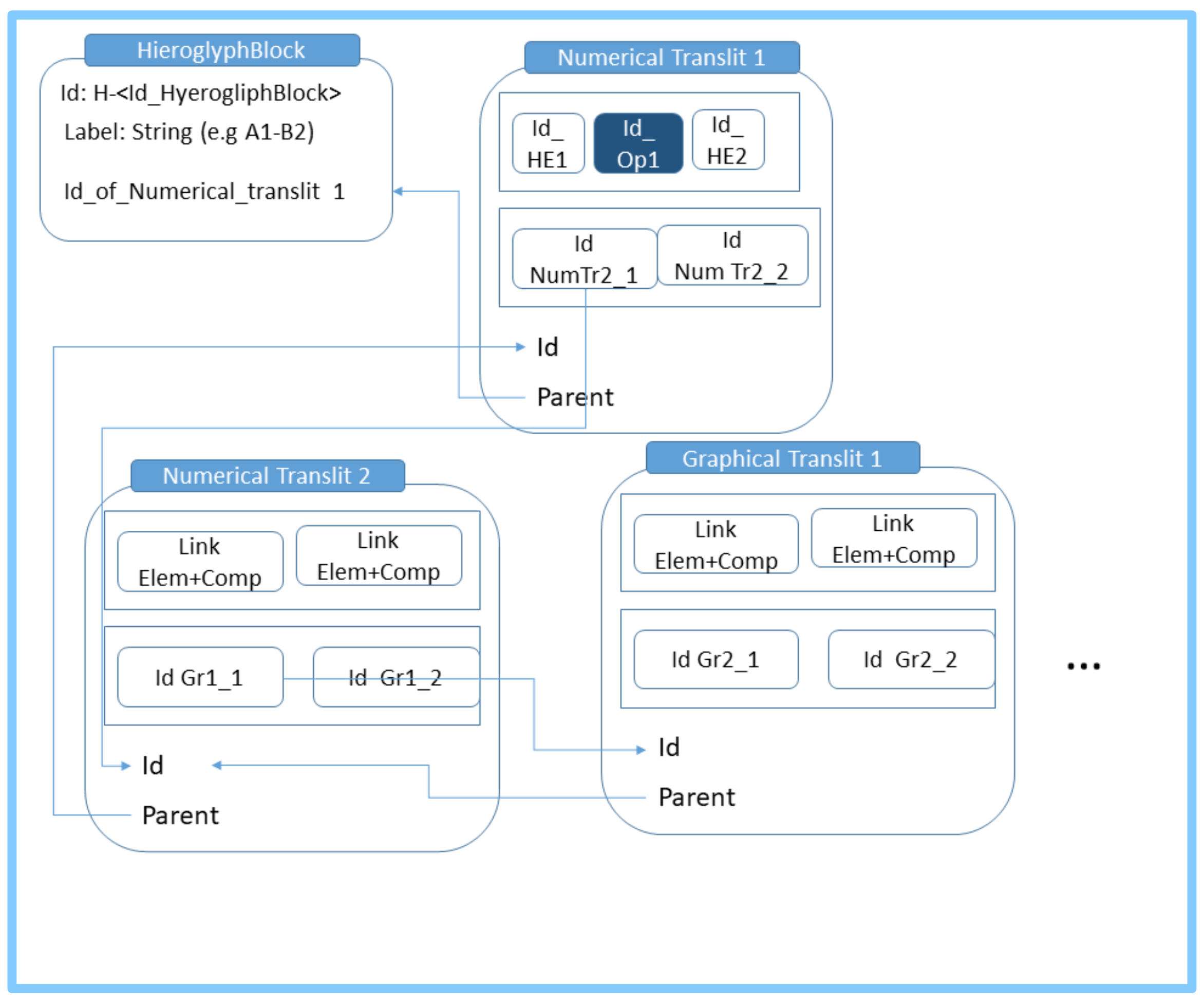
Maya

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 Zeichen ergeben. **HAB** ist eine angenommene Lesung, weil wir davon ausgehen, dass 548 und 74 ein komplexes Zeichen ergeben
Alternative 3	**TZIK HAB XOK**	–	–	–	**PAW.PILLOW-na-ja**	**TZIK** ist eine angenommene Lesung, weil wir davon ausgehen, dass 124 und 74 ein komplexes Zeichen ergeben. **HAB** ist eine angenommene Lesung, weil wir davon ausgehen, dass 548 und 28 ein komplexes Zeichen ergeben



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


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
…