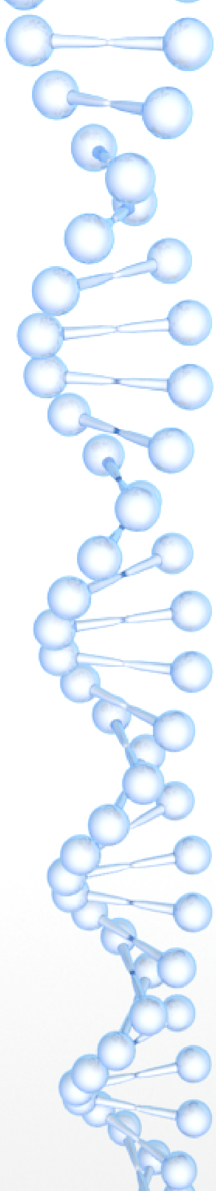


Paolo Monella
paolo.monella@unipa.it

An ontology for digital graphematics and philology

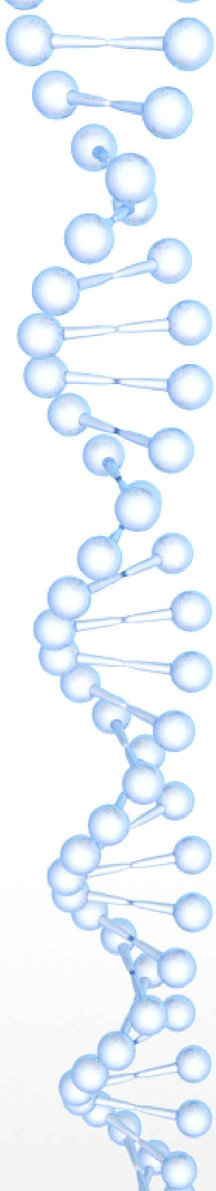
[ve]dph
Venice Centre for Digital and Public Humanities

Università Ca'Foscari Venezia
Dipartimento di Studi Umanistici

Die (hyper-)diplomatische Transkription und ihre Erkenntnispotentiale
Bergische Universität Wuppertal (BUW), 6 February 2020

Università degli Studi di Palermo

Outline

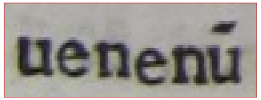


Outline

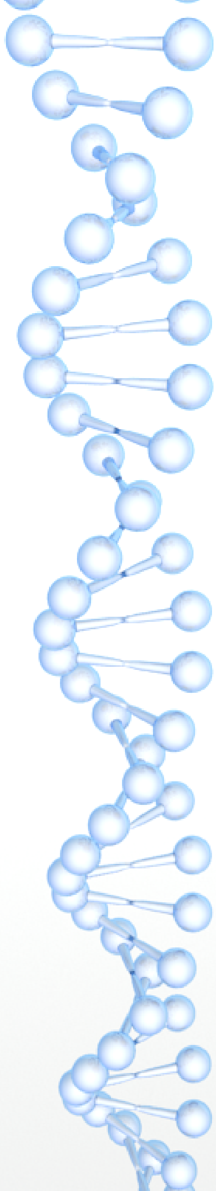
- **Interoperability**
  of digital scholarly editions (DSEs)
  based on diplomatic transcriptions
- **Digital modelling (ontology)**
  of pre-modern writing systems
  - **Graphemes / allographs**
  - **Allographs**:
    capitals, ligatures, positional variants, emphasis etc.
- **In practice**:
  how can grapheme/allograph modelling
  make my DSE more interoperable?
- **Open issues**

Interoperability: the issue

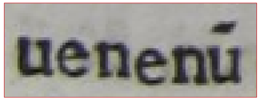
Interoperability: the issue



uenenú



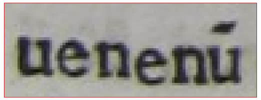
Interoperability: the issue



uenenū

- uenenū

Interoperability: the issue

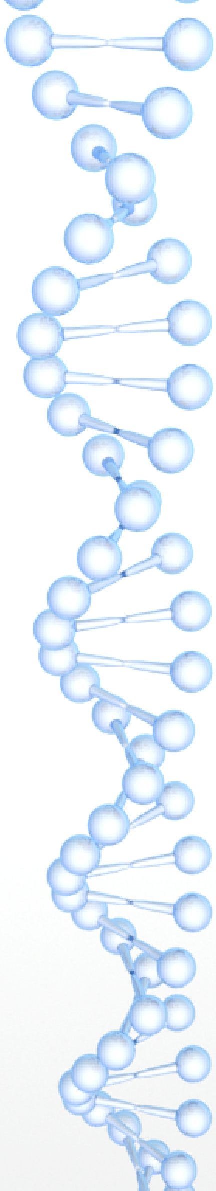


uenenū

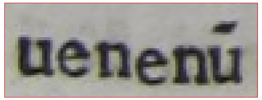
- uenenū

Diplomatic

- Historical documentation
- Visualization
- Processing
  - (Erkenntnispotentiale)

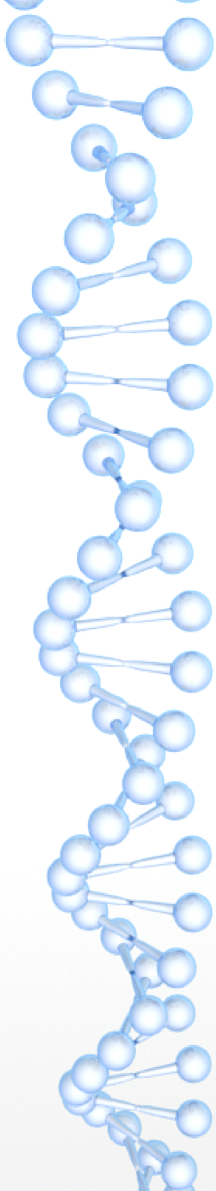


Interoperability: the issue

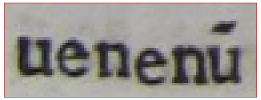


uenenū

- uenenū



Interoperability: the issue



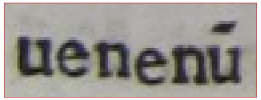
uenenũ

- uenenū

- venenum

Interoperability: the issue

- Processing
  - Search
  - Collation
  - NLP (lemma, PoS etc.)
  - Statistics (dist. reading)

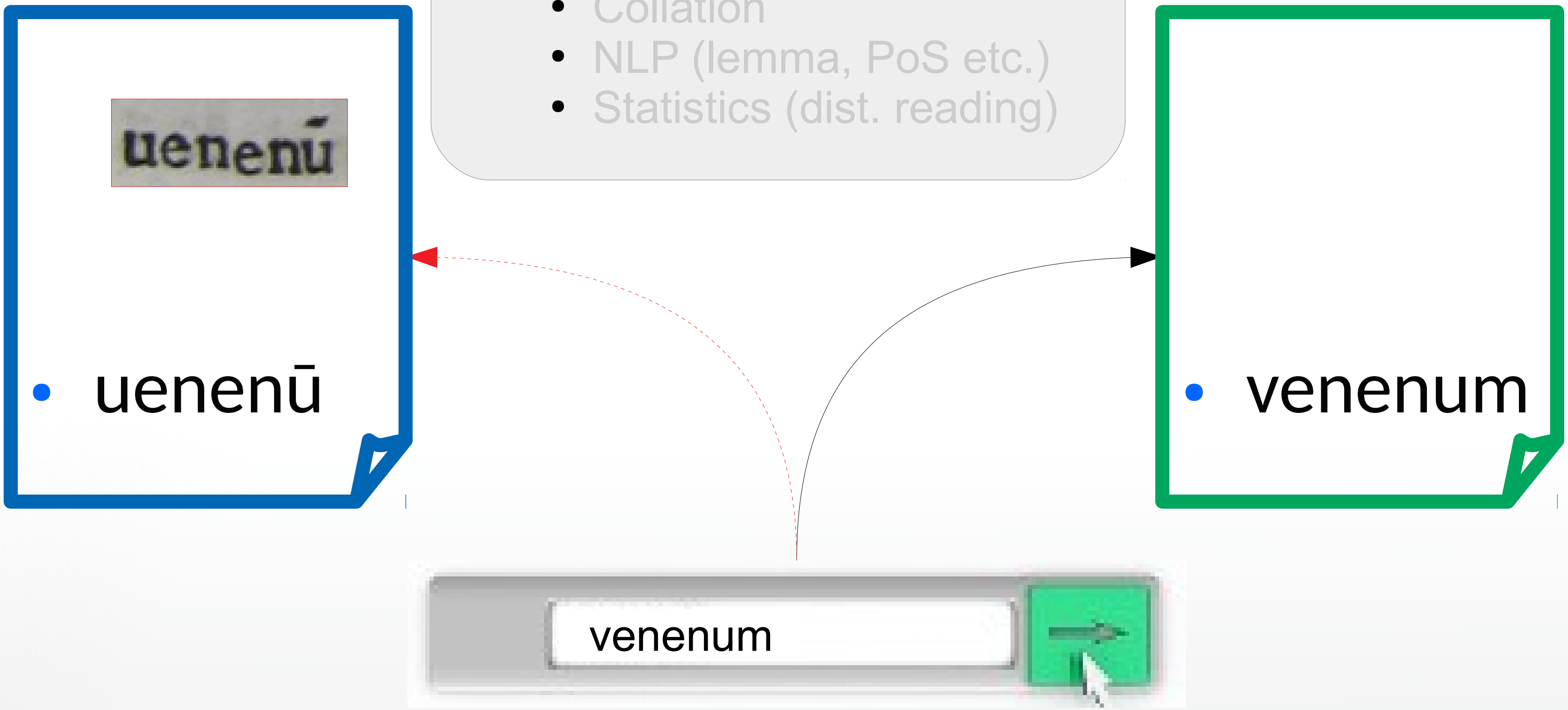


- uenenū

- venenum

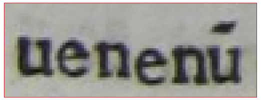
Interoperability: the issue

- Processing
  - **Search**
  - Collation
  - NLP (lemma, PoS etc.)
  - Statistics (dist. reading)



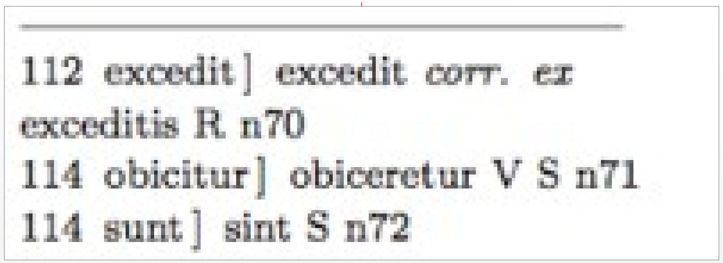
Interoperability: the issue

- Processing
  - Search
  - **Collation**
  - NLP (lemma, PoS etc.)
  - Statistics (dist. reading)

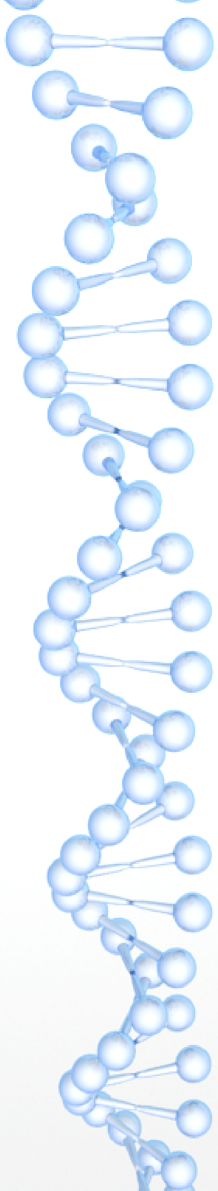


- uenenū

- venenum

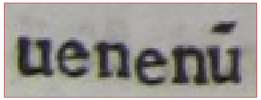


112 excedit] excedit *corr. ex* exceditis R n70
114 obicitur] obiceretur V S n71
114 sunt] sint S n72



Interoperability: the issue

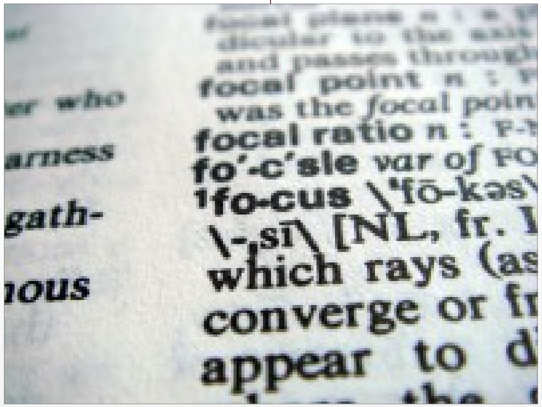
- Processing
  - Search
  - Collation
  - **NLP (lemma, PoS etc.)**
  - Statistics (dist. reading)



uenenũ

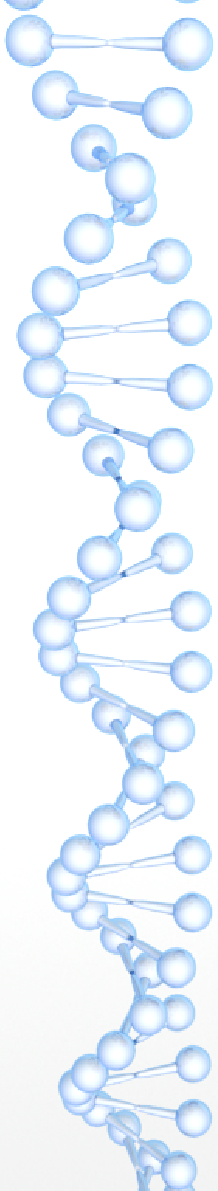
- uenenū

- venenum



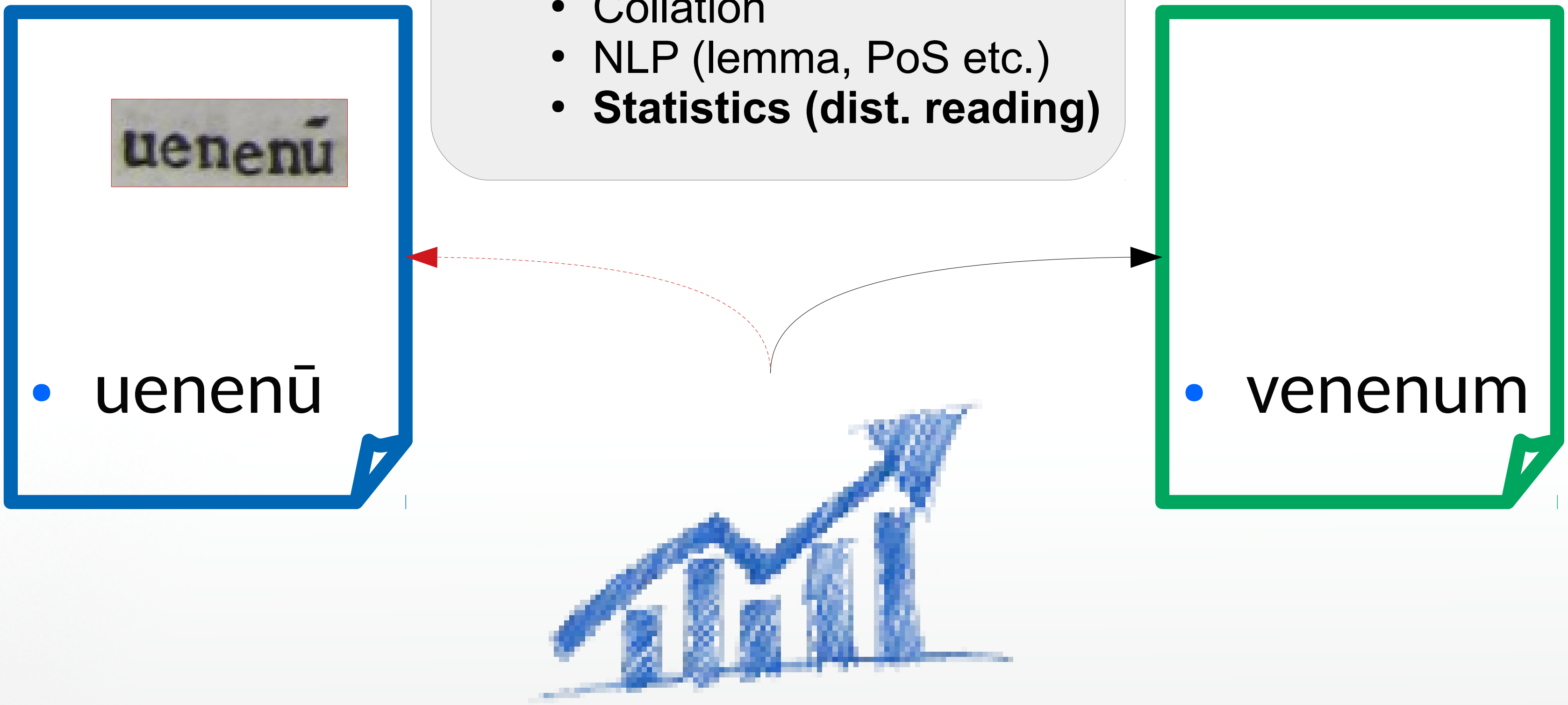
er who
arness
gath-
ious

focal plane n : a p
dicular to the axis
and passes through
focal point n : F
was the focal poin
focal ratio n : F-N
fo'c'sle var of FO
¹fo·cus \'fō-kəs\
-,sī\ [NL, fr. L
which rays (as
converge or fr
appear to d



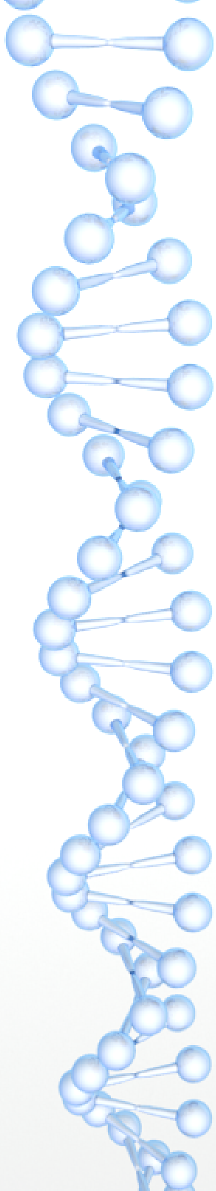
Interoperability: the issue

- Processing
  - Search
  - Collation
  - NLP (lemma, PoS etc.)
  - **Statistics (dist. reading)**



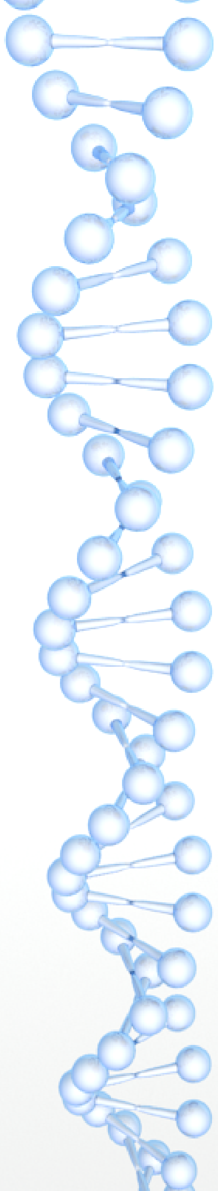
- uenenū

- venenum



Interoperability: the issue

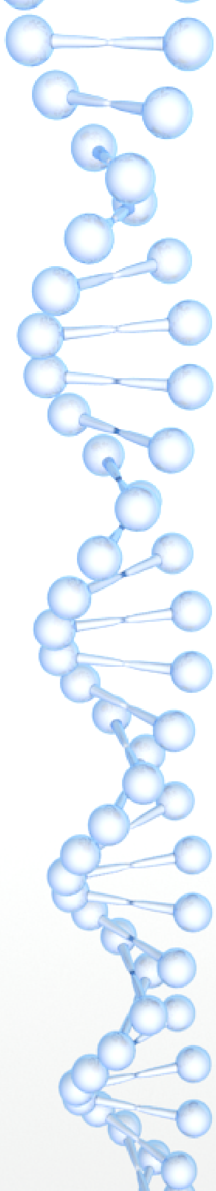
- ***My focus: European Medieval handwriting***
  - ...and early print (imitating handwriting)



Interoperability: the issue

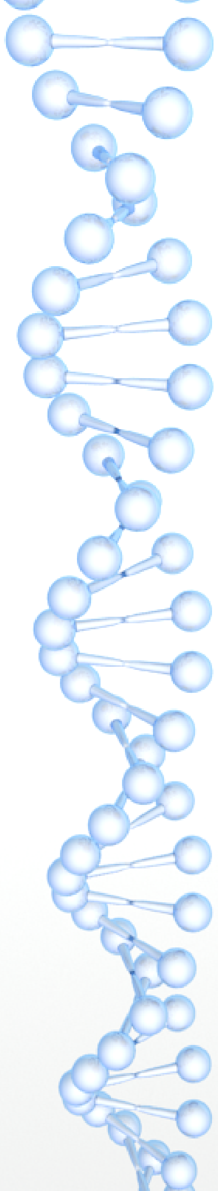
- ***My focus: European Medieval handwriting***
  - ...and early print (imitating handwriting)
  - Pre-Gutenberg (and shortly after)
- Alphabetic writing systems (so far)
  - **Latin** script (Italian, English...), Greek, Cyrillic...
  - No non-alphabetic (Cuneiform, Arabic, Chinese etc.)

Interoperability: current solutions



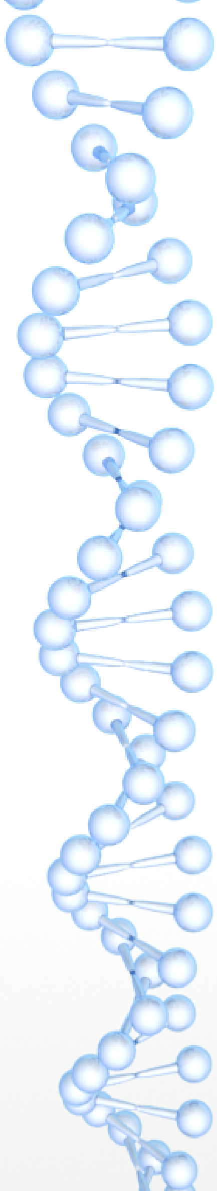
Unicode (TEI’s recommendation)

- Solution for new digital texts
- Not enough for pre-modern writing systems
  - Allographs
    - ſ (U+017F)   /   s (U+0073; ASCII 115)
    - Have I encoded that they correspond to each other (variants of grapheme <s>)?



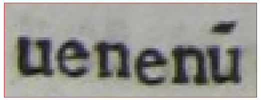
Unicode (TEI’s recommendation)

- Solution for new digital texts
- Not enough for pre-modern writing systems
  - Allographs
    - **ſ** (U+017F)   /   **s** (U+0073; ASCII 115)
    - Have I encoded that they correspond to each other (variants of grapheme <s>)?
  - Ligatures
    - **&** (U+0026; ASCII 38)
    - Have I encoded that it is equivalent to “e + t” in that MS?
  - Grapheme set
    - **u** (U+0075; ASCII 117)
    - Have I encoded whether it “covers” (or not) <u> *and* <v>?



Diplomatic/normalized: the surrender?

- venenum



uenenũ

- uenenū

Normalized

- Processing
  - Search
  - Collation
  - NLP (lemma, PoS etc.)
  - Statistics (distant reading)...

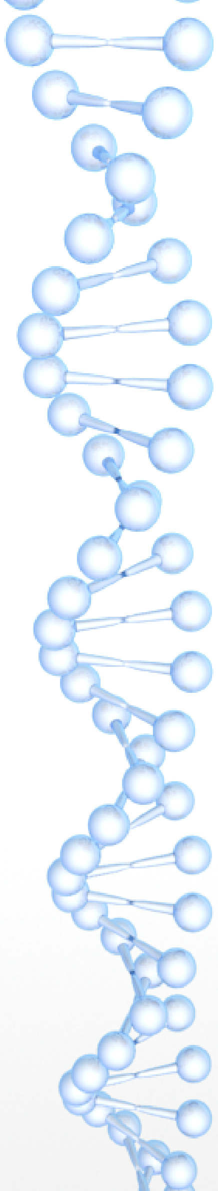
Diplomatic

- Historical documentation
- Visualization
- Processing
  - (Erkenntnispotentiale)

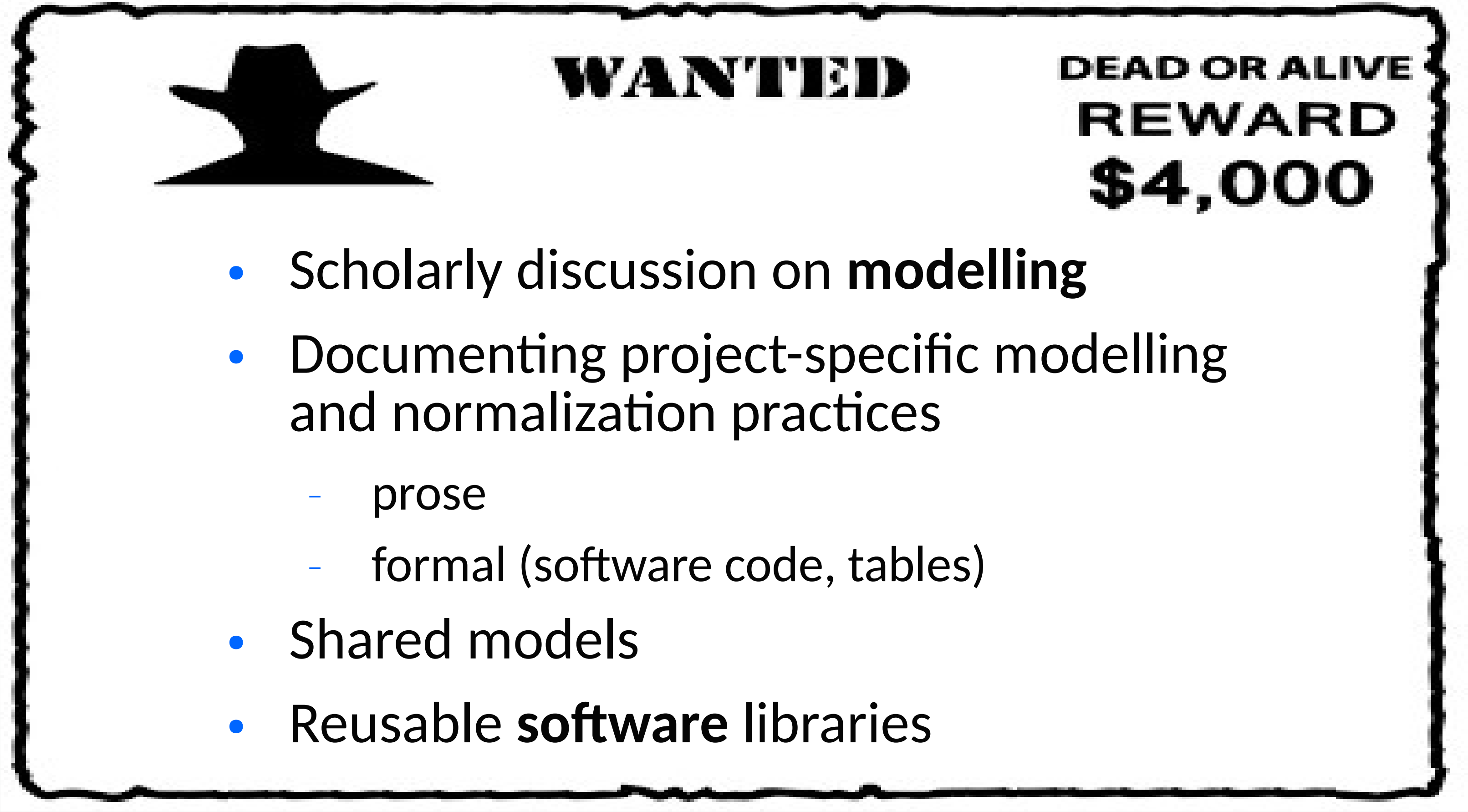
Project-specific solutions

- Disposable home-made solutions
- Normalization software and strategies
- TEI: theory-agnostic

Interoperability through modelling

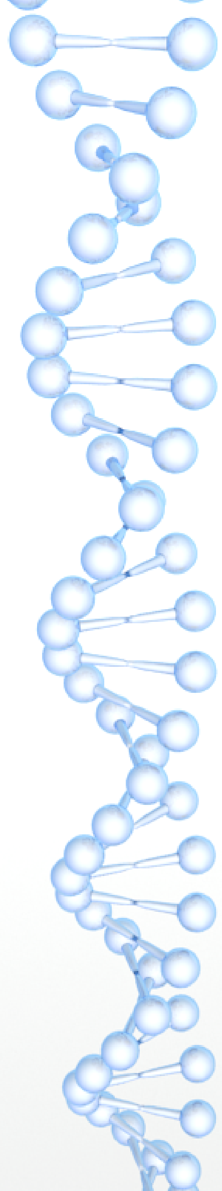


Interoperability through modelling

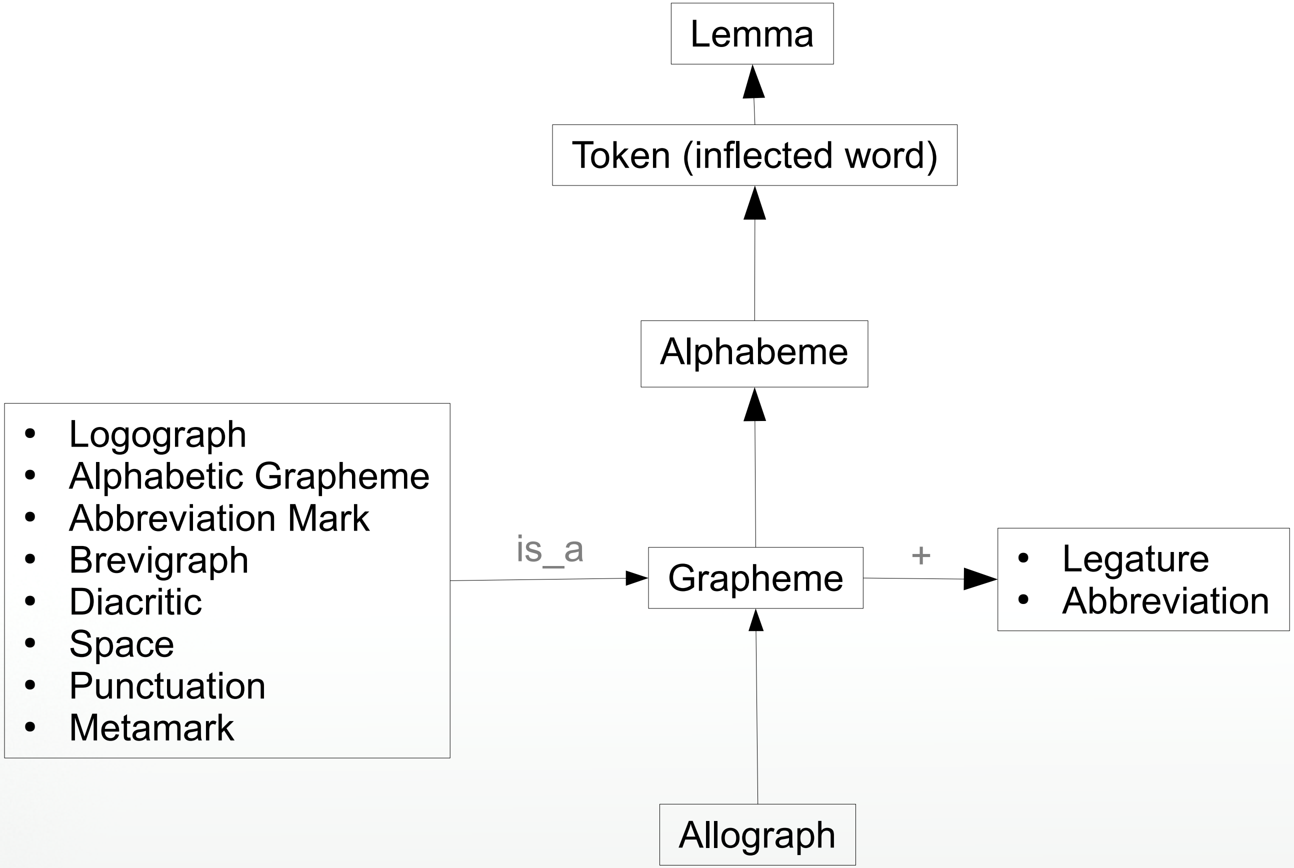


- Scholarly discussion on **modelling**
- Documenting project-specific modelling and normalization practices
  - prose
  - formal (software code, tables)
- Shared models
- Reusable **software** libraries

An ontology
for digital graphematics
and philology



Ontology



Lemma

Token (inflected word)

Alphabeme

- Logograph
- Alphabetic Grapheme
- Abbreviation Mark
- Brevigraph
- Diacritic
- Space
- Punctuation
- Metamark

is_a

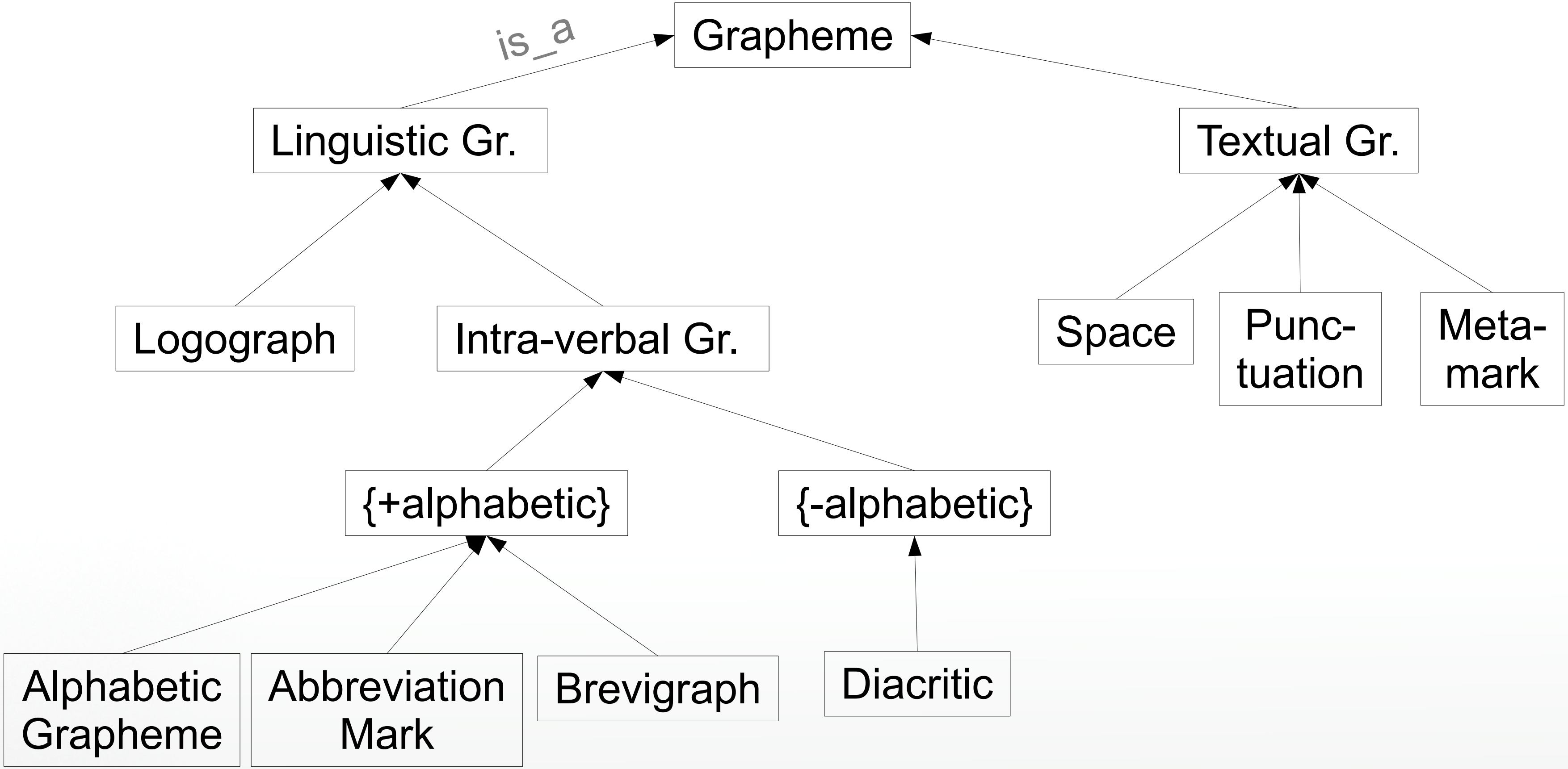
Grapheme

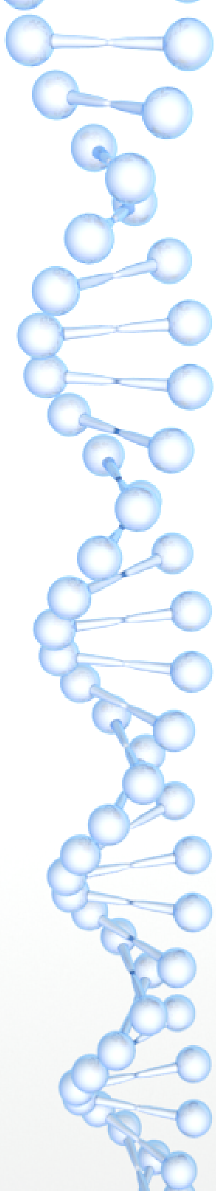
+

- Legature
- Abbreviation

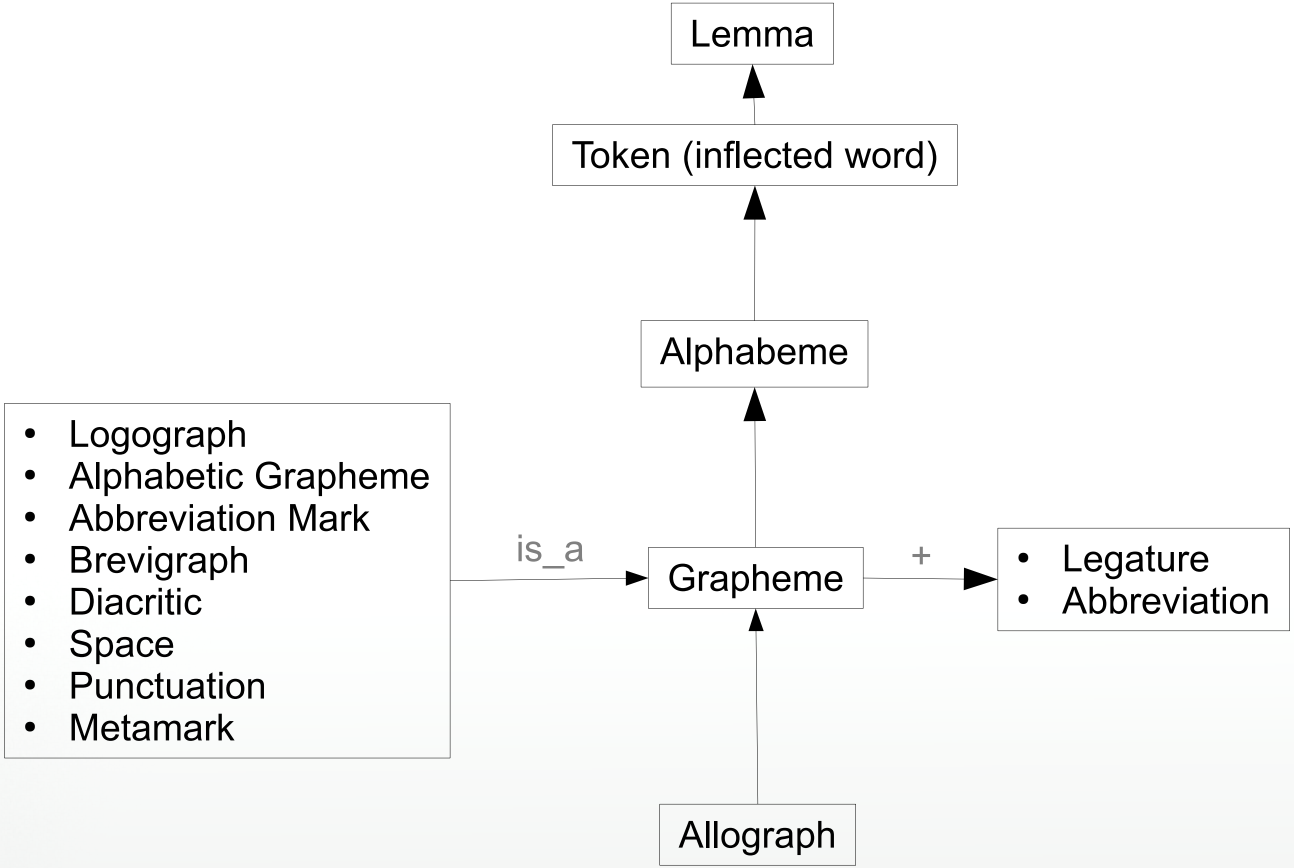
Allograph

Ontology


is_a
Grapheme
Linguistic Gr.
Textual Gr.
Logograph
Intra-verbal Gr.
Space
Punc-tuation
Meta-mark
{+alphabetic}
{-alphabetic}
Alphabetic Grapheme
Abbreviation Mark
Brevigraph
Diacritic



Ontology



Lemma

Token (inflected word)

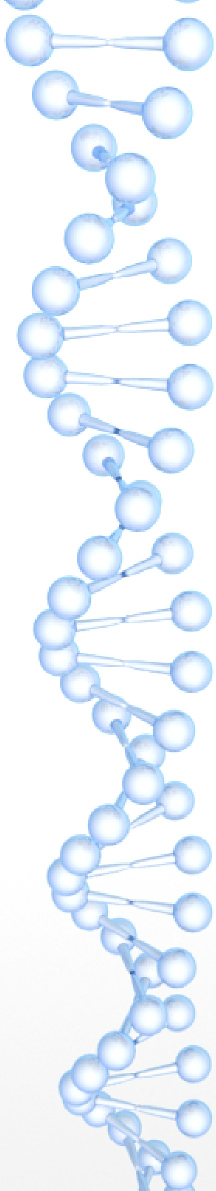
Alphabeme

- Logograph
- Alphabetic Grapheme
- Abbreviation Mark
- Brevigraph
- Diacritic
- Space
- Punctuation
- Metamark

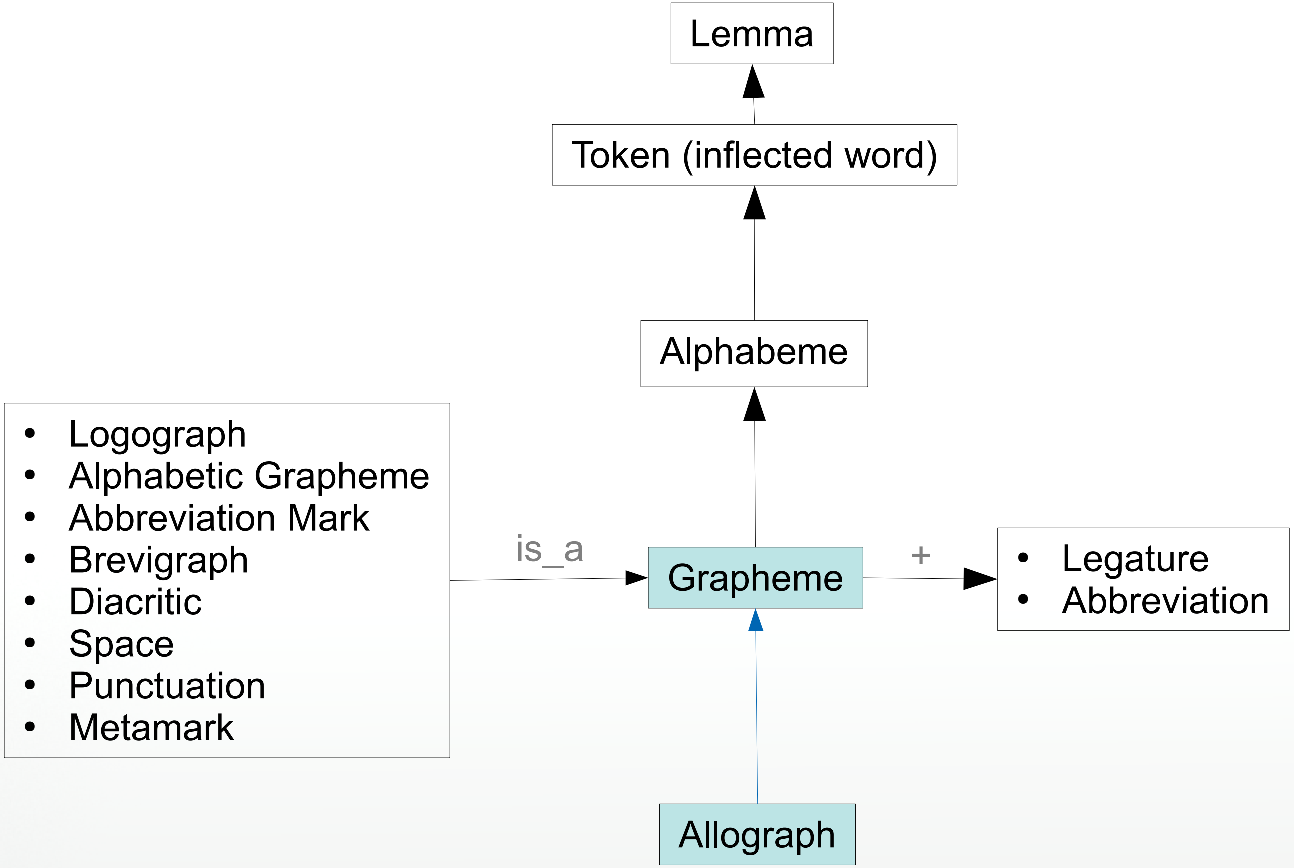
Grapheme

- Legature
- Abbreviation

Allograph



Ontology



Lemma

Token (inflected word)

Alphabeme

- Logograph
- Alphabetic Grapheme
- Abbreviation Mark
- Brevigraph
- Diacritic
- Space
- Punctuation
- Metamark

is_a

Grapheme

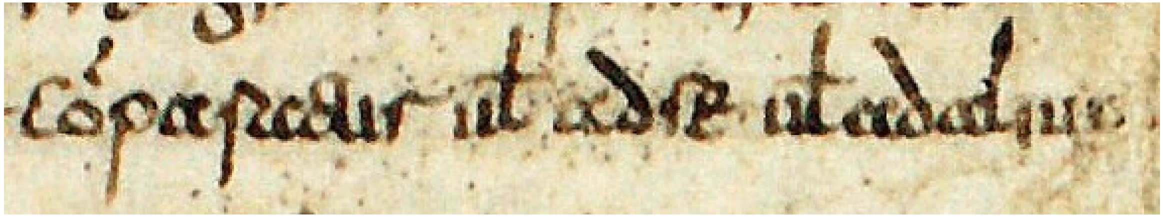
+

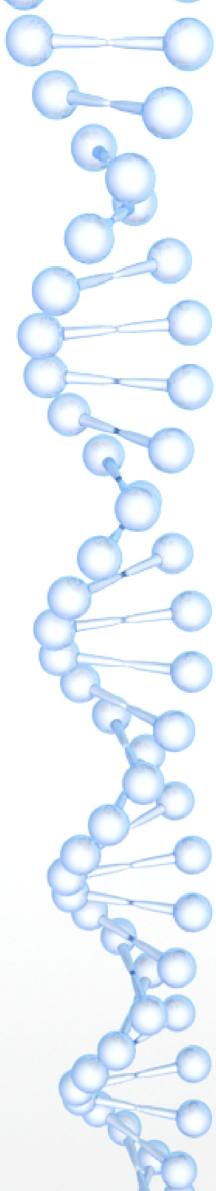
- Legature
- Abbreviation

Allograph

Digital modelling for pre-modern writing systems

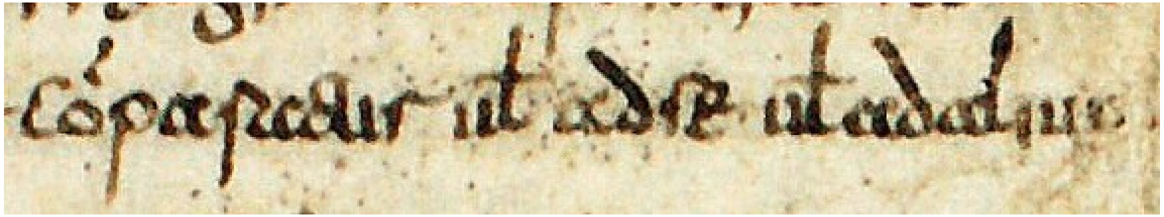
Digital modelling





Digital modelling

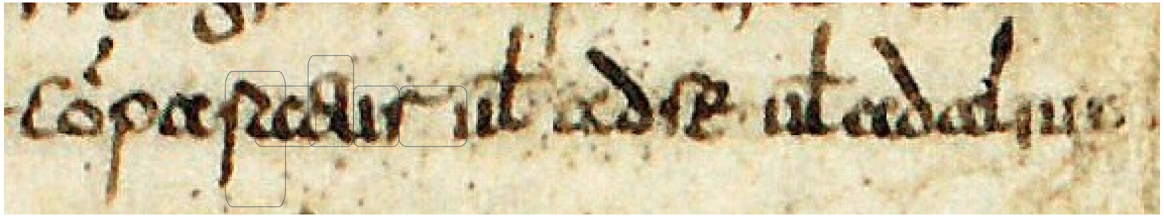
- Comparatur vel ad se vel ad alium
  He is compared to himself or to another



- cỏparaƐur uł adſe uładalium

Digital modelling

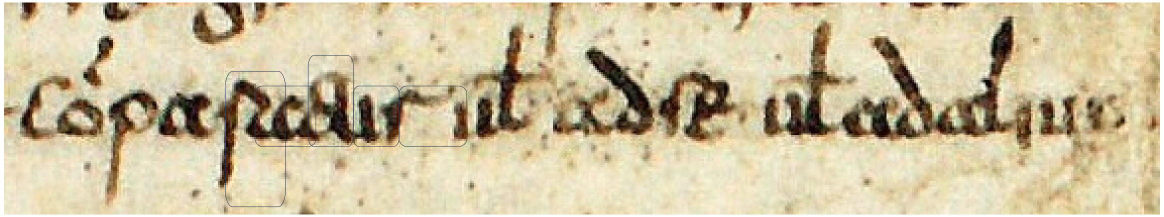
- Comparatur vel ad se vel ad alium
  He is compared to himself or to another



- cȯparaƐur uł adſe uładalium

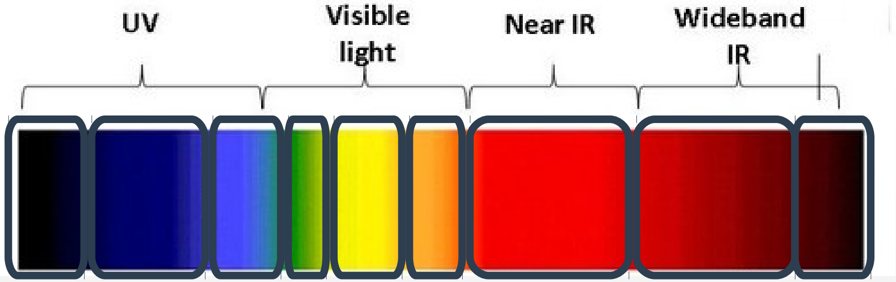
Digital modelling

- Comparatur vel ad se vel ad alium
  He is compared to himself or to another

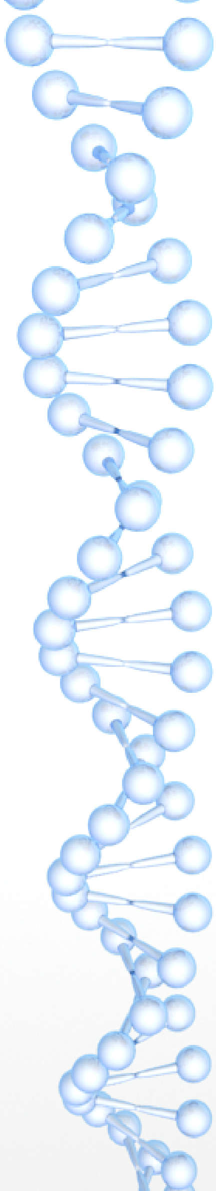


- cỏparaƐur uł adſe uładalium

Digital modelling

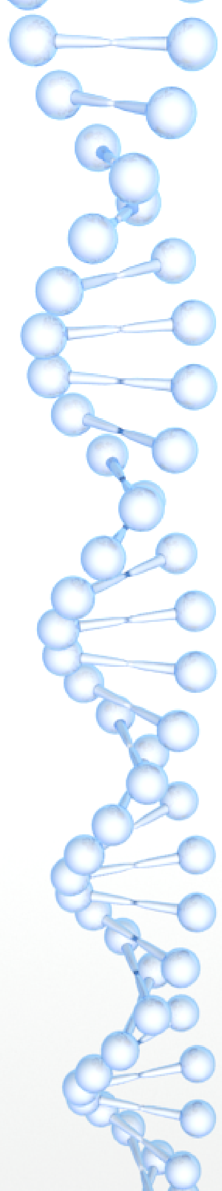


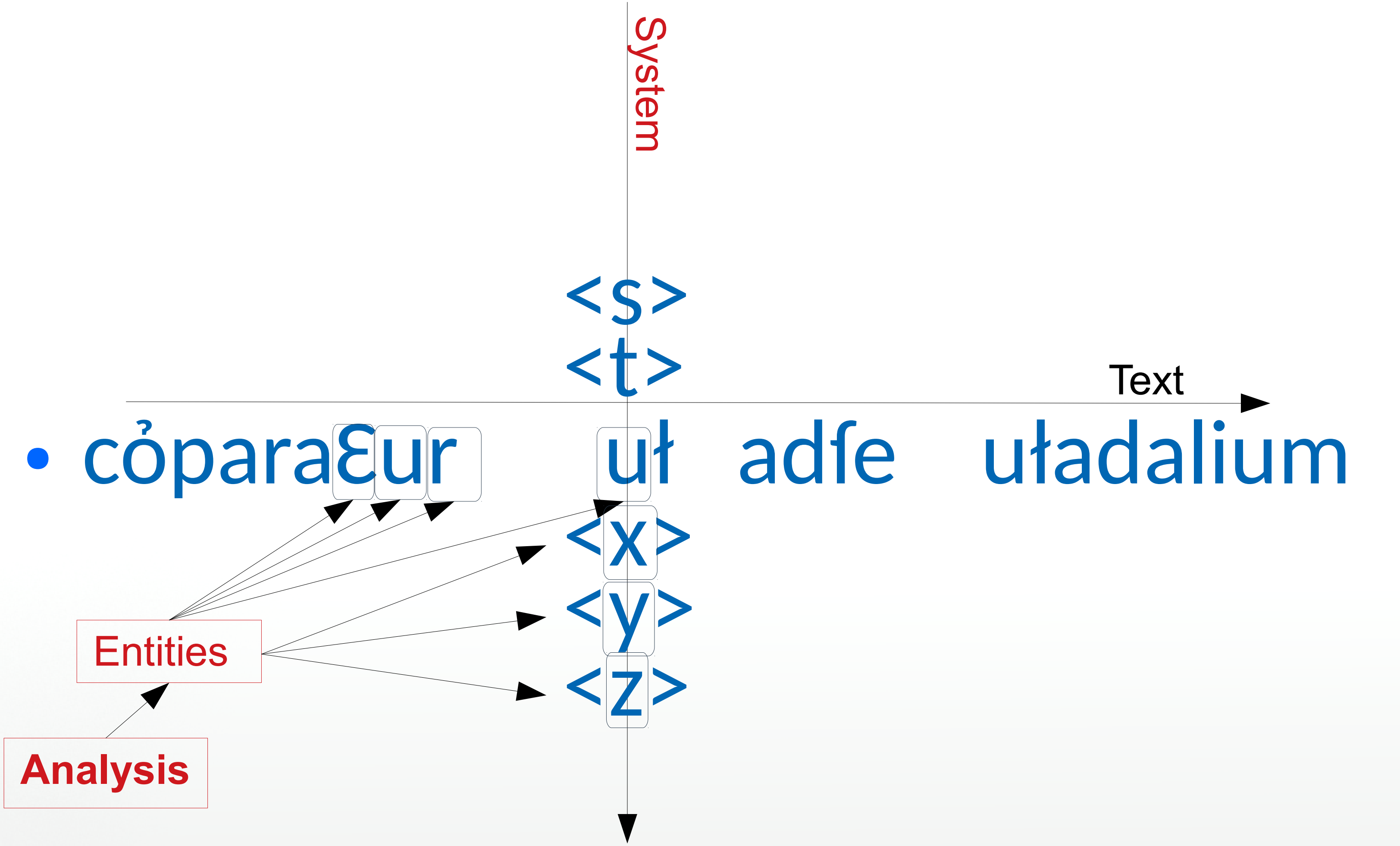
Digital modelling



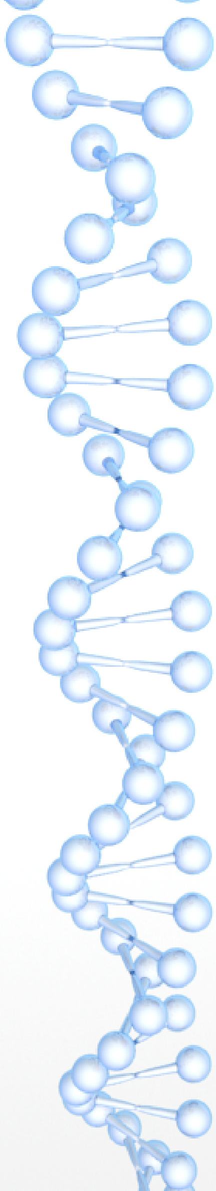
- cỏparaƐur uł adſe uładalium

A structural approach to digital modelling

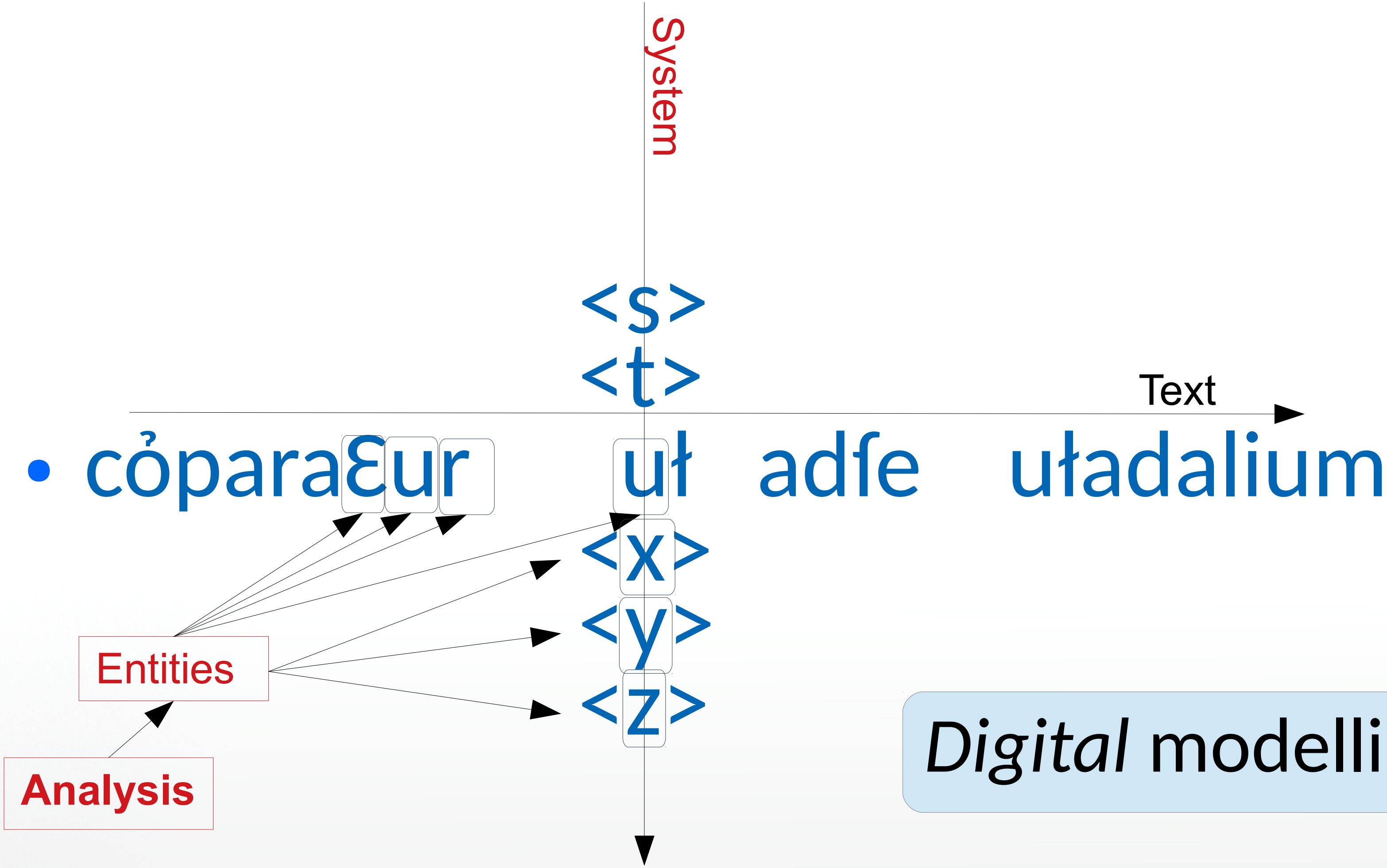




- cỏparaƐur uł adſe uładalium



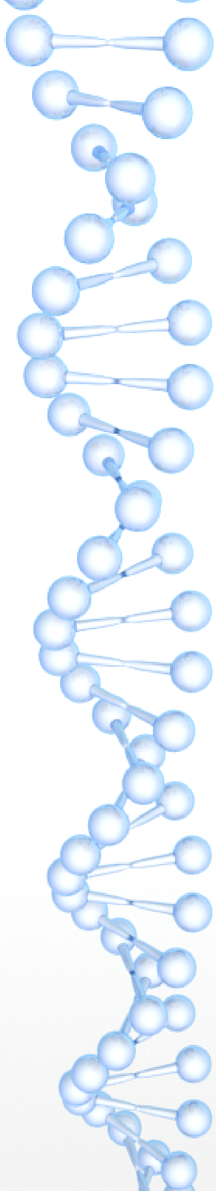
A structural approach to digital modelling



- cỏparaƐur uł adſe uładalium

Digital modelling

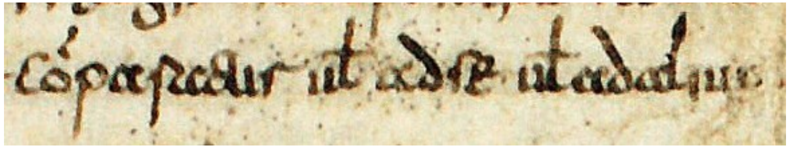
Graphemes/allographs

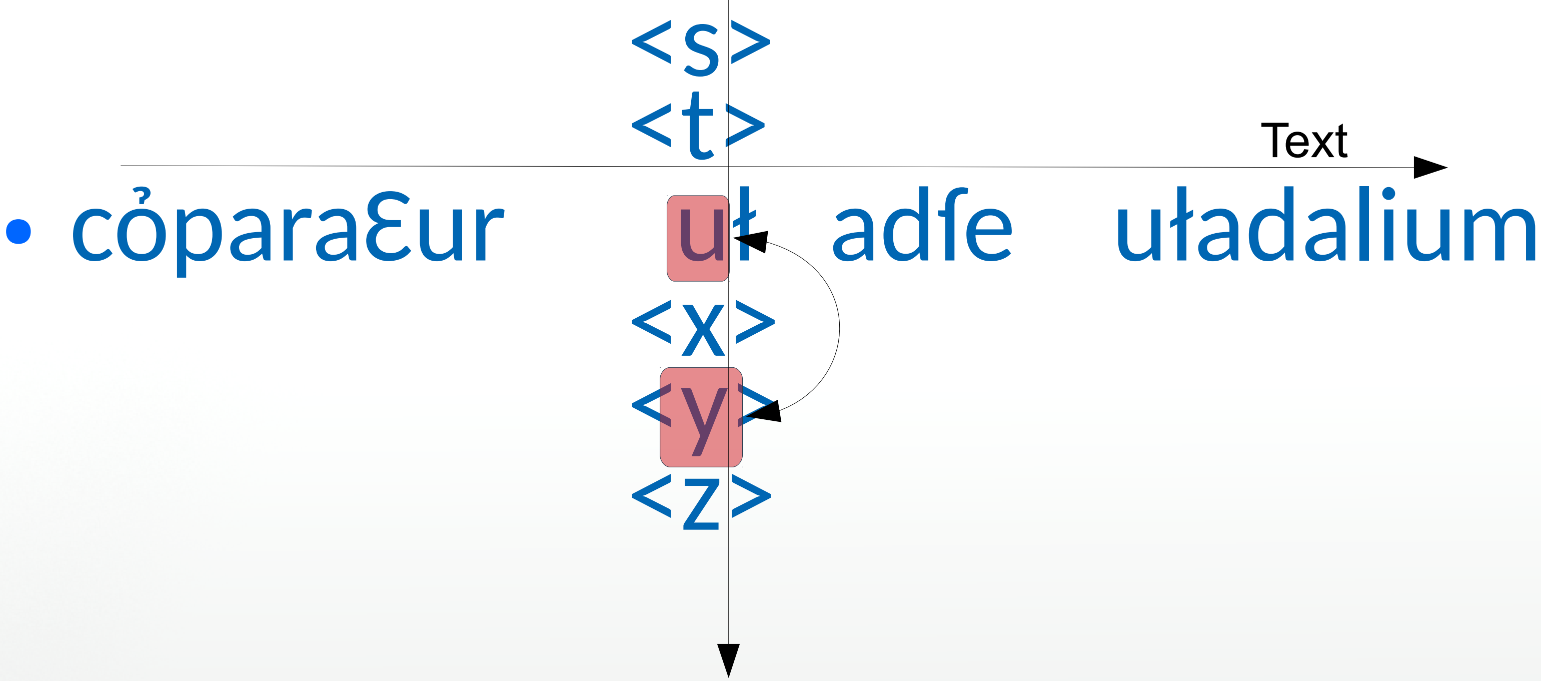


Graphemes/allographs: the commutation test

Comparatur vel ad se vel ad alium

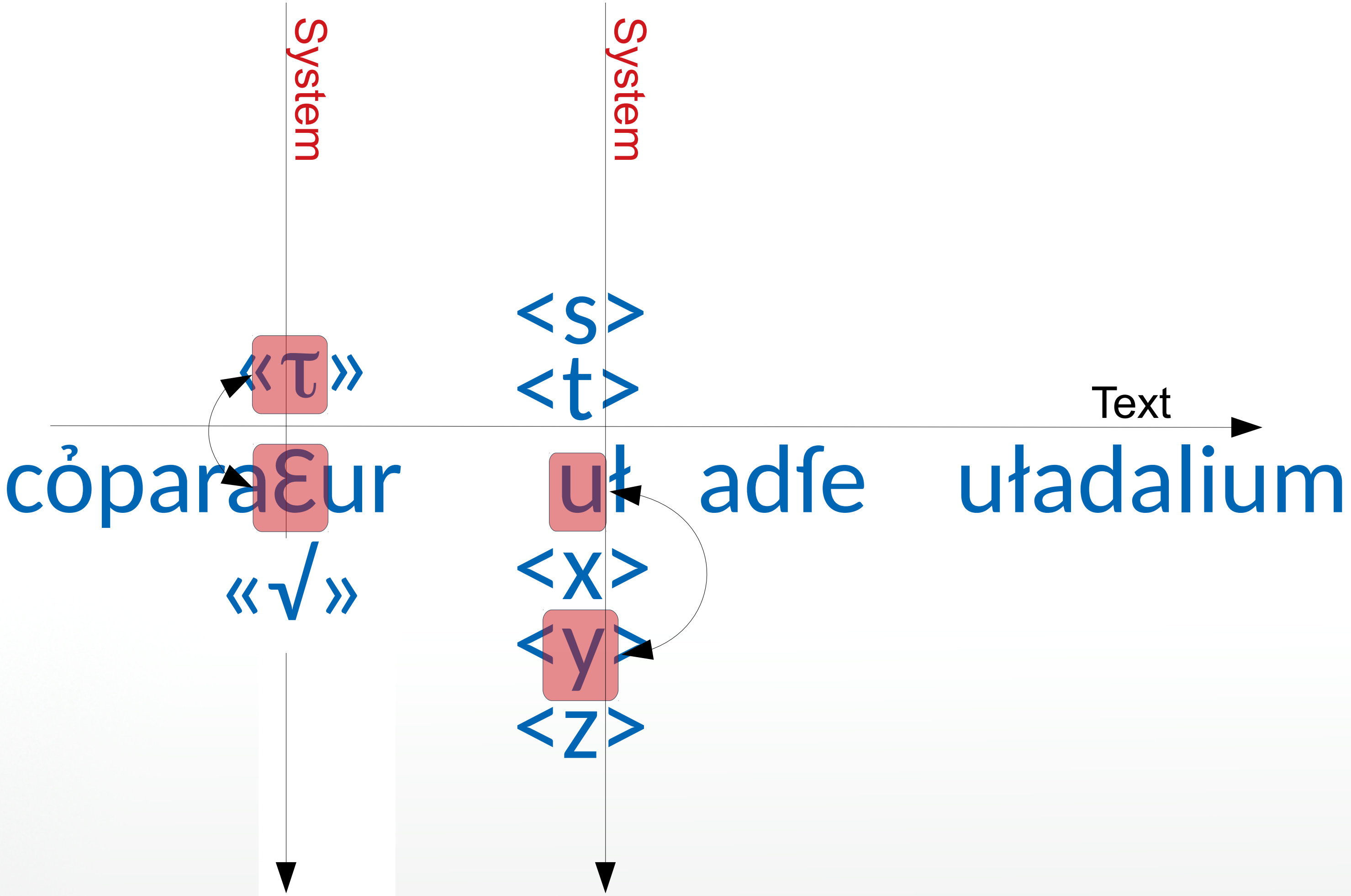
He is compared to himself or to another





- cỏparaƐur uł adſe uładalium

Graphemes/allographs: the commutation test



System

System

<s>

«τ»

<t>

Text

- cỏparaƐur uł adſe uładalium

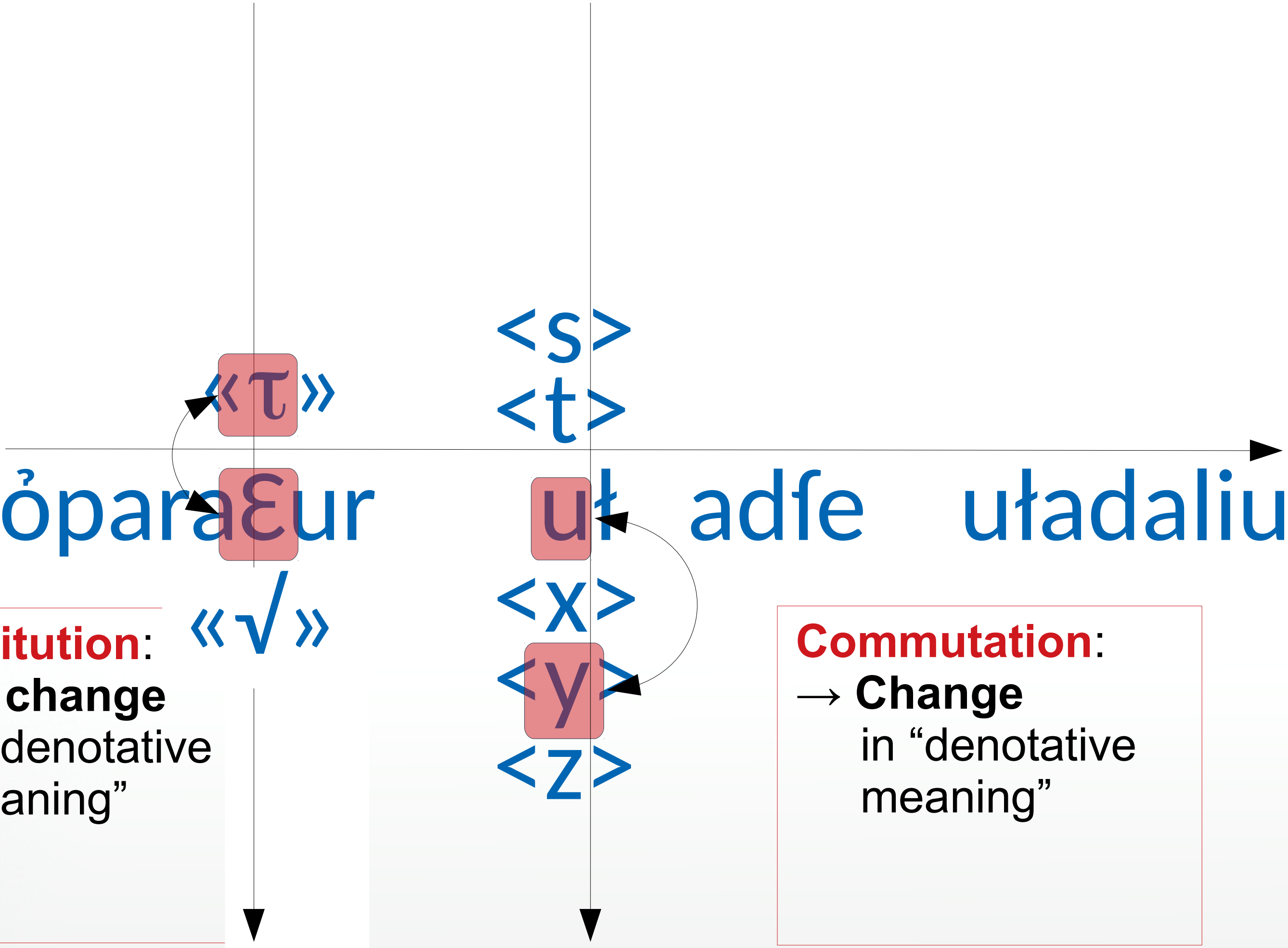
«√»

<x>

<y>

<z>

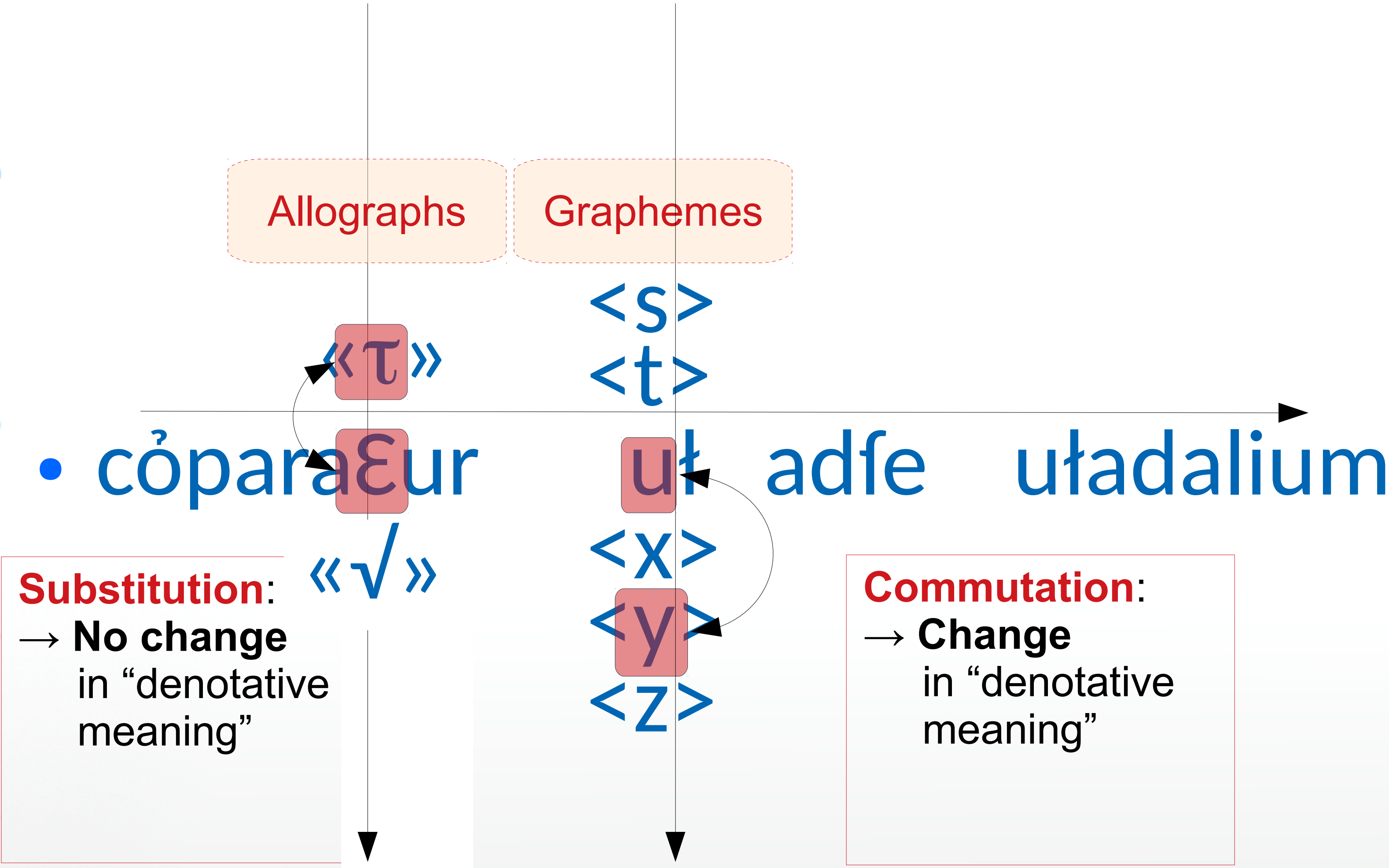
Graphemes/allographs: the commutation test



Substitution:
→ **No change**
in “denotative meaning”

Commutation:
→ **Change**
in “denotative meaning”

Graphemes/allographs: the commutation test



Allographs

Graphemes

<s>

«τ»

<t>

- cỏparaƐur uł adſe uładalium

«√»

<x>

<y>

<z>

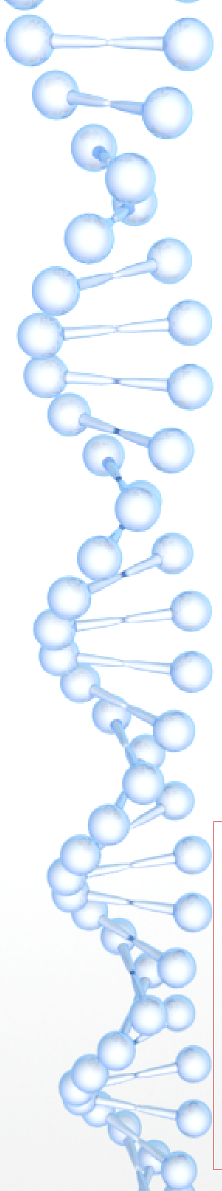
Substitution:

→ **No change**
in “denotative meaning”

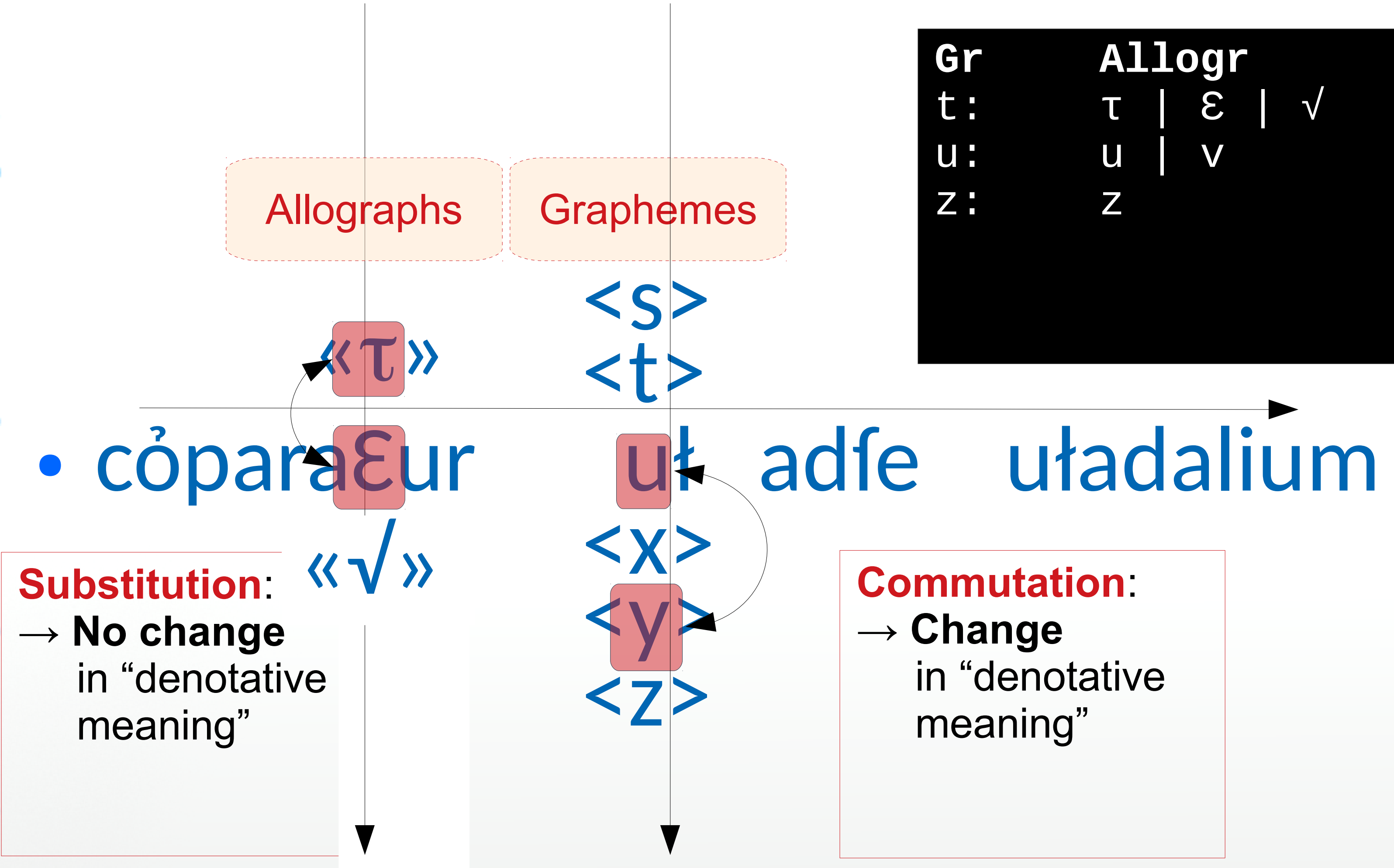
Commutation:

→ **Change**
in “denotative meaning”

Graphemes/allographs: the commutation test



```
Gr       Allogr
t:       τ | ℇ | √
u:       u | v
z:       z
```



Substitution:
→ **No change** in “denotative meaning”

Commutation:
→ **Change** in “denotative meaning”

Graphemes / allographs: what to transcribe?

- What the project wants!
  - based on its scientific interests
  - (and on time / money)
- But: framed in a larger model

Saussure, pertinence and the scribe’s toolbox

Saussure, pertinence and the scribe’s toolbox

MS A

a b c d e f g h i l m n o p q r s t u z · ;

Ceci n'est pas un linguiste.

MS B

a b c d e f g h i j l m n o p q r s t u v z . , ; : !

Saussure, pertinence and the scribe’s toolbox

OCR from Teubner

a b c d e f g h i l m n o p q r s t u z · ;

Ceci n'est pas un linguiste.

OCR from Loeb

a b c d e f g h i j l m n o p q r s t u v z . , ; : !

Saussure, pertinence and the scribe’s toolbox

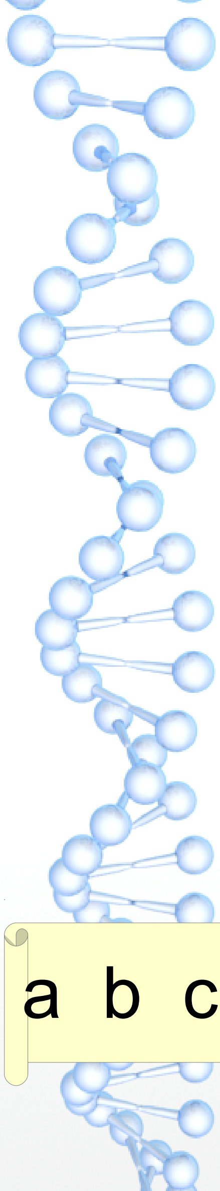
- The toolbox of the scribe
  - Definition of graphemes, allographs…
- Writing systems as autonomous semiotic systems (Sampson)
  - Not as epiphenomena of oral language (phonemes)
  - Mandarin / cantonese
  - “Opaque” orthographies (English)
    - “knight”, “aile”, “read”, “read” (past tense)
  - Medieval MSS: pronunciation?

a b c d e f g h **i j** l m n o p q r s t **u v** z **. , ; : !**

Saussure, pertinence and the scribe’s toolbox

- “In language there are only differences” (Saussure)
  - “But the statement that everything in language is negative is true only **if the signified and the signifier are considered separately**; when we consider the sign in its totality, we have something that is **positive** in its own class”

a b c d e f g h **i j** l m n o p q r s t **u v** z **. , ; : !**



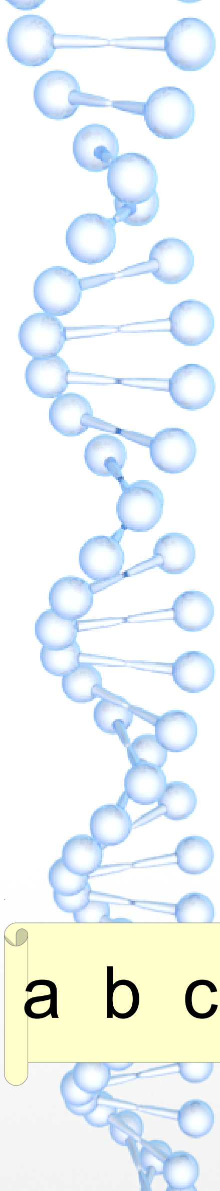
Saussure, pertinence and the scribe’s toolbox

- Can we define the scribe’s (graphematic, signifier) toolbox under complete ignorance of the linguistic (meaning, signified) dimension?

a b c d e f g h **i j** l m n o p q r s t **u v** z **. , ; : !**

Saussure, pertinence and the scribe's toolbox

- Can we define the scribe's toolbox under complete ignorance of the linguistic dimension?

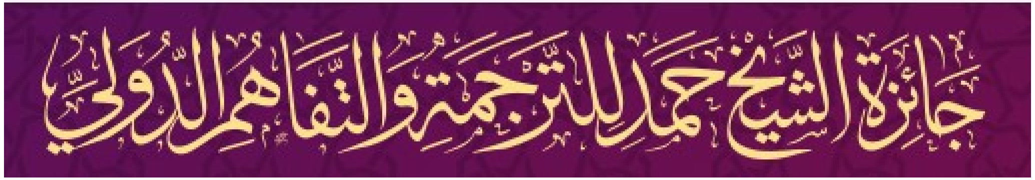


a b c d e f g h **i j** l m n o p q r s t **u v** z **. , ; : !**

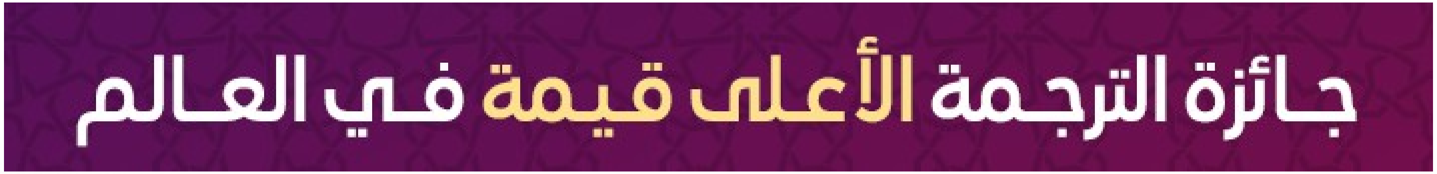
Saussure, pertinence and the scribe’s toolbox

- Can we define the scribe’s toolbox under complete ignorance of the linguistic dimension?

Segmentation



جائزة الشيخ حمد للترجمة والتفاهم الدولي



جائزة الترجمة الأعلى قيمة في العالم

a b c d e f g h i j l m n o p q r s t u v z . , ; : !

Saussure, pertinence and the scribe’s toolbox

- Can we define the scribe’s toolbox under complete ignorance of the linguistic dimension?

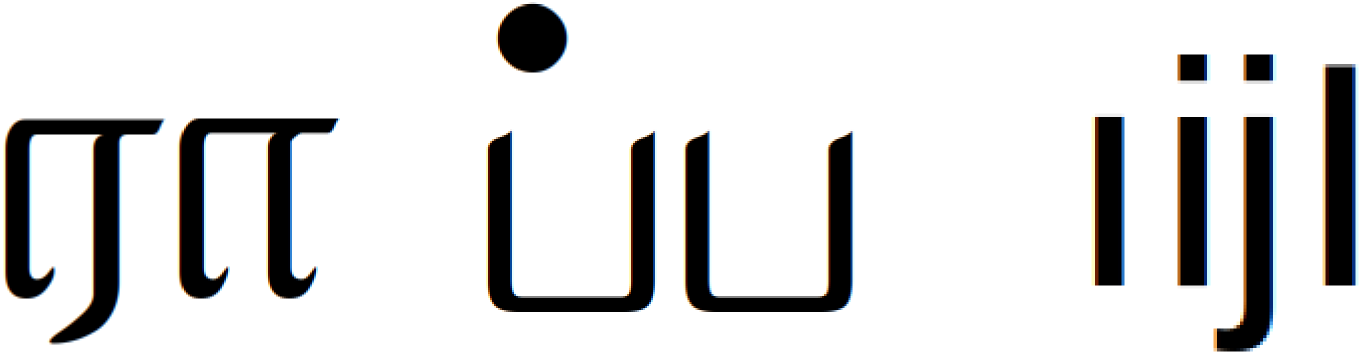
Devanāgarī

அடைந்தது பேனா பிடிக்கும் கை தடியையும் பிடிக்கும் -
ஈ.வி.கே.எஸ். இளங்கோவன் வரும் தேர்தலில் ஐபேக்
நிறுவனத்துடன் இணைந்து பணியாற்ற உள்ளோம்: ஸ்டாலின்
பிரசாந்த் கிஷோரின் நிறுவனத்துடன் இணைந்து

a b c d e f g h **i j** l m n o p q r s t **u v** z **. , ; : !**

Saussure, pertinence and the scribe’s toolbox

- Can we define the scribe’s toolbox under complete ignorance of the linguistic dimension?

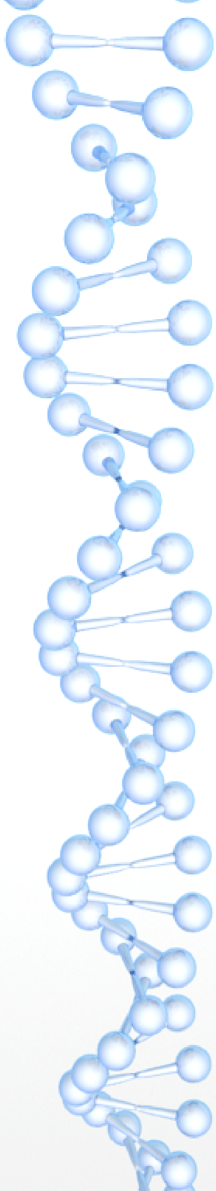


Devanāgarī

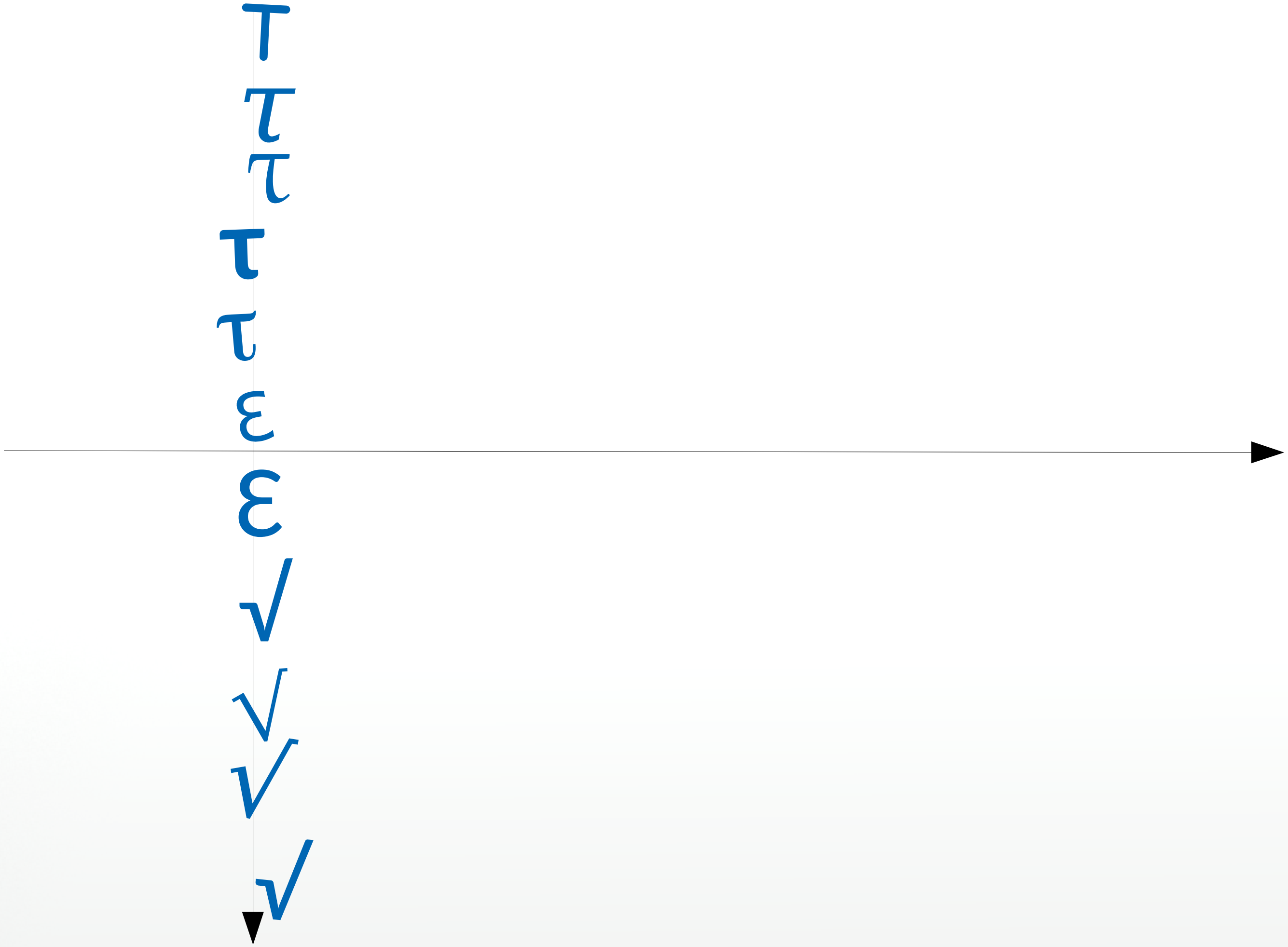
Turkish, Latin, Italian, English

a b c d e f g h **i j** l m n o p q r s t **u v** z **. , ; : !**

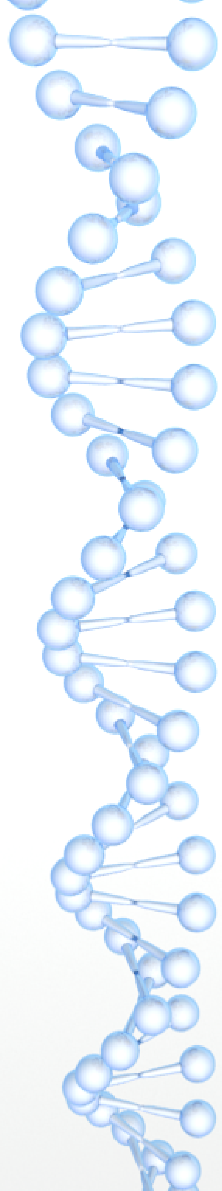
Can allographs have a distinctive value?

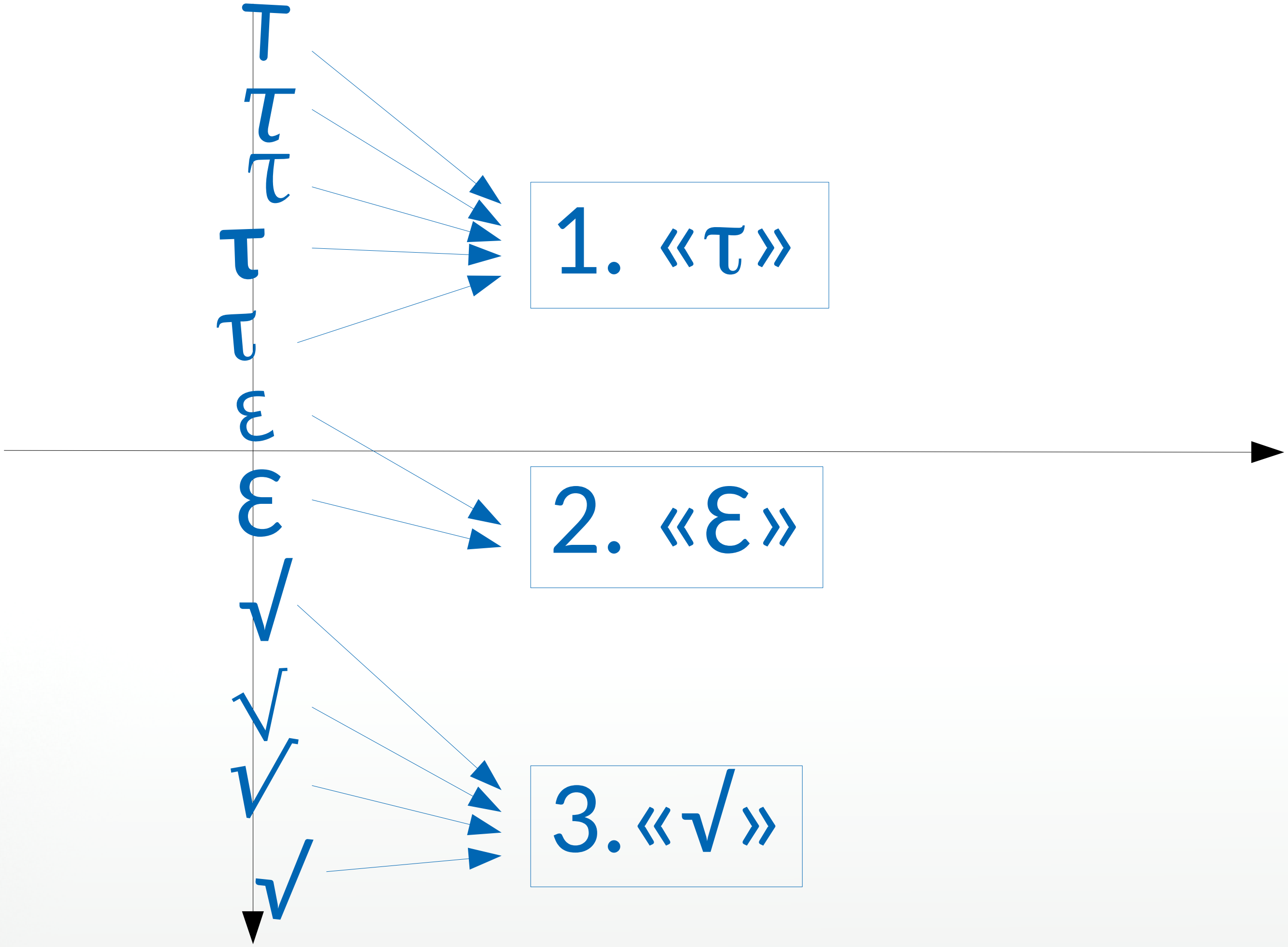


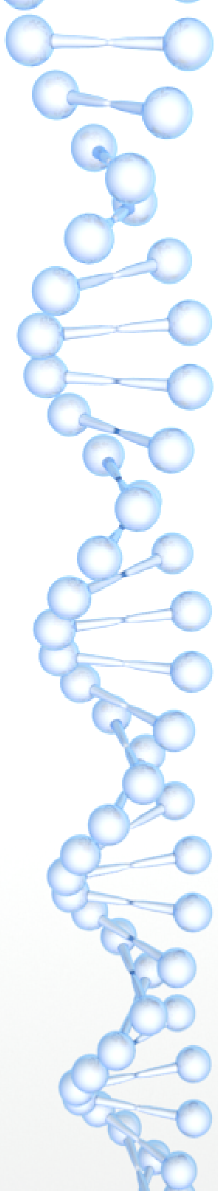
Allographs



Allographs







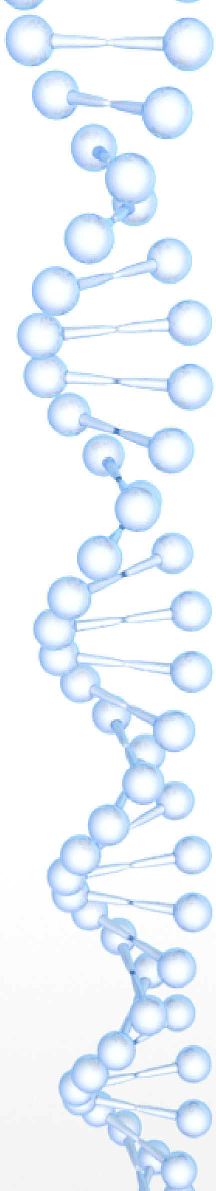
Capitals: allographs or graphemes?

• Cool (CA) is a cool town	*Geographical name*
• Smith is a good smith	*Proper name*
• ODD files are odd files	*Acronym*

⚠

OK for contemporary Western writing systems

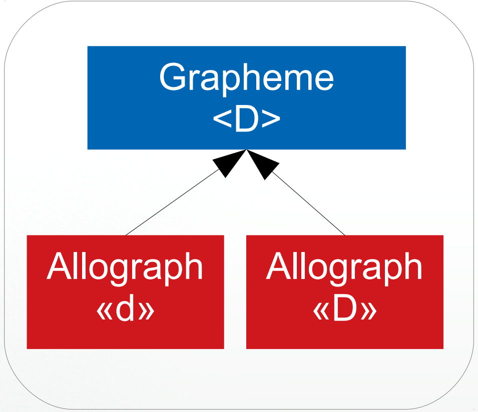
Not for classical/medieval handwriting (see later)



Capitals: allographs or graphemes?

• Cool (CA) is a cool town	*Geographical name*
• Smith is a good smith	*Proper name*
• ODD files are odd files	*Acronym*

R. Mordenti

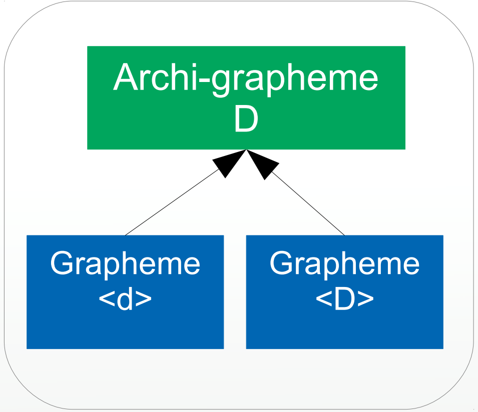


Grapheme <D>

Allograph «d»

Allograph «D»

F. Neuber

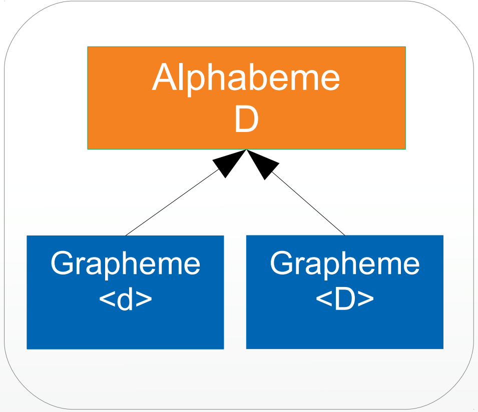


Archi-grapheme D

Grapheme <d>

Grapheme <D>

P. Monella



Alphabeme D

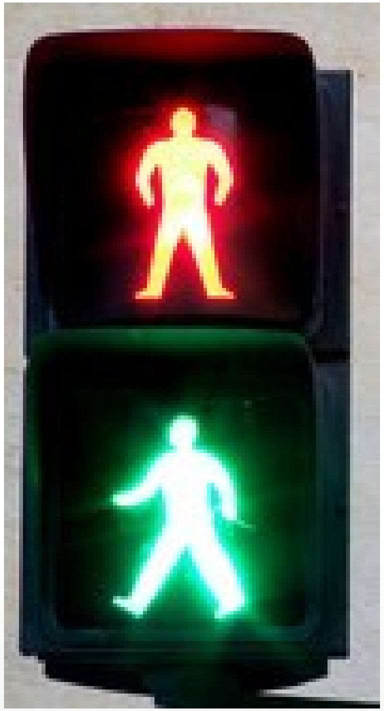
Grapheme <d>

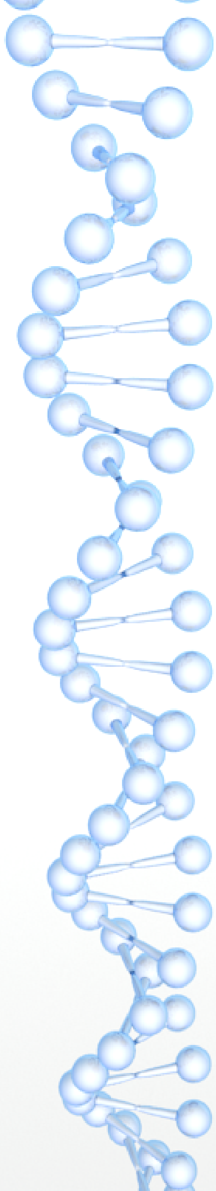
Grapheme <D>

Sentence segmentation: distinctive value for meaning of the whole text

- I go because I have to. Stay here!
I go because I have to stay here!

Capitals



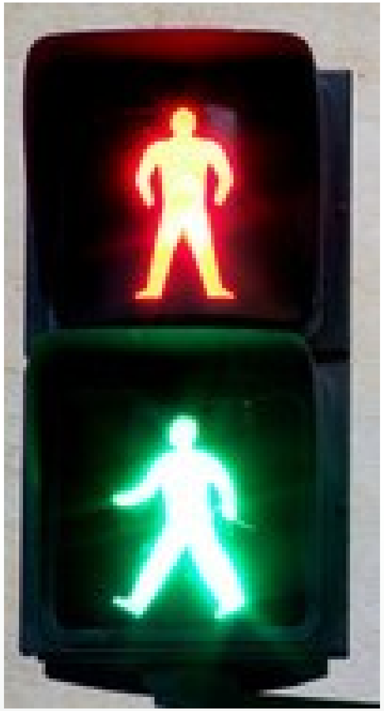


Sentence segmentation: distinctive value for meaning of the whole text

- I go because I have to. Stay here!
  I go because I have to stay here!

Punctuation

Capitals



Word segmentation: distinctive value for meaning of the whole text

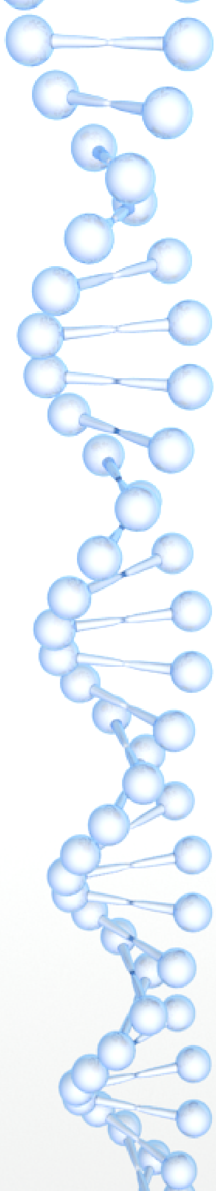
- σαῦρος, ſucceſs, daſs (daß)

Word segmentation: distinctive value for meaning of the whole text

- σαῦρος, ſucceſs, daſs (daß)

  Paulus suſtinet me (*Paolo holds me up*)
  Paulus ſus tinet me (*Paolo the pig holds me*)

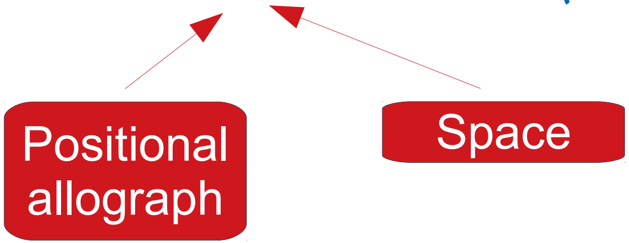
Positional allograph



Word segmentation: distinctive value for meaning of the whole text

- σαῦρος, ſucceſs, daſs (daß)

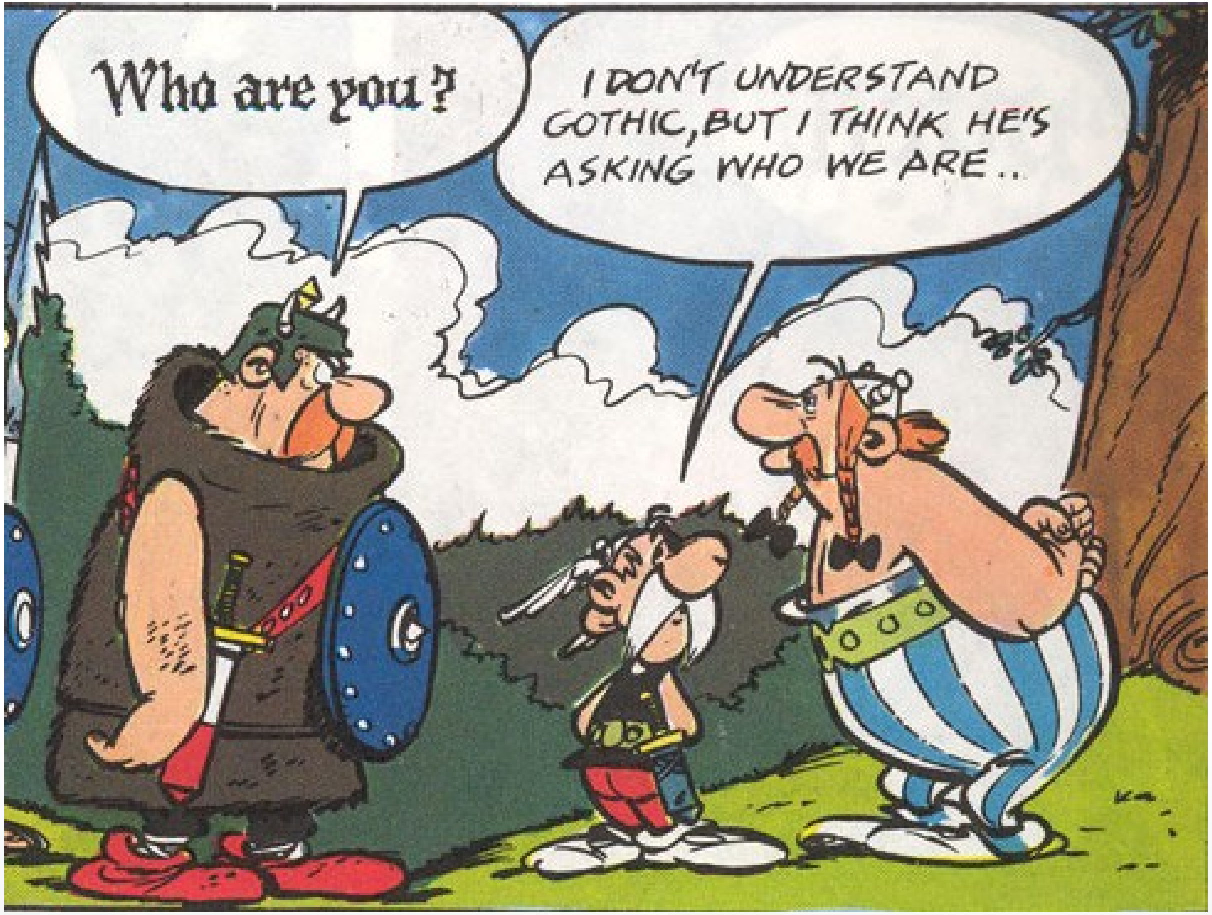
Paulus suſtinet me *(Paolo holds me up)*
Paulus ſus tinet me *(Paolo the pig holds me)*



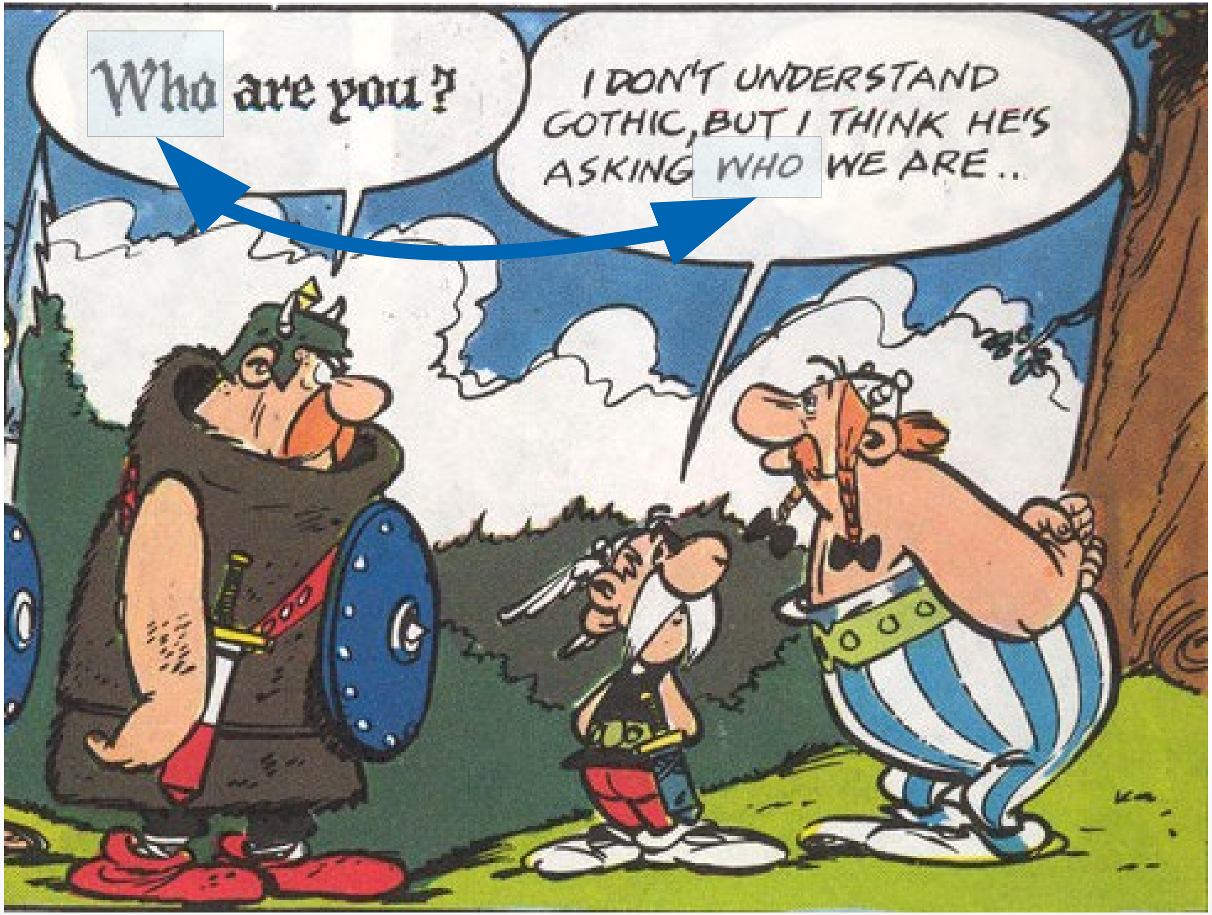
Positional allograph

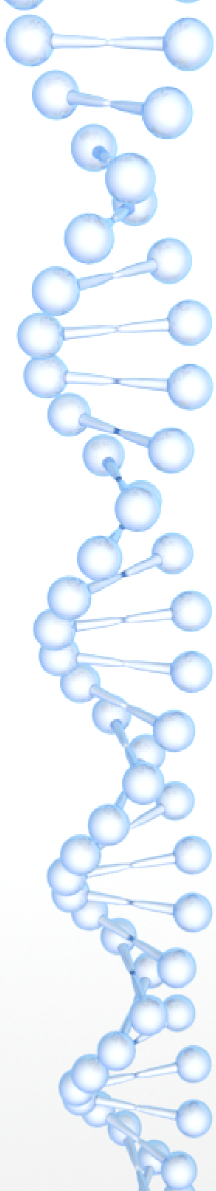
Space

Connotators

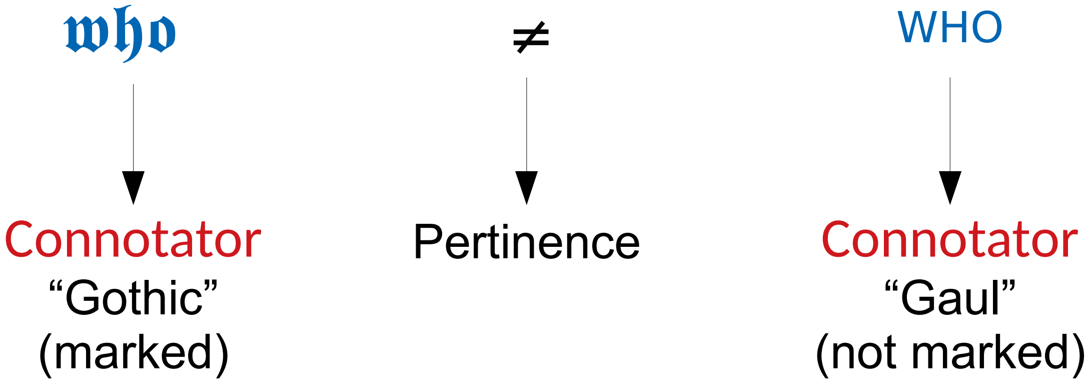

Who are you?
I DON'T UNDERSTAND GOTHIC, BUT I THINK HE'S ASKING WHO WE ARE..

Connotators


Who are you?
I DON'T UNDERSTAND GOTHIC, BUT I THINK HE'S ASKING WHO WE ARE..



Connotators



who

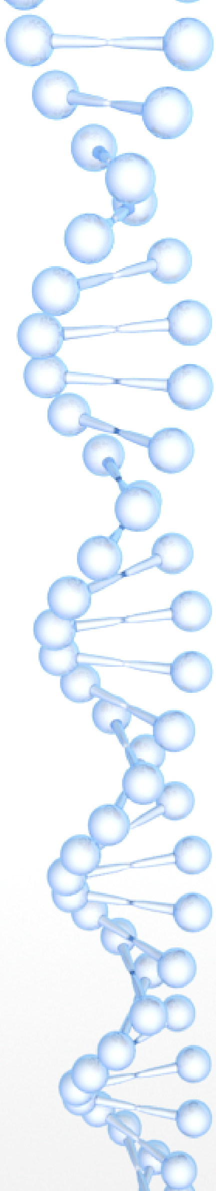
≠

WHO

Connotator
“Gothic”
(marked)

Pertinence

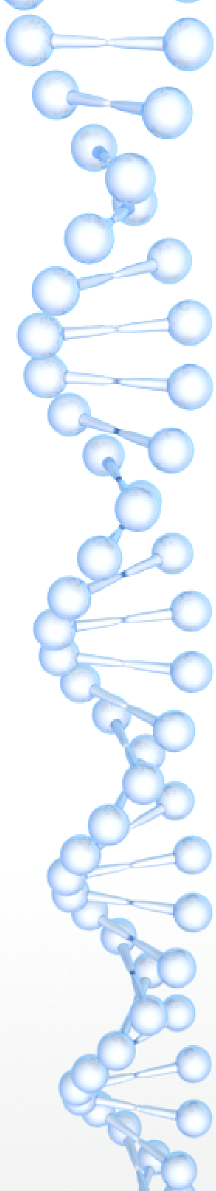
Connotator
“Gaul”
(not marked)



Connotators

Connotators, pertinent for the writer

• *graphemes* as entities	*Emphasis*
• the Evangelist wrote	*Respect*

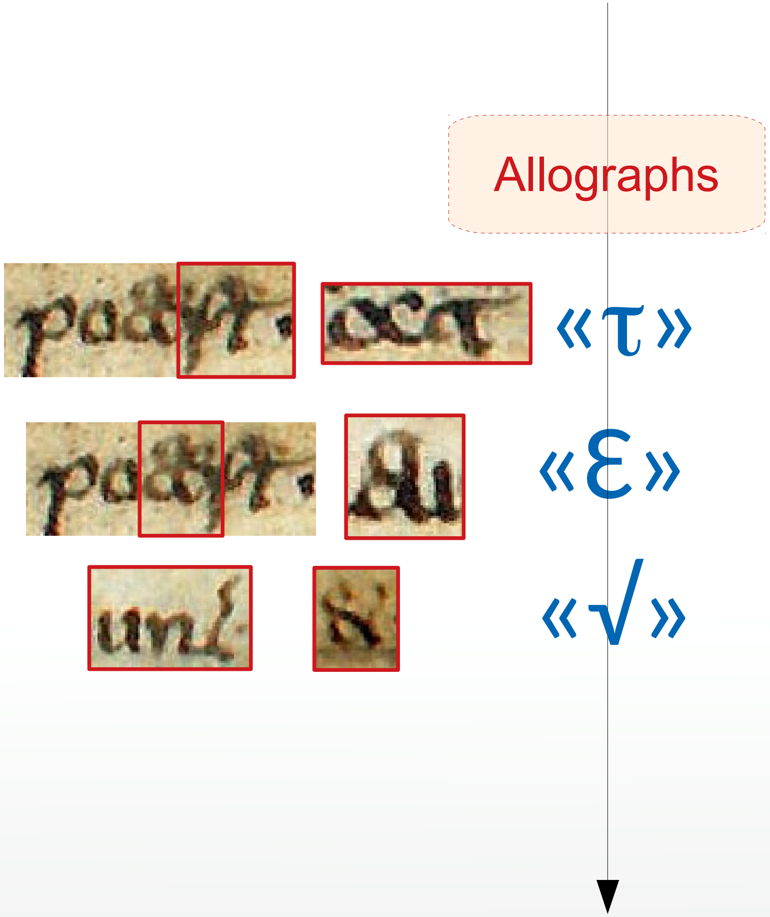


Distinctive value (pertinence) of allographs?

- **Pertinent** differences define entities (graphemes, allographs)
  - Distinctive value
  - -etic *vs* -emic

(Non-)pertinent allographs: positional variants

- Complementary distribution
  - Hjelmslev’s “varieties”



Allographs

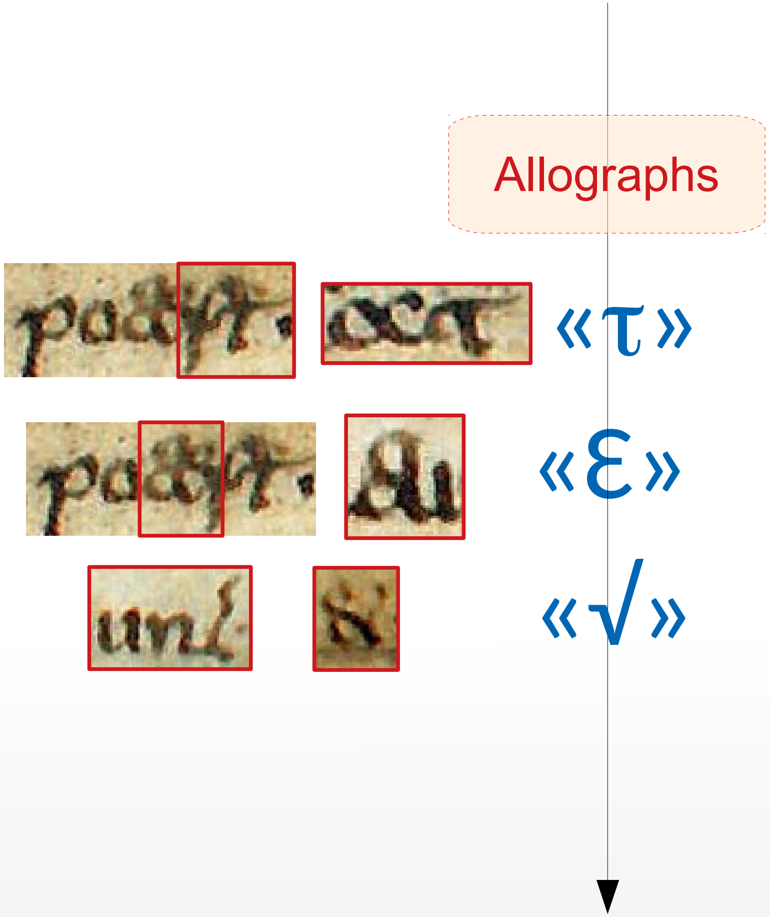
«τ»

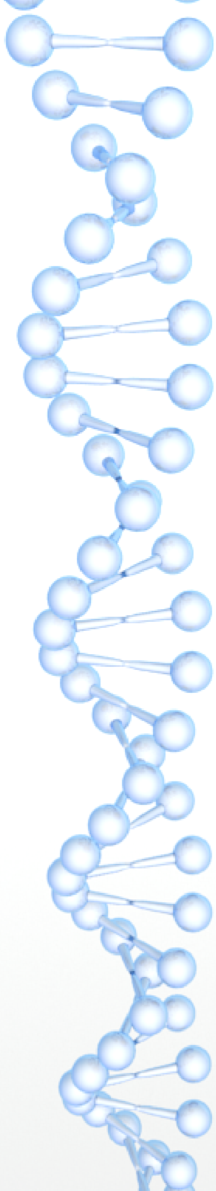
«ε»

«√»

(Non-)pertinent allographs: positional variants

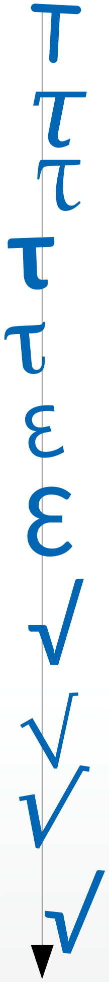
- Ligatures
- **Non-pertinent** for the writer
- Connotators, **pertinent** for (some) readers
  - editors, paleographers, codicologists, historians studying a MS / book
  - (Beneventan vs Caroline script, print font, ſ / s)

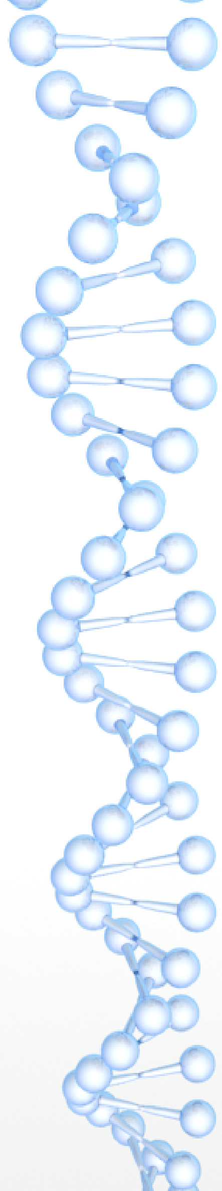




(Non-)pertinent allographs: free variants

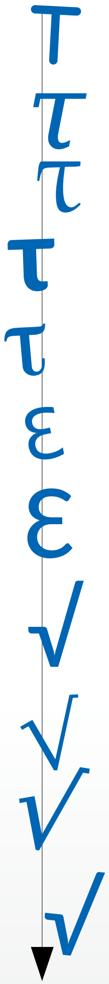
- **Non-pertinent** for the writer
- Connotators, **pertinent** for (some) readers
  - editors, paleographers, codicologists, historians studying a MS / book
  - (Beneventan vs Caroline script, print font, ſ / s)

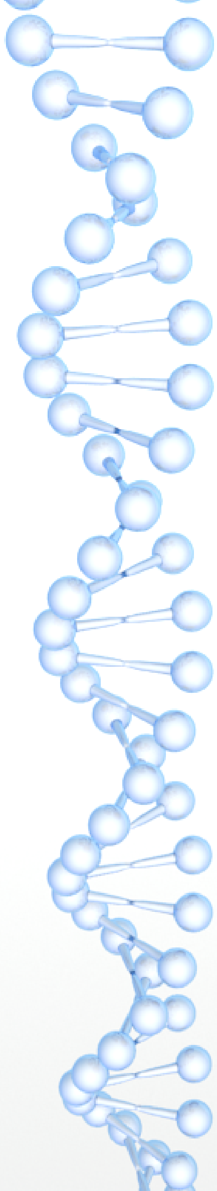




(Non-)pertinent allographs: free variants

- Infinite
- Continuum → discrete
  - It is difficult to draw boundaries
  - *Digital* (=discrete) modelling
- Hjelmlev: metasemiology

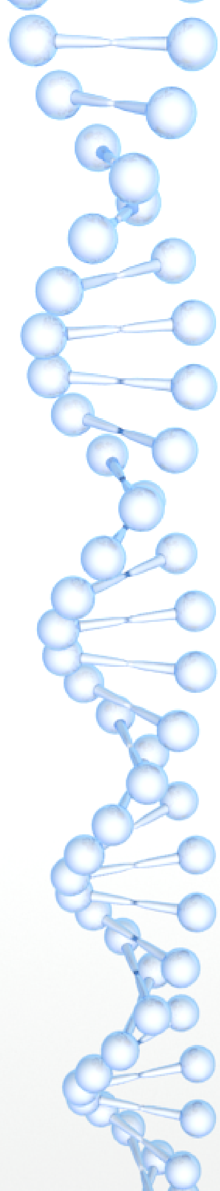




Distinctive value (pertinence) of allographs?

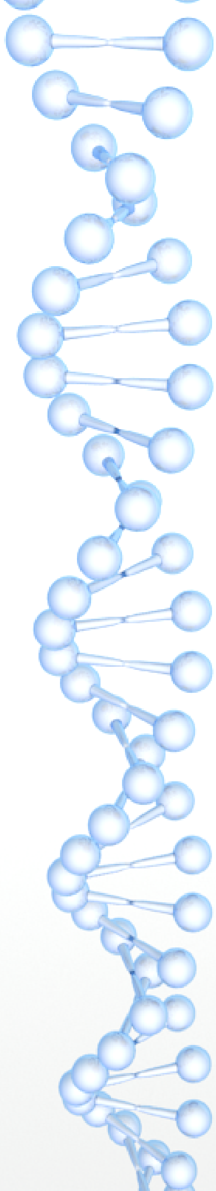
- **Graphemes** change **denotative** meaning
  - fame *vs* name
  - Hjelmslev: denotative semiotics
- **Allographs** can have **other forms of distinctive value** (pertinence)
  - For the writer
    - who *vs* WHO
    - Hjelmslev: connotative semiotics
  - For the reader (digital editor)
    - Digital editors can set their own pertinence (transcription) criteria
      - based on their scientific interests
      - E.g.: fraktur font → political connotation in WW1

In practice: how can grapheme/allograph modelling make my DSE more interoperable?



In practice: how can grapheme/allograph modelling make my DSE more interoperable?

- How can a structural **digital modelling** of the graphemes/allographs distinction make my DSE more **interoperable**?



In practice: how can grapheme/allograph modelling make my DSE more interoperable?

OCR/HTT
(witness A)

Manual (selective)
transcription
(witness B)

In practice: how can grapheme/allograph modelling make my DSE more interoperable?

Allographic transcription

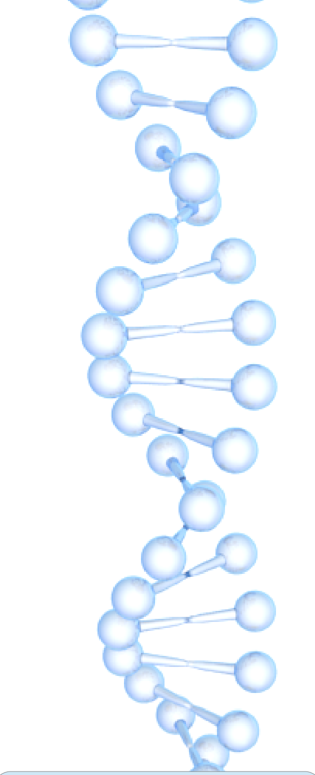
Vnτer <hi>dem</hi> schloss

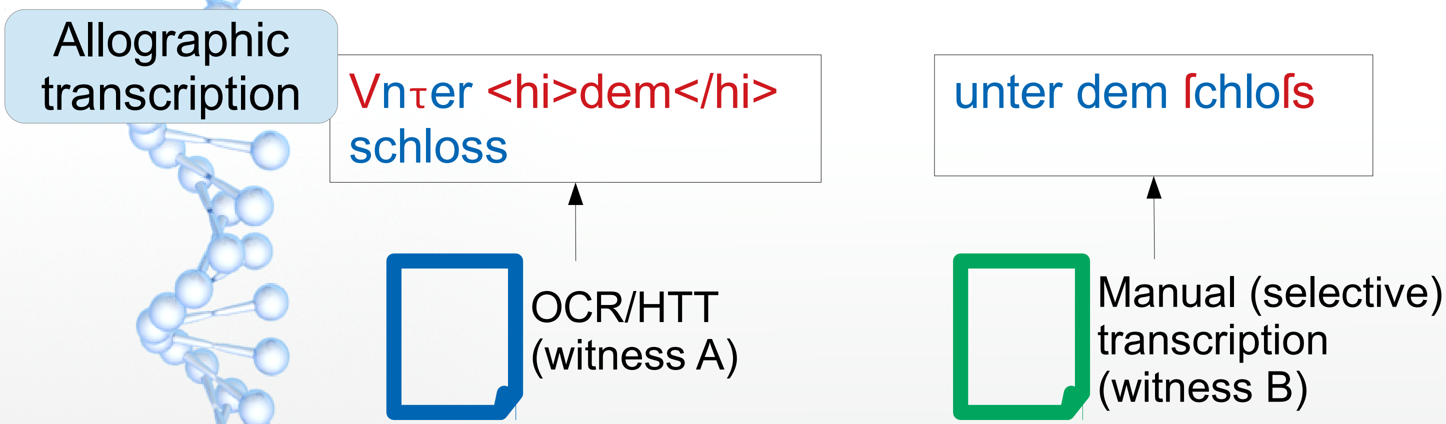
unter dem ſchloſs

OCR/HTT (witness A)

Manual (selective) transcription (witness B)

In practice: how can grapheme/allograph modelling make my DSE more interoperable?





Allographic transcription

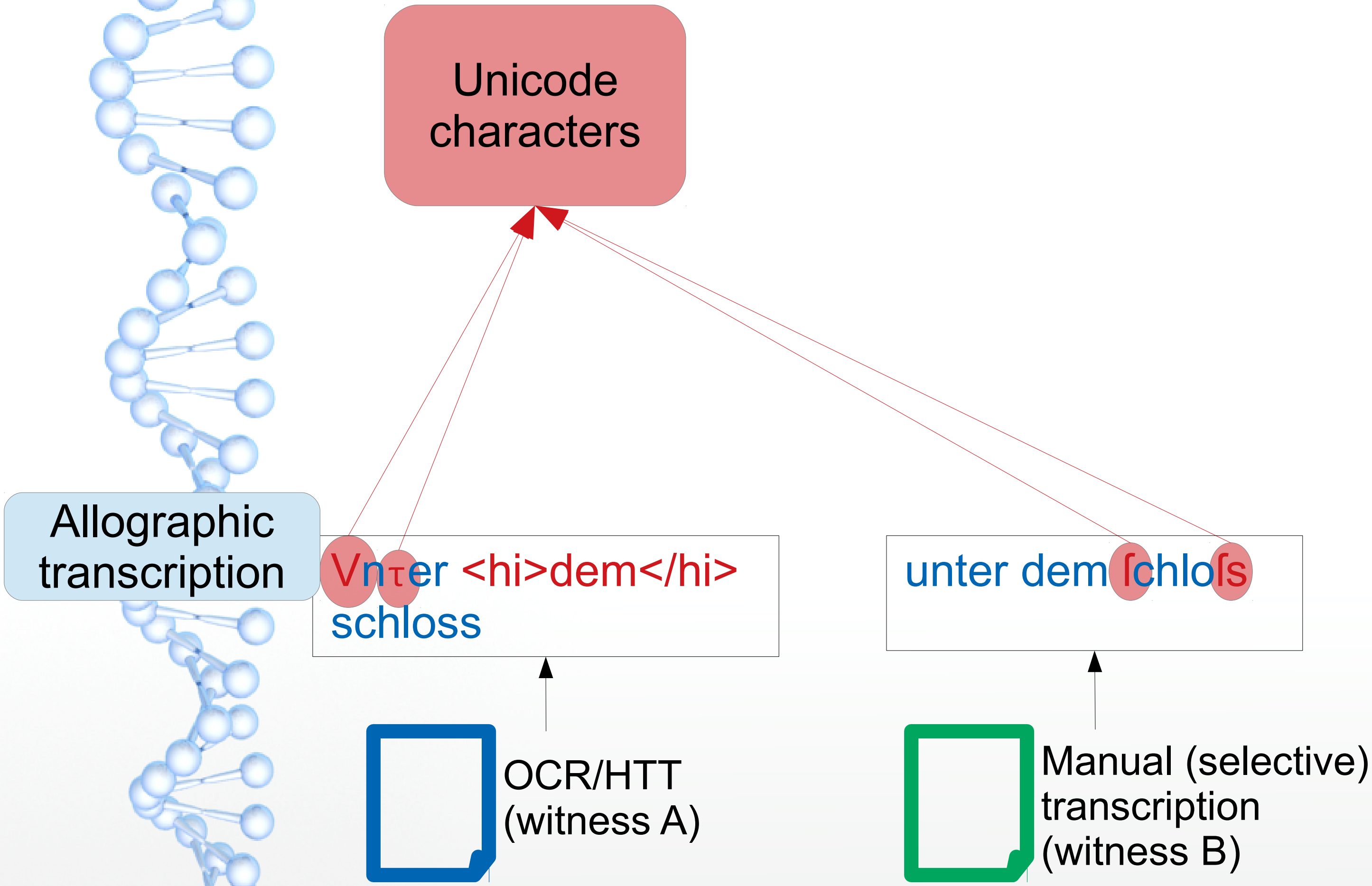
Vnτer <hi>dem</hi> schloss

unter dem ſchloſs

OCR/HTT (witness A)

Manual (selective) transcription (witness B)

In practice: how can grapheme/allograph modelling make my DSE more interoperable?



Unicode characters

Allographic transcription

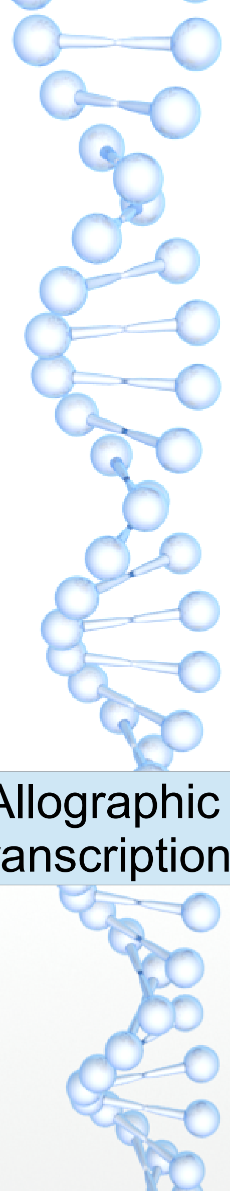
Vnτer <hi>dem</hi> schloss

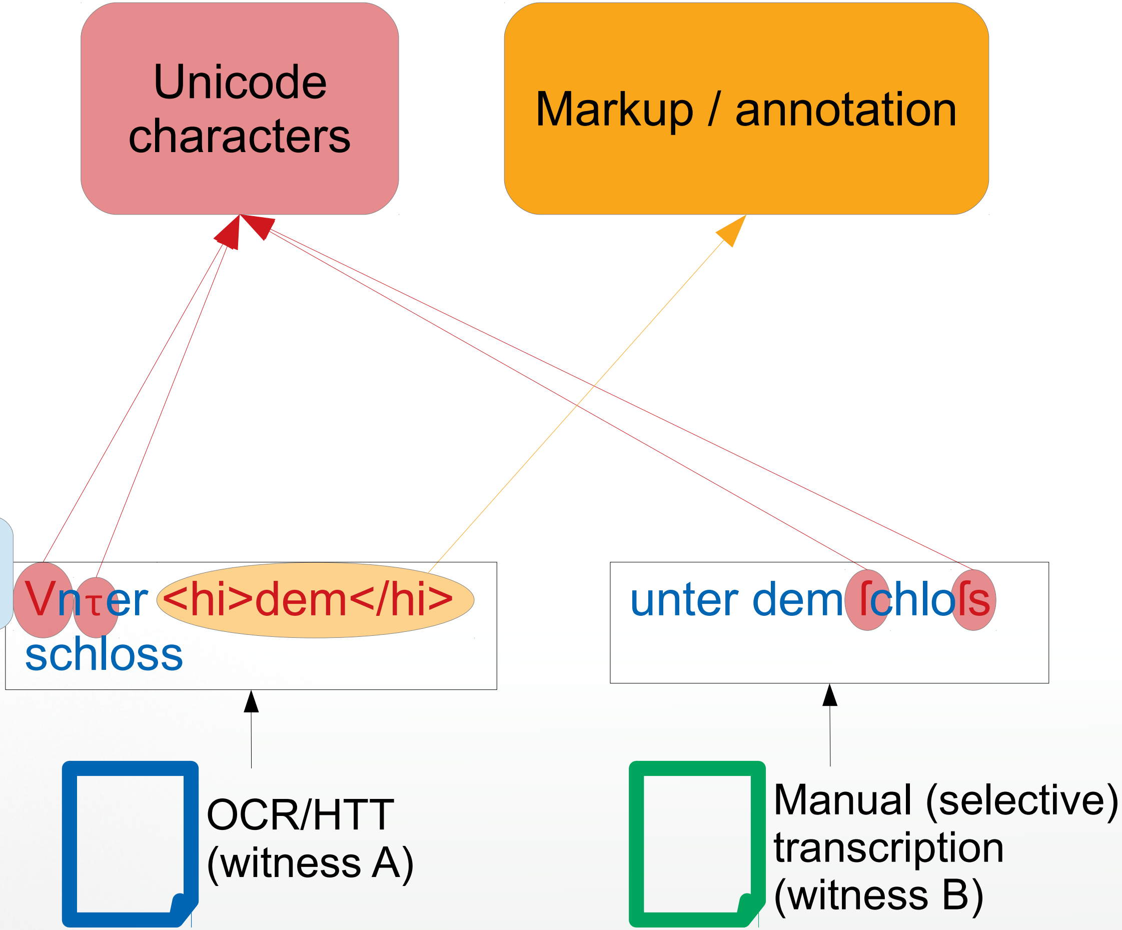
unter dem ſchloſs

OCR/HTT (witness A)

Manual (selective) transcription (witness B)

In practice: how can grapheme/allograph modelling make my DSE more interoperable?





Unicode characters

Markup / annotation

Allographic transcription

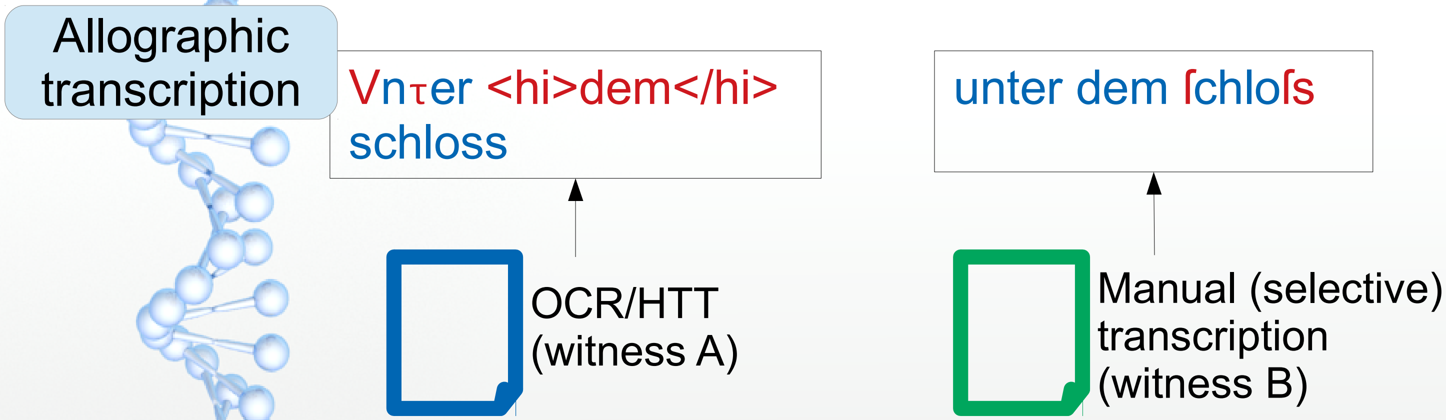
Vnτer <hi>dem</hi> schloss

unter dem ſchloſs

OCR/HTT (witness A)

Manual (selective) transcription (witness B)

In practice: how can grapheme/allograph modelling make my DSE more interoperable?



Allographic transcription

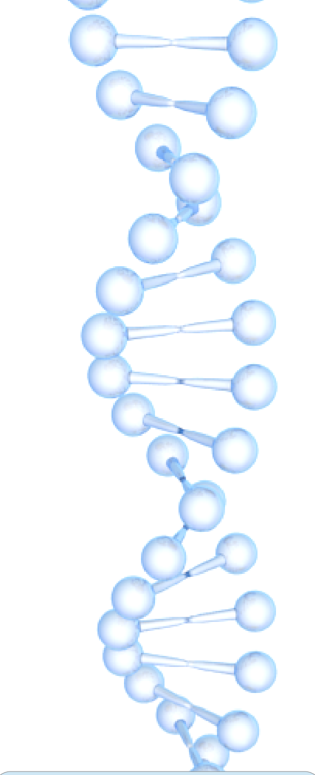
Vnτer <hi>dem</hi> schloss

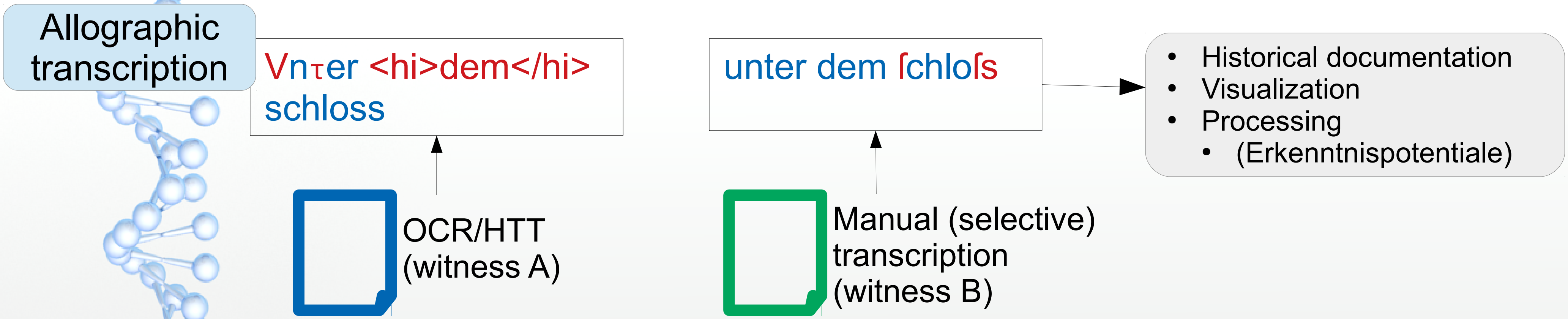
unter dem ſchloſs

OCR/HTT (witness A)

Manual (selective) transcription (witness B)

In practice: how can grapheme/allograph modelling make my DSE more interoperable?





Allographic transcription

Vnτer <hi>dem</hi> schloss

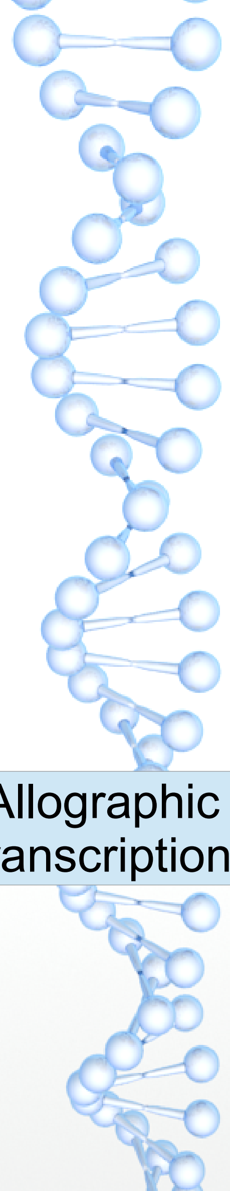
OCR/HTT (witness A)

unter dem ſchloſs

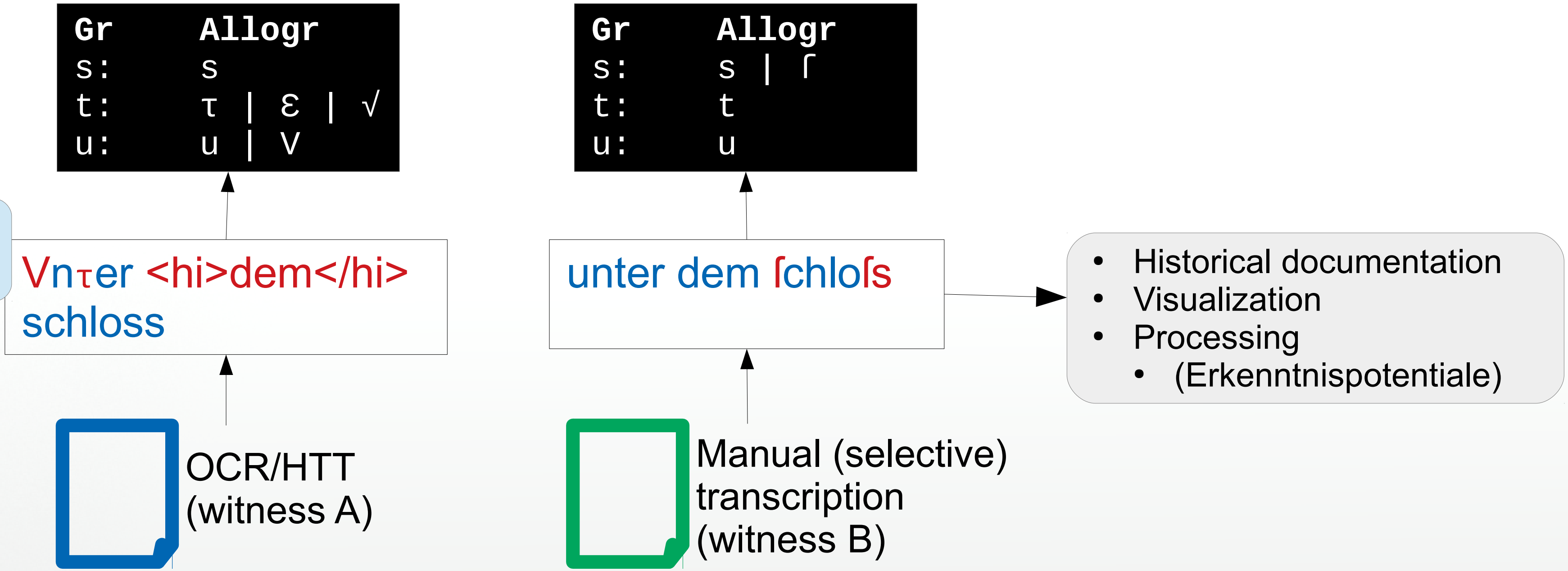
Manual (selective) transcription (witness B)

- Historical documentation
- Visualization
- Processing
  - (Erkenntnispotentiale)

In practice: how can grapheme/allograph modelling make my DSE more interoperable?



Allographic transcription



Gr	Allogr
s:	s
t:	τ | ε | √
u:	u | V

Vnτer <hi>dem</hi> schloss

OCR/HTT (witness A)

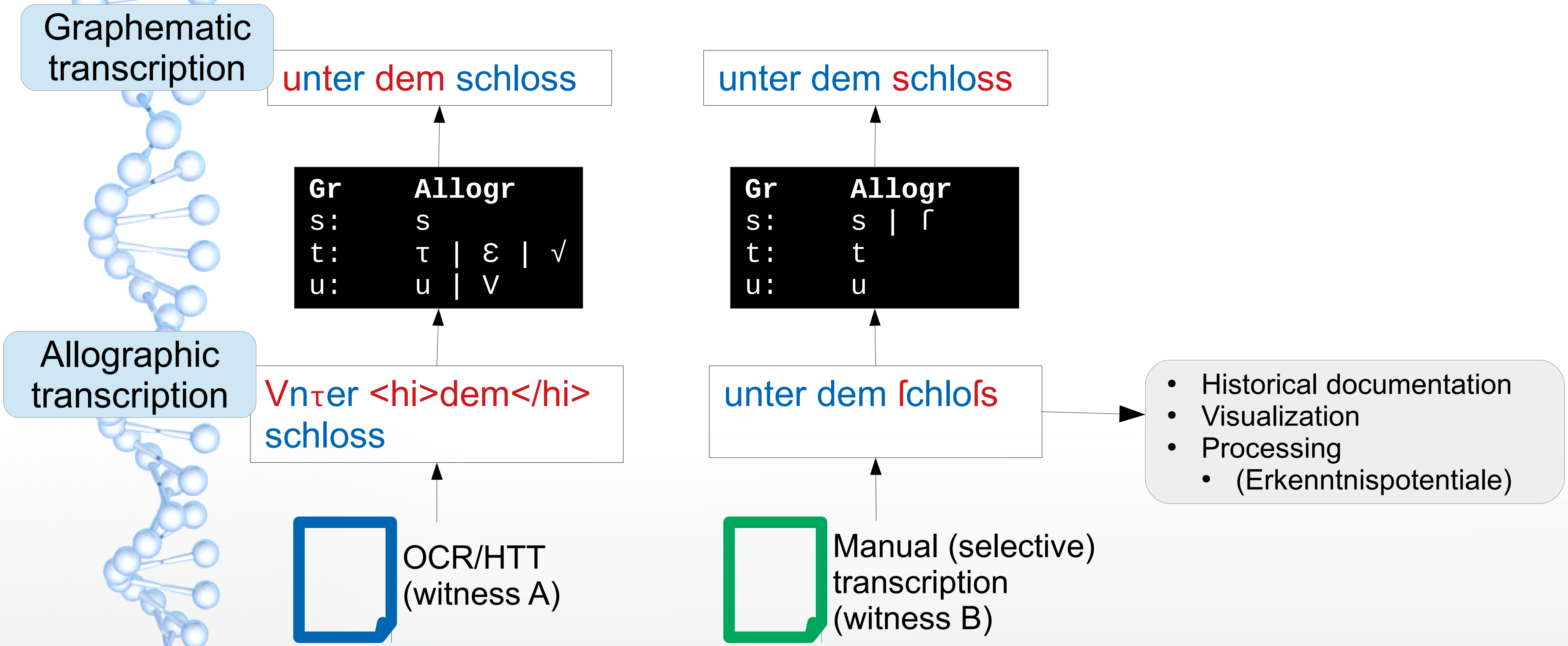
Gr	Allogr
s:	s | ſ
t:	t
u:	u

unter dem ſchloſs

Manual (selective) transcription (witness B)

- Historical documentation
- Visualization
- Processing
  - (Erkenntnispotentiale)

In practice: how can grapheme/allograph modelling make my DSE more interoperable?



Graphematic transcription

unter dem schloss

unter dem schloss

Gr	Allogr
s:	s
t:	τ | Ɛ | √
u:	u | V

Gr	Allogr
s:	s | ſ
t:	t
u:	u

Allographic transcription

Vnτer <hi>dem</hi> schloss

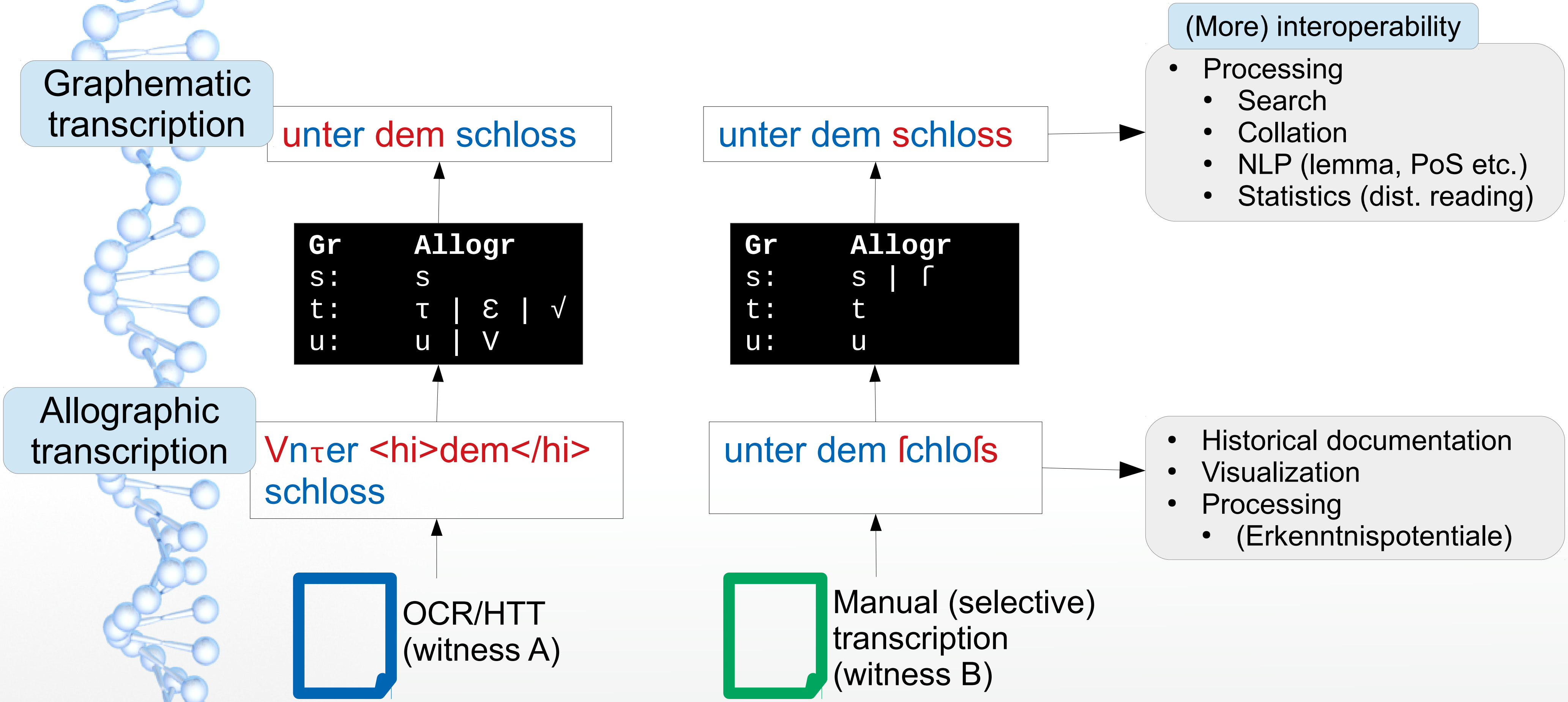
unter dem ſchloſs

- Historical documentation
- Visualization
- Processing
  - (Erkenntnispotentiale)

OCR/HTT (witness A)

Manual (selective) transcription (witness B)

In practice: how can grapheme/allograph modelling make my DSE more interoperable?



Graphematic transcription

unter dem schloss

```
Gr      Allogr
s:      s
t:      τ | ε | √
u:      u | V
```

Allographic transcription

Vnτer <hi>dem</hi> schloss

OCR/HTT (witness A)

unter dem schloss

```
Gr      Allogr
s:      s | ſ
t:      t
u:      u
```

unter dem ſchloſs

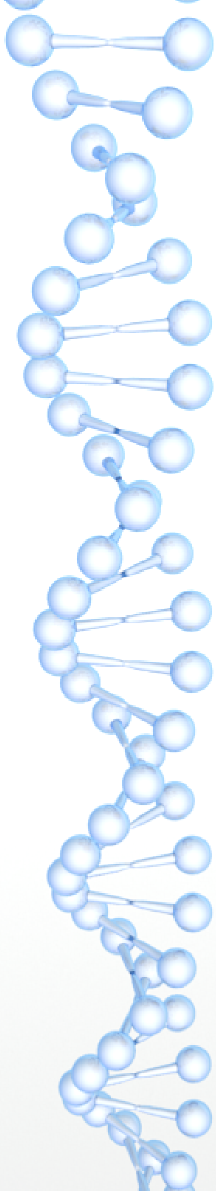
Manual (selective) transcription (witness B)

(More) interoperability

- Processing
  - Search
  - Collation
  - NLP (lemma, PoS etc.)
  - Statistics (dist. reading)

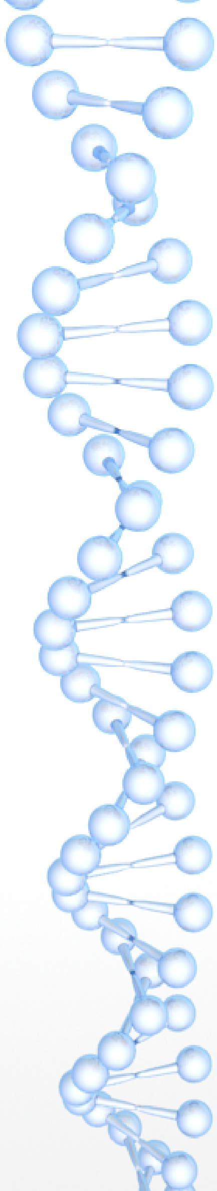
- Historical documentation
- Visualization
- Processing
  - (Erkenntnispotentiale)

Open issues



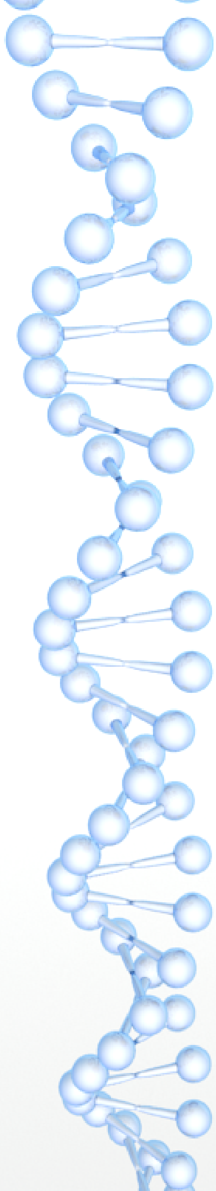
Open issues

- Individual allographs
  - Distinctive value / pertinence (capitals, punctuation etc.)
  - Ligature segmentation one (&) or two (et)?



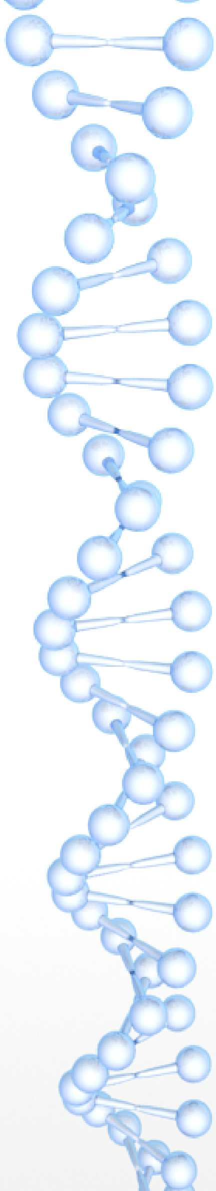
Open issues

- Allographic words
  - Spelling (wife / wyffe)



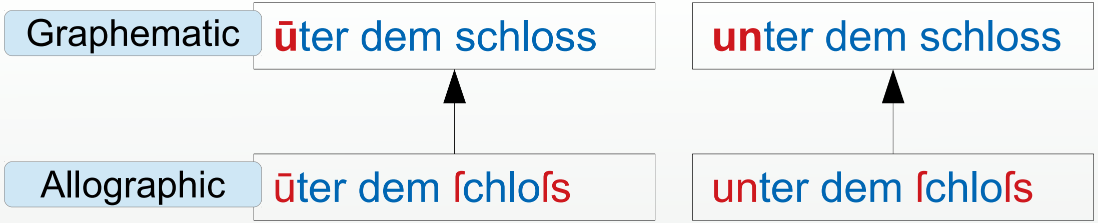
Open issues

- Allographic words
  - Spelling (wife / wyffe)
  - Abbreviations (ūter / unter)



Open issues

- Allographic words
  - Spelling (wife / wyffe)
  - Abbreviations (ūter / unter)



Graphematic

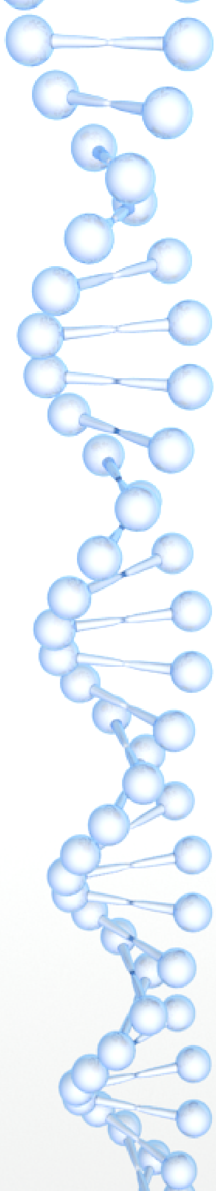
ūter dem schloss

unter dem schloss

Allographic

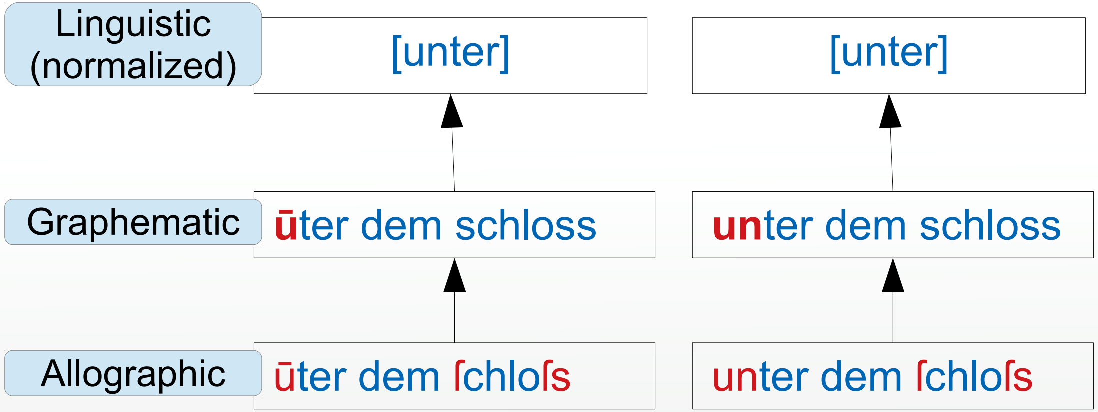
ūter dem ſchloſs

unter dem ſchloſs



Open issues

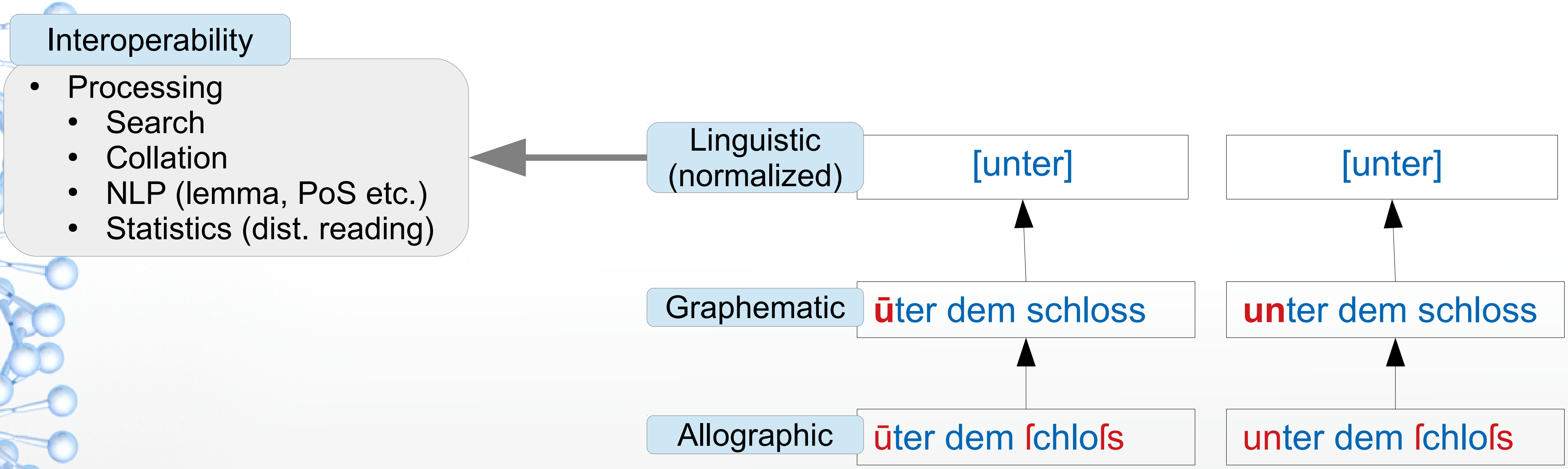
- Allographic words
  - Spelling (wife / wyffe)
  - Abbreviations (ūter / unter)



Linguistic (normalized)	[unter]	[unter]
Graphematic	ūter dem schloss	unter dem schloss
Allographic	ūter dem ſchloſs	unter dem ſchloſs

Open issues

- Allographic words
  - Spelling (wife / wyffe)
  - Abbreviations (ūter / unter)



Interoperability

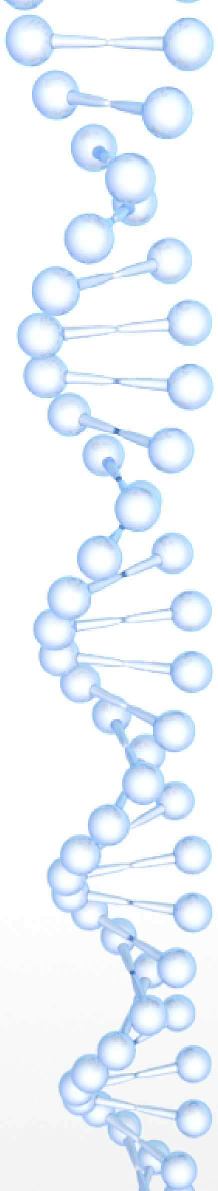
- Processing
  - Search
  - Collation
  - NLP (lemma, PoS etc.)
  - Statistics (dist. reading)

Linguistic (normalized) | [unter] | [unter]

Graphematic | ūter dem schloss | unter dem schloss

Allographic | ūter dem ſchloſs | unter dem ſchloſs

Outline



Outline

- **Interoperability**
  of digital scholarly editions (DSEs)
  based on diplomatic transcriptions
- **Digital modelling (ontology)**
  of pre-modern writing systems
  - **Graphemes / allographs**
  - **Allographs**:
    capitals, ligatures, positional variants, emphasis etc.
- **In practice**:
  how can grapheme/allograph modelling
  make my DSE more interoperable?
- **Open issues**