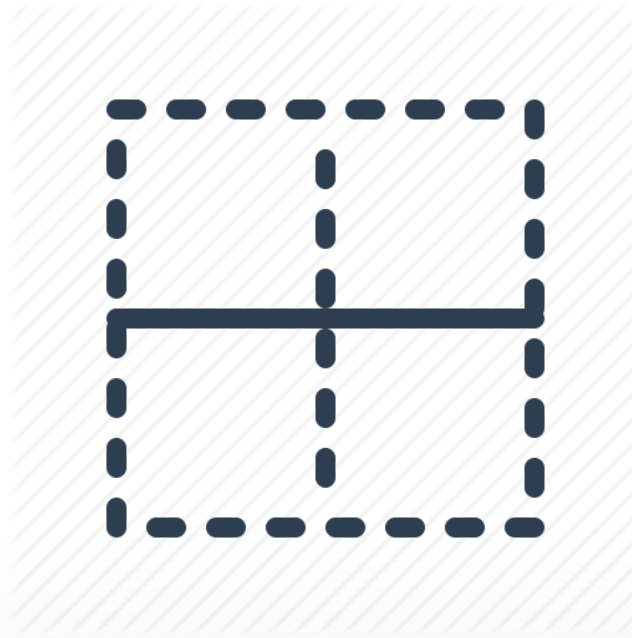


Paolo Monella
paolo.monella@unipa.it

An ontology for digital graphematics and philology



[ve]dph
Venice Centre for
Digital and Public
Humanities

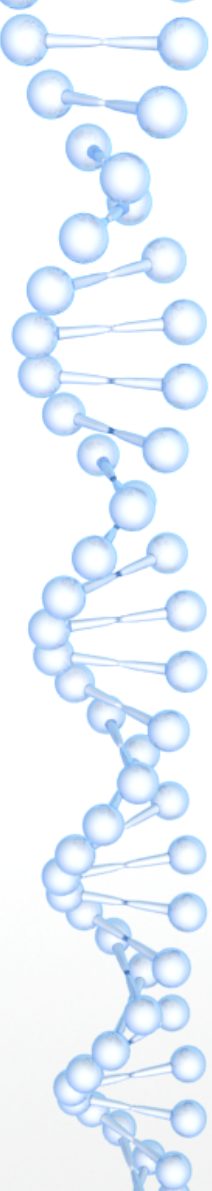


Die (hyper-)diplomatische Transkription und ihre Erkenntnispotentiale
Bergische Universität Wuppertal (BUW), 6 February 2020





Outline



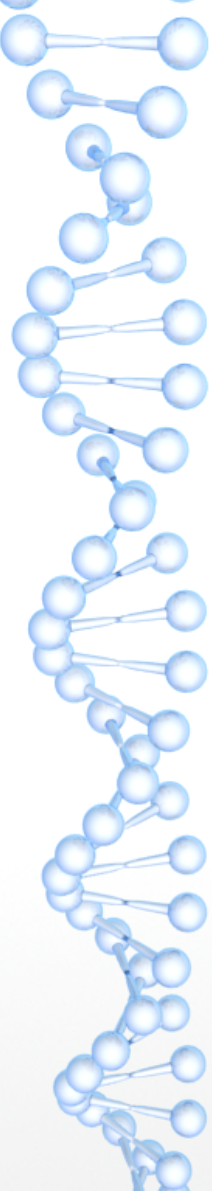
Outline

- **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**

A white speech bubble with a dark blue background. The bubble has a rectangular body and a triangular tail pointing downwards and to the left.

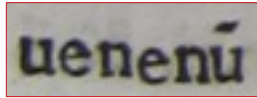
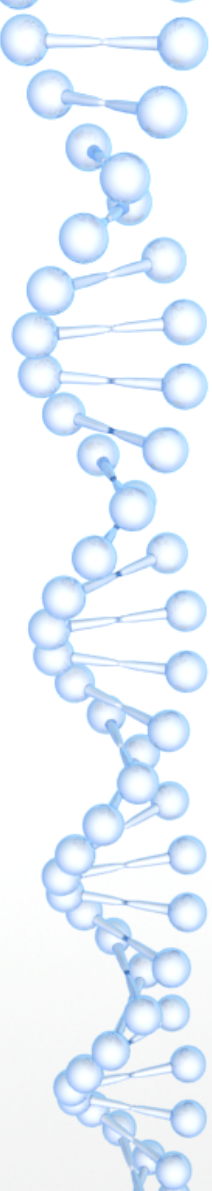
Interoperability: the issue

Interoperability: the issue



uenenú

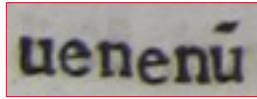
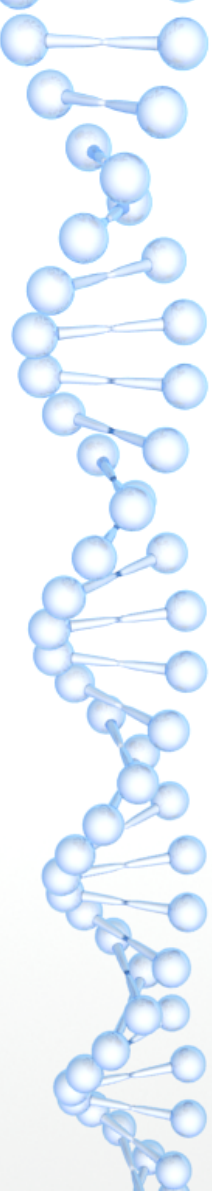
Interoperability: the issue



uenenū

- uenenū

Interoperability: the issue

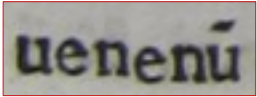
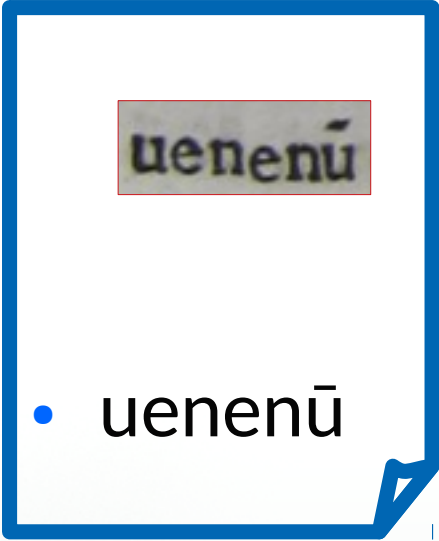
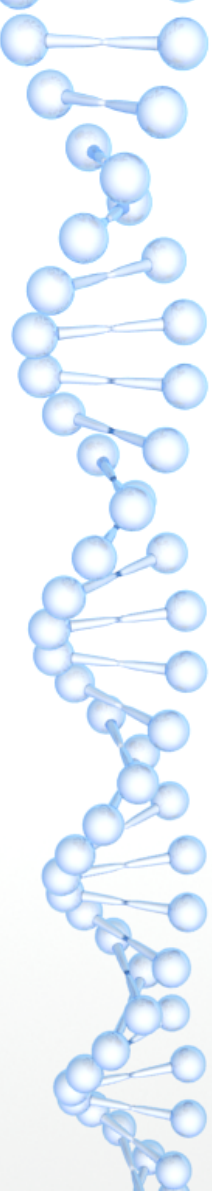


- uenenū

Diplomatic

- Historical documentation
- Visualization
- Processing
 - (Erkenntnispotentiale)

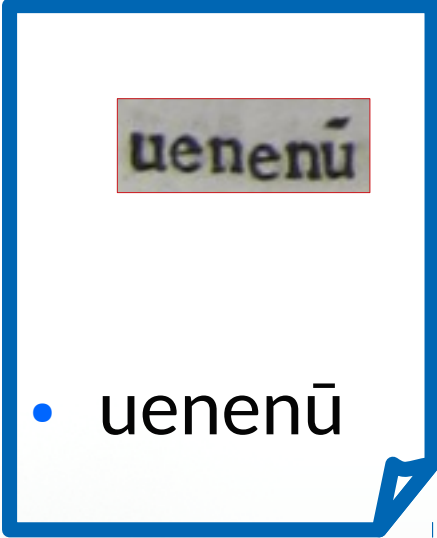
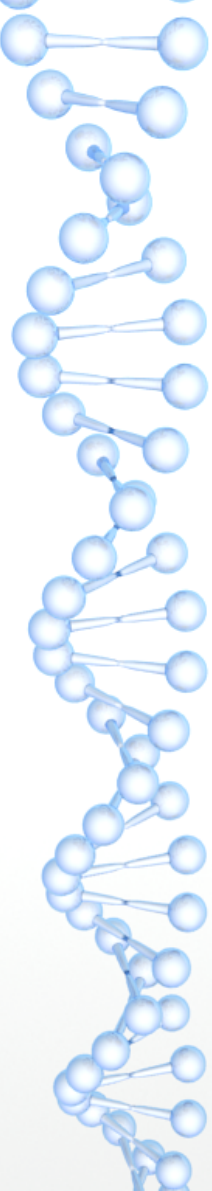
Interoperability: the issue



uenenū

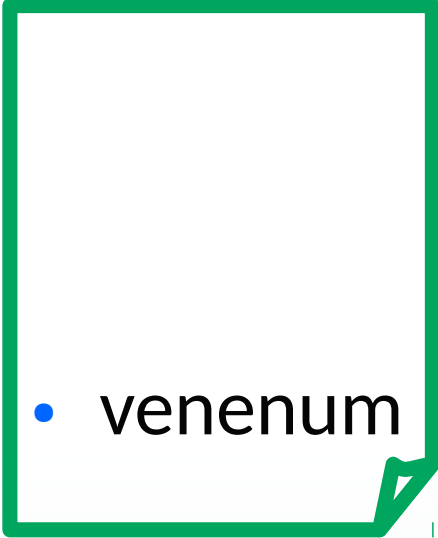
- uenenū

Interoperability: the issue



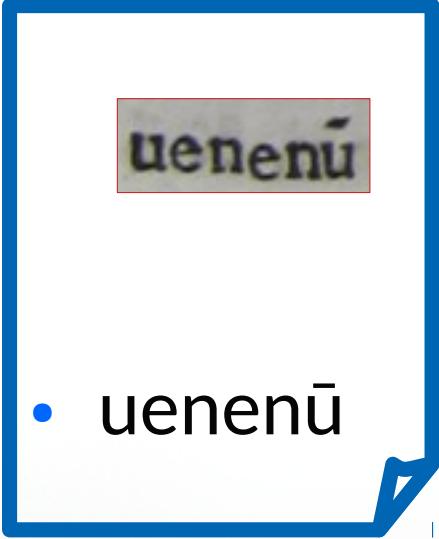
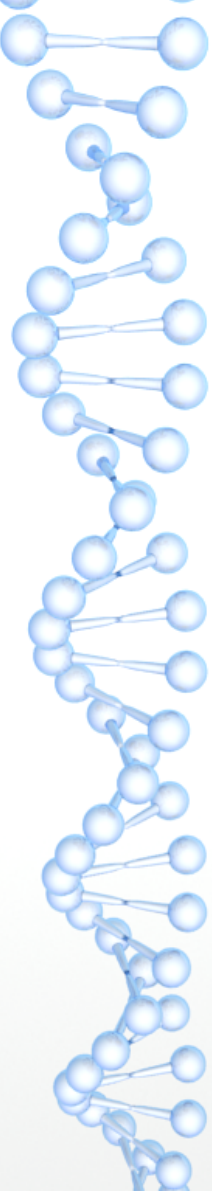
uenenū

- uenenū

- 
- venenum

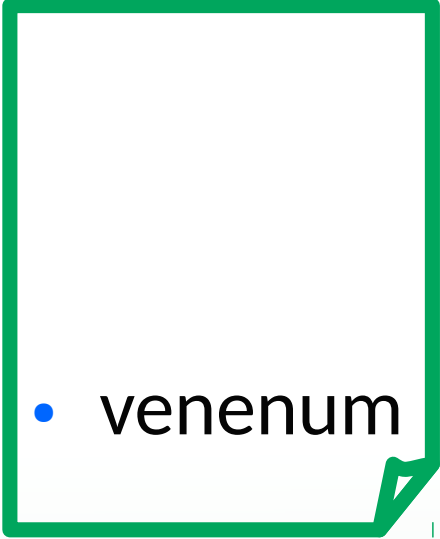
Interoperability: the issue

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

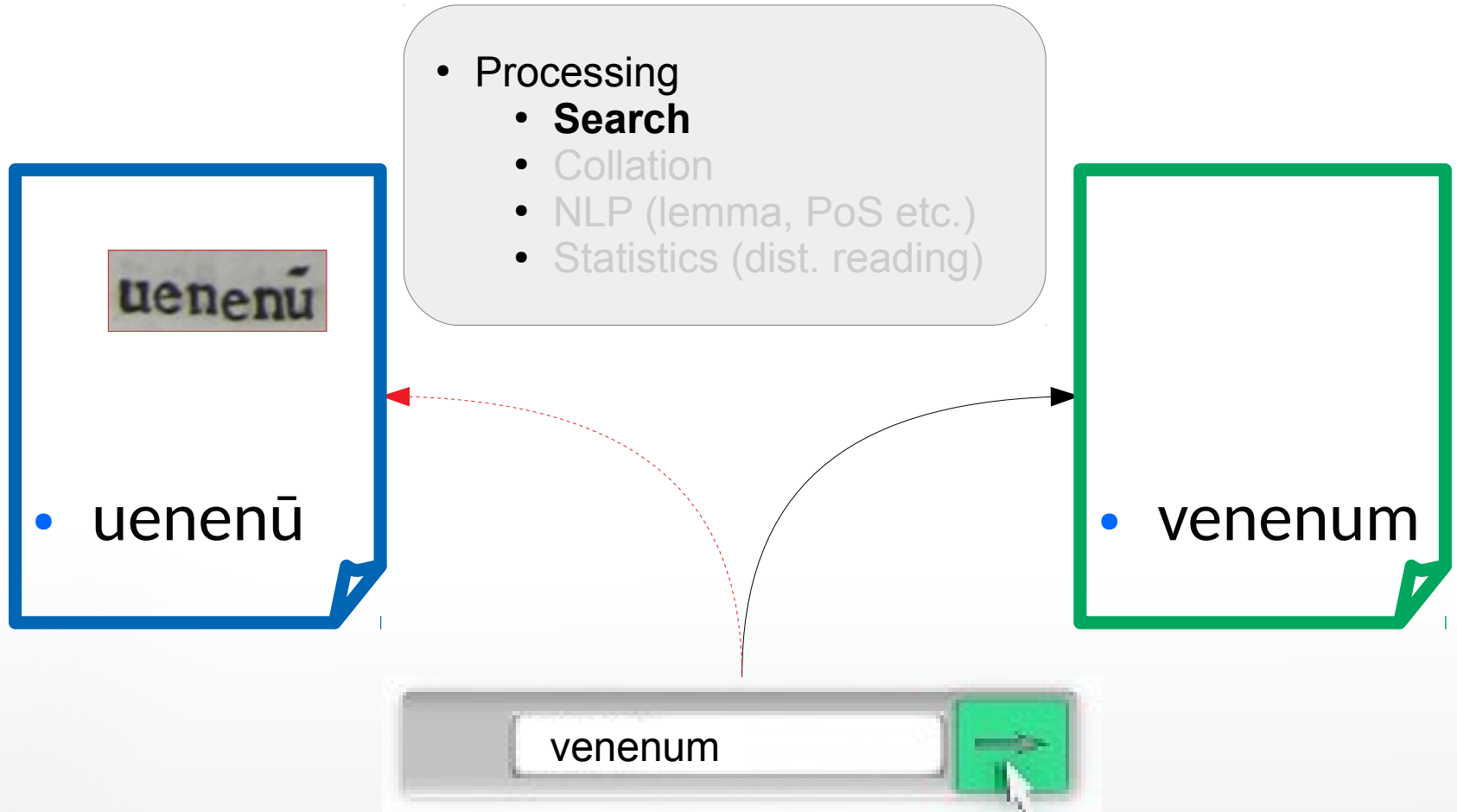


uenenū

- uenenū

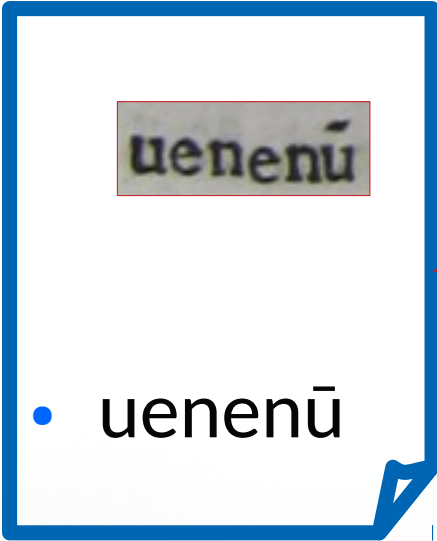
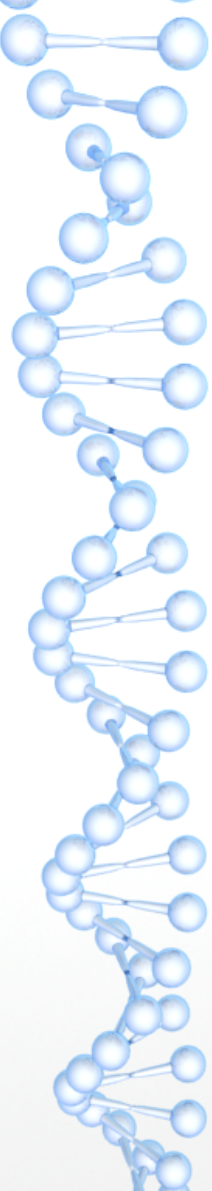
- 
- venenum

Interoperability: the issue



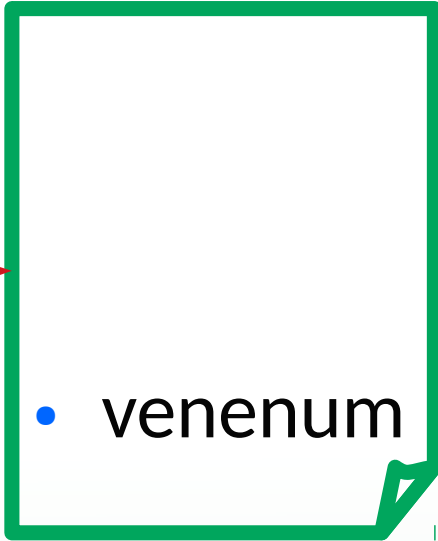
Interoperability: the issue

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



uenenū

• uenenū

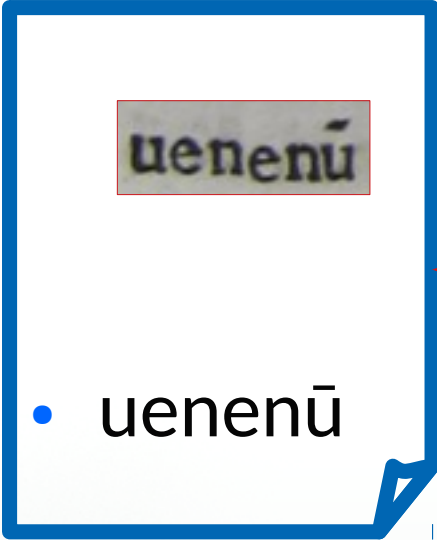
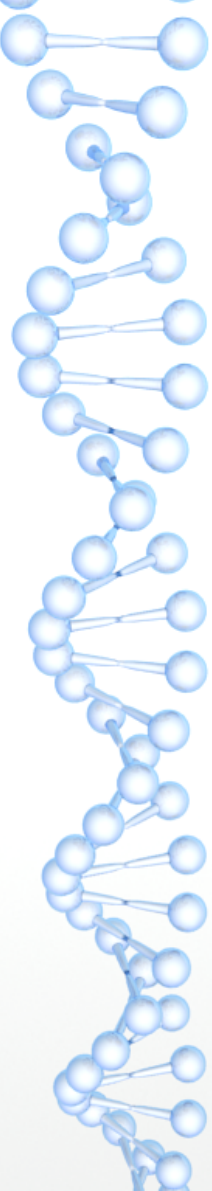


• venenum

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

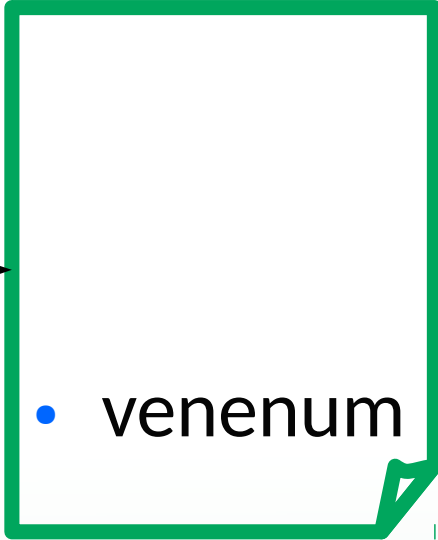
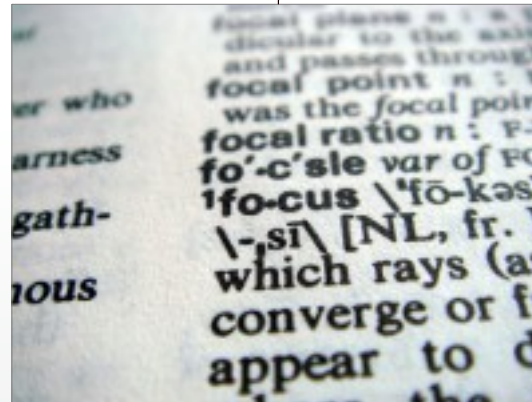
Interoperability: the issue

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



uenenū

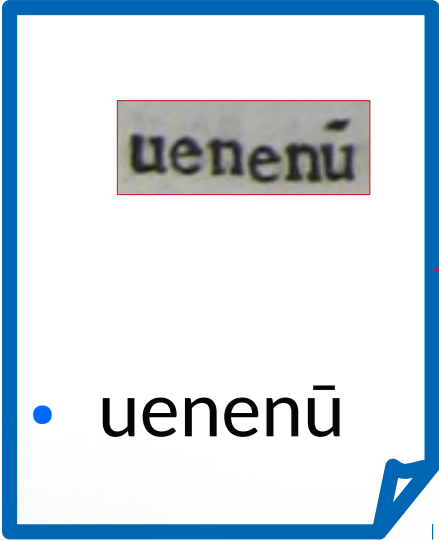
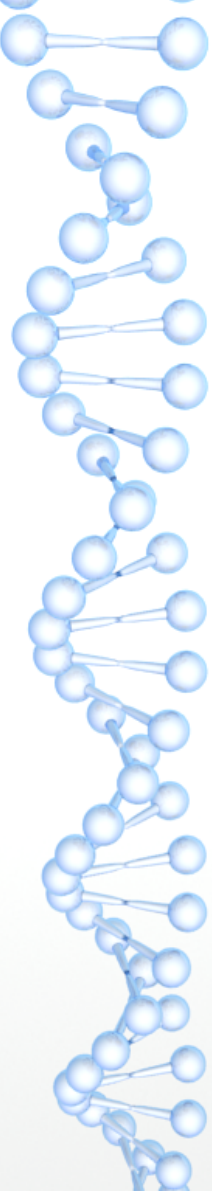
• uenenū



• venenum

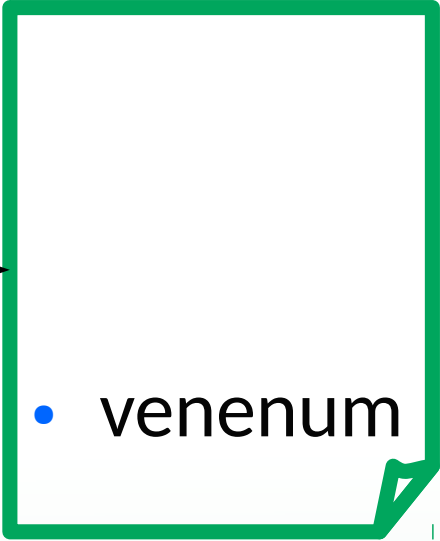
Interoperability: the issue

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

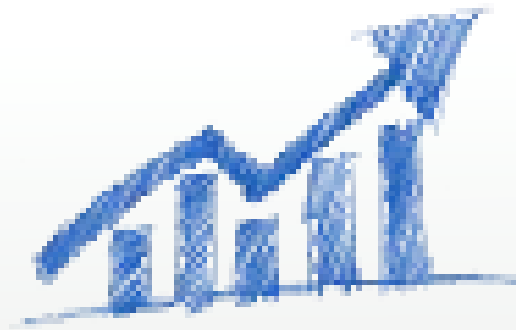


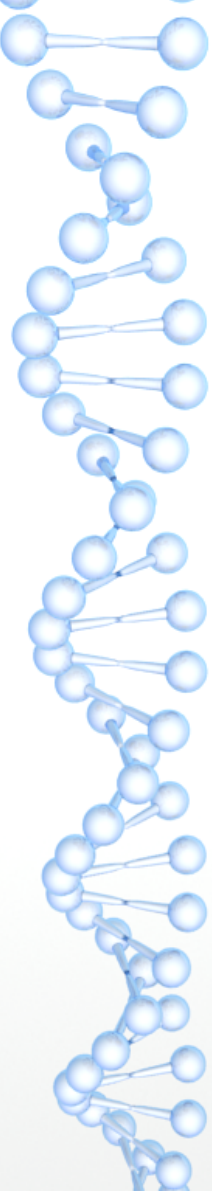
uenenū

• uenenū



• venenum



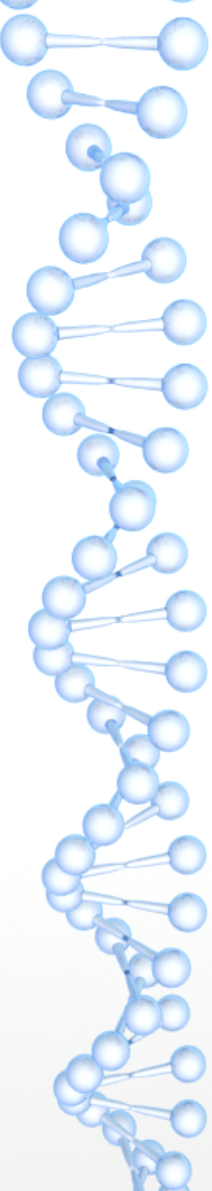


Interoperability: the issue

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

A white speech bubble with a dark blue background. The bubble has a rectangular body and a triangular tail pointing towards the bottom-left corner.

**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>)?

Project-specific solutions

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



A white speech bubble with a dark blue background. The bubble has a rectangular body and a triangular tail pointing towards the bottom-left corner.

**Interoperability
through modelling**

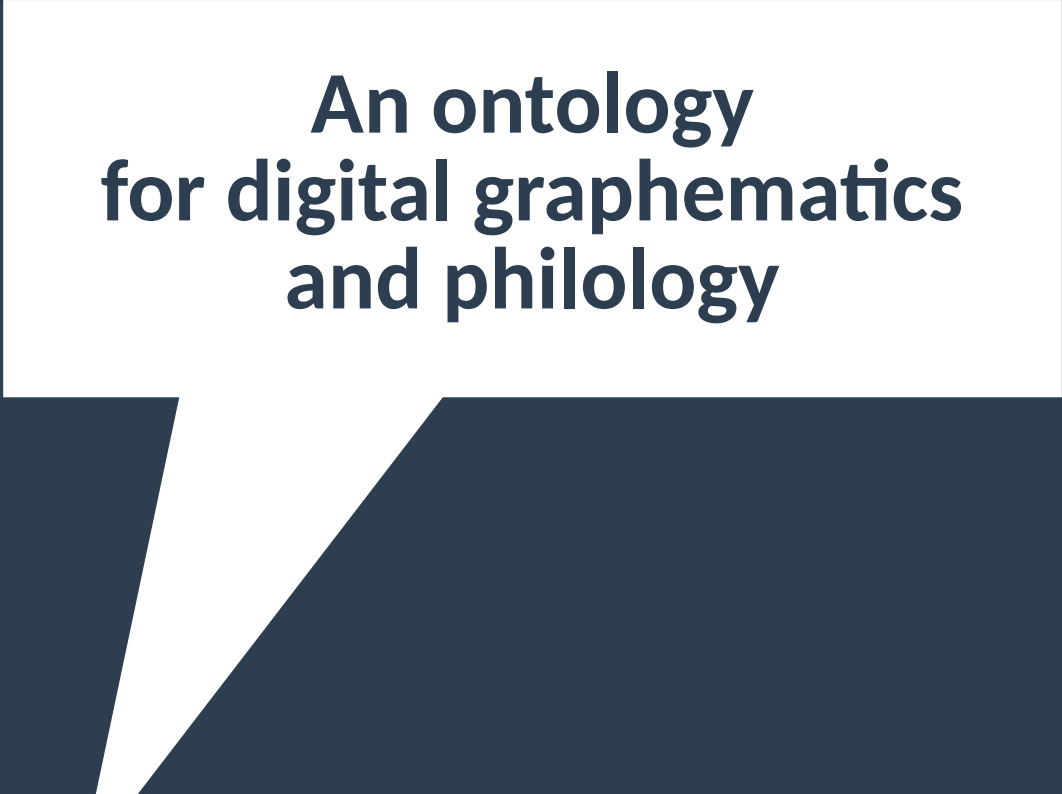
Interoperability through modelling



WANTED

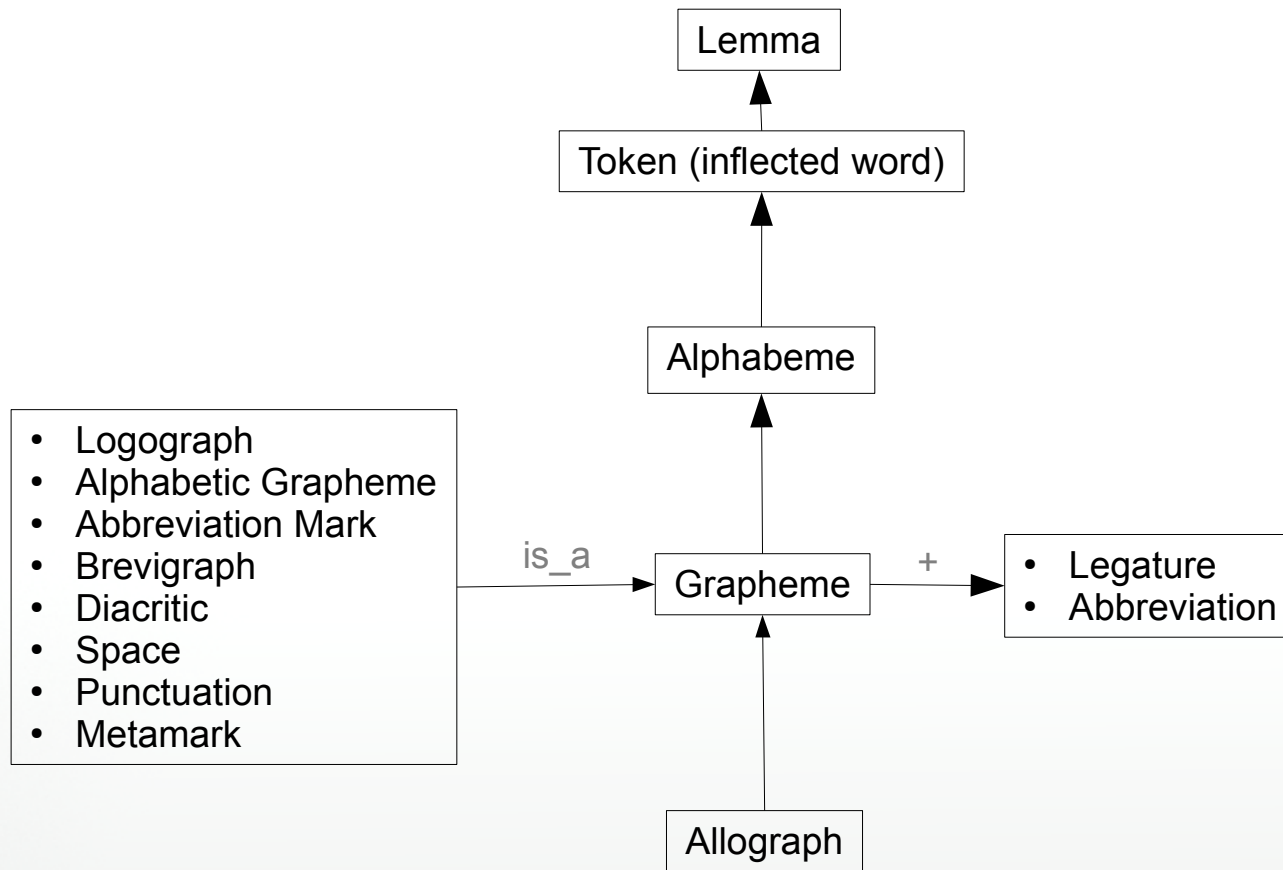
**DEAD OR ALIVE
REWARD
\$4,000**

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

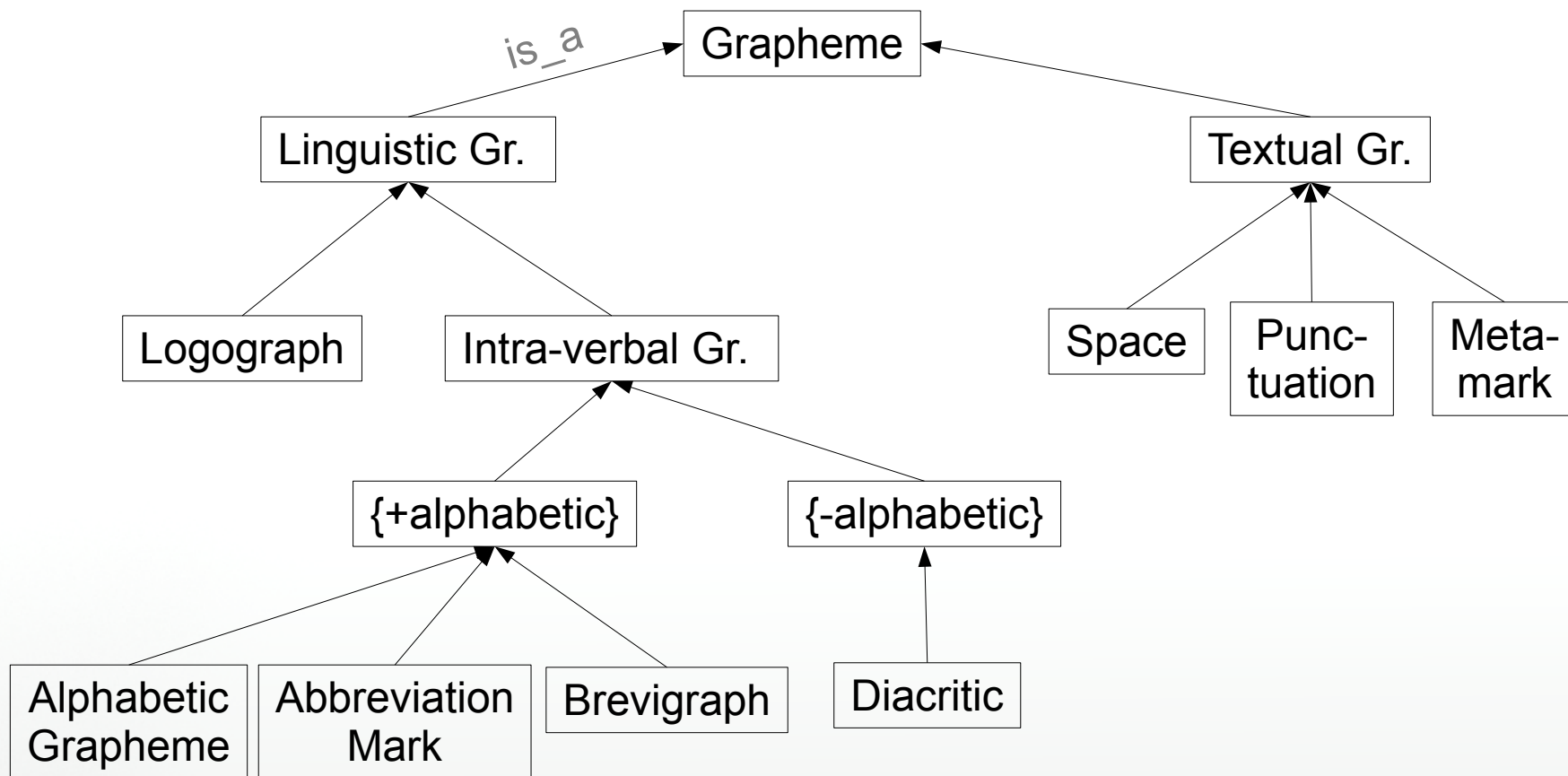


**An ontology
for digital graphematics
and philology**

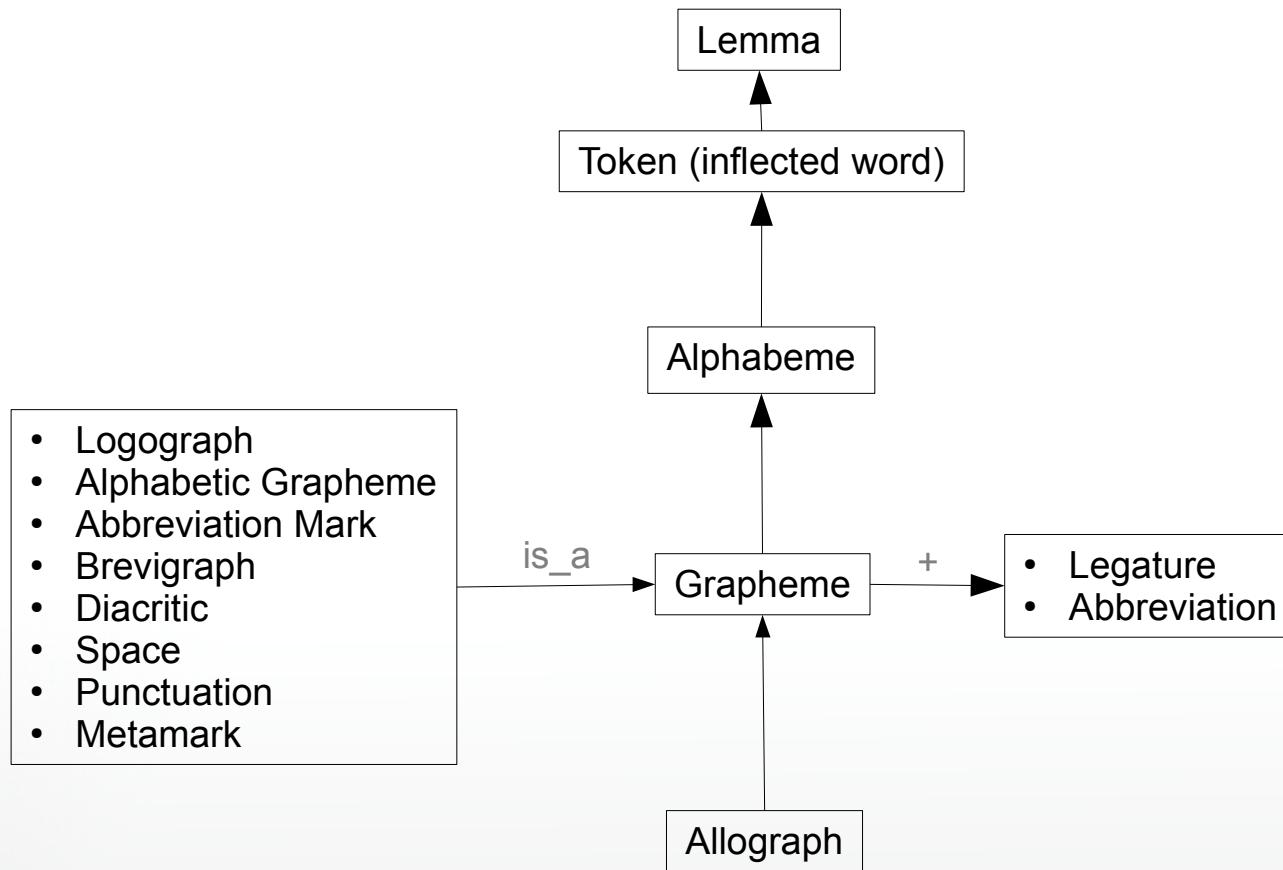
Ontology



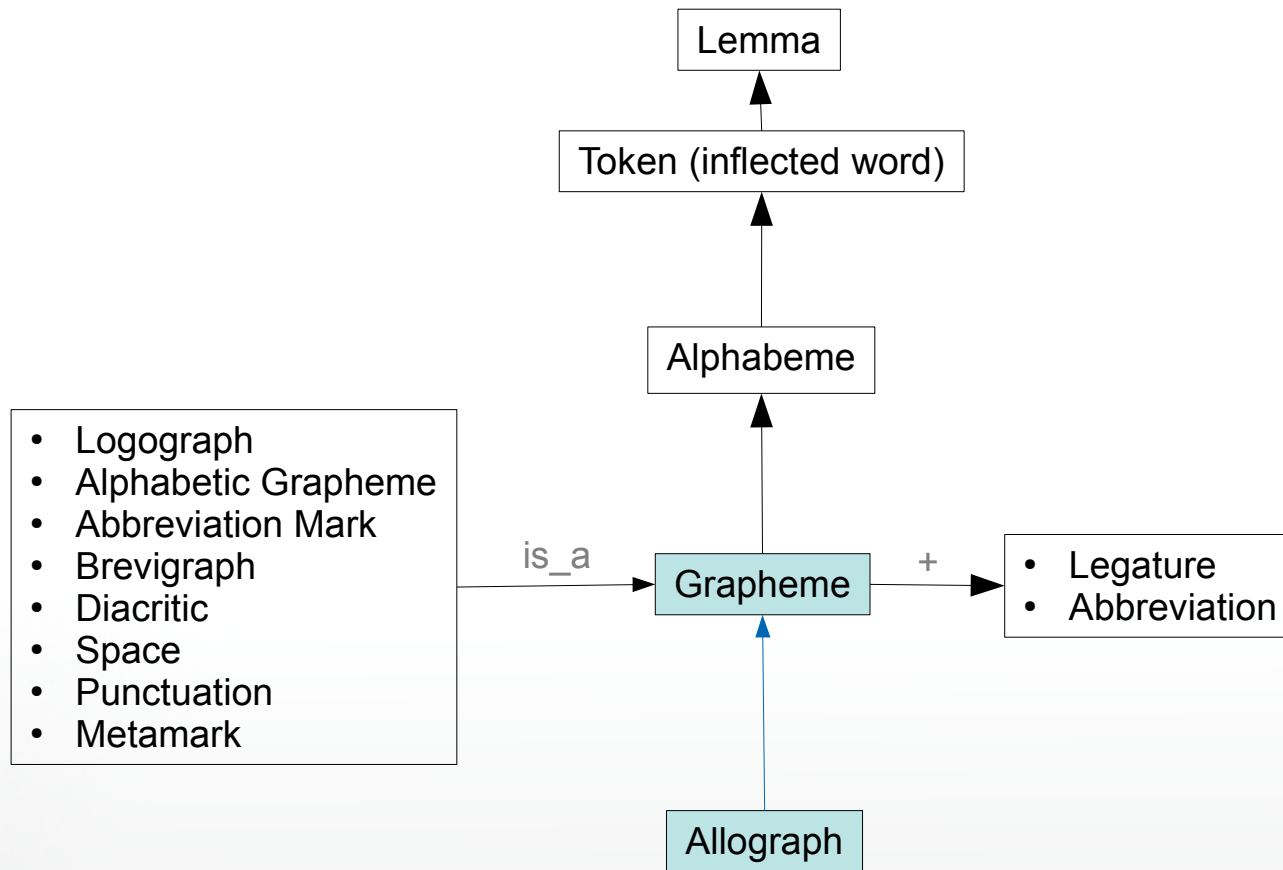
Ontology



Ontology



Ontology



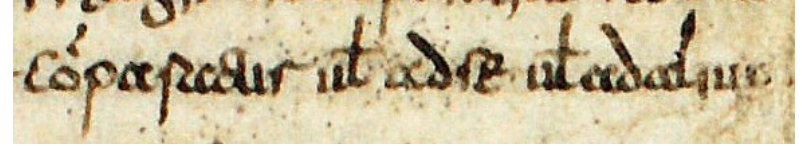


Graphemes/allographs

Graphemes/allographs: the commutation test

System

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



<s>

<t>

Text

- cōparaEUR

u

adfe

uɫadaliu

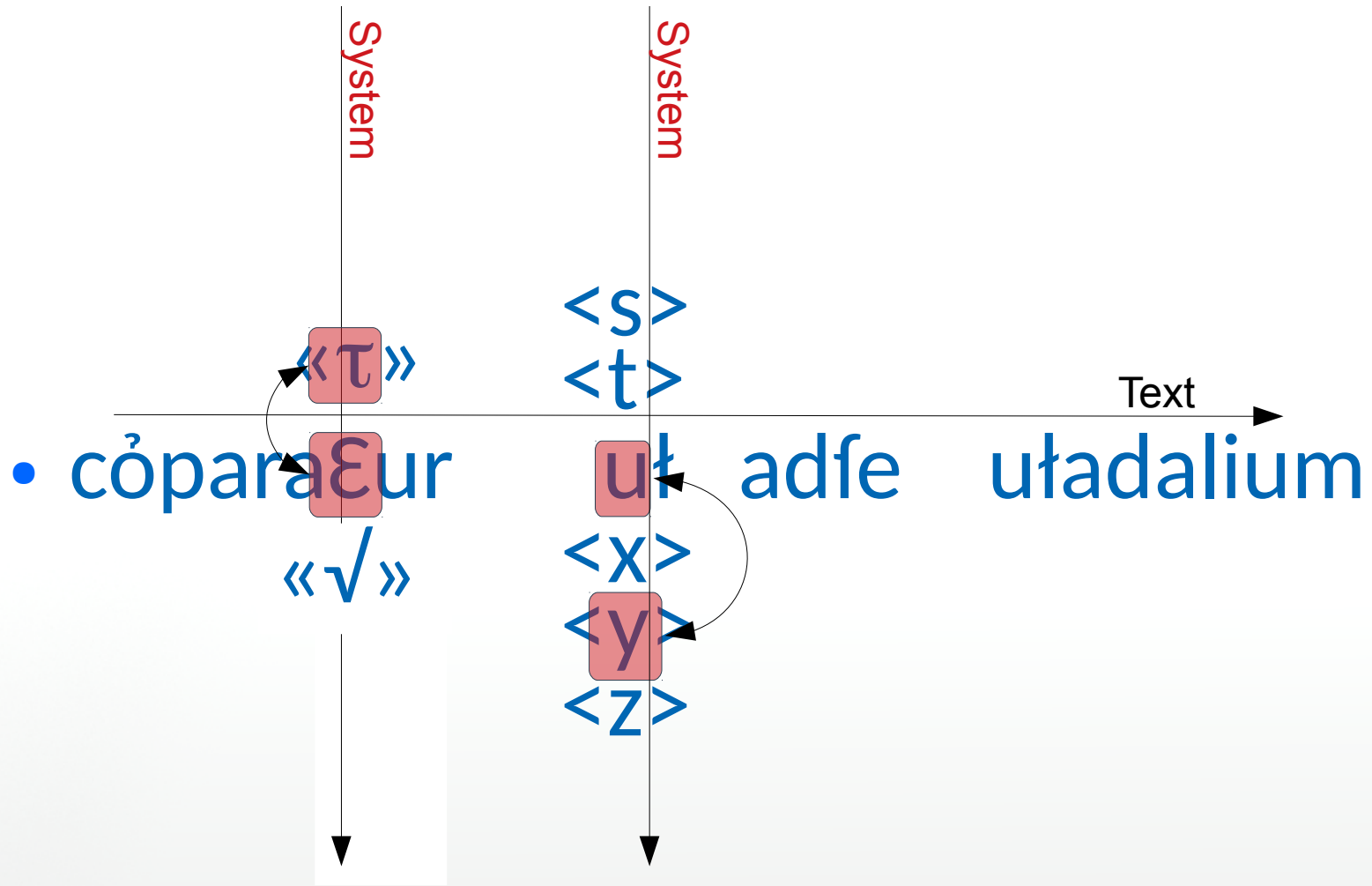
<x>

<y>

<z>



Graphemes/allographs: the commutation test



Graphemes/allographs: the commutation test

• còparaEur

Substitution:

→ No change
in “denotative
meaning”

«τ»

«√»

<s>
<t>

uł

<x>

<y>

<z>

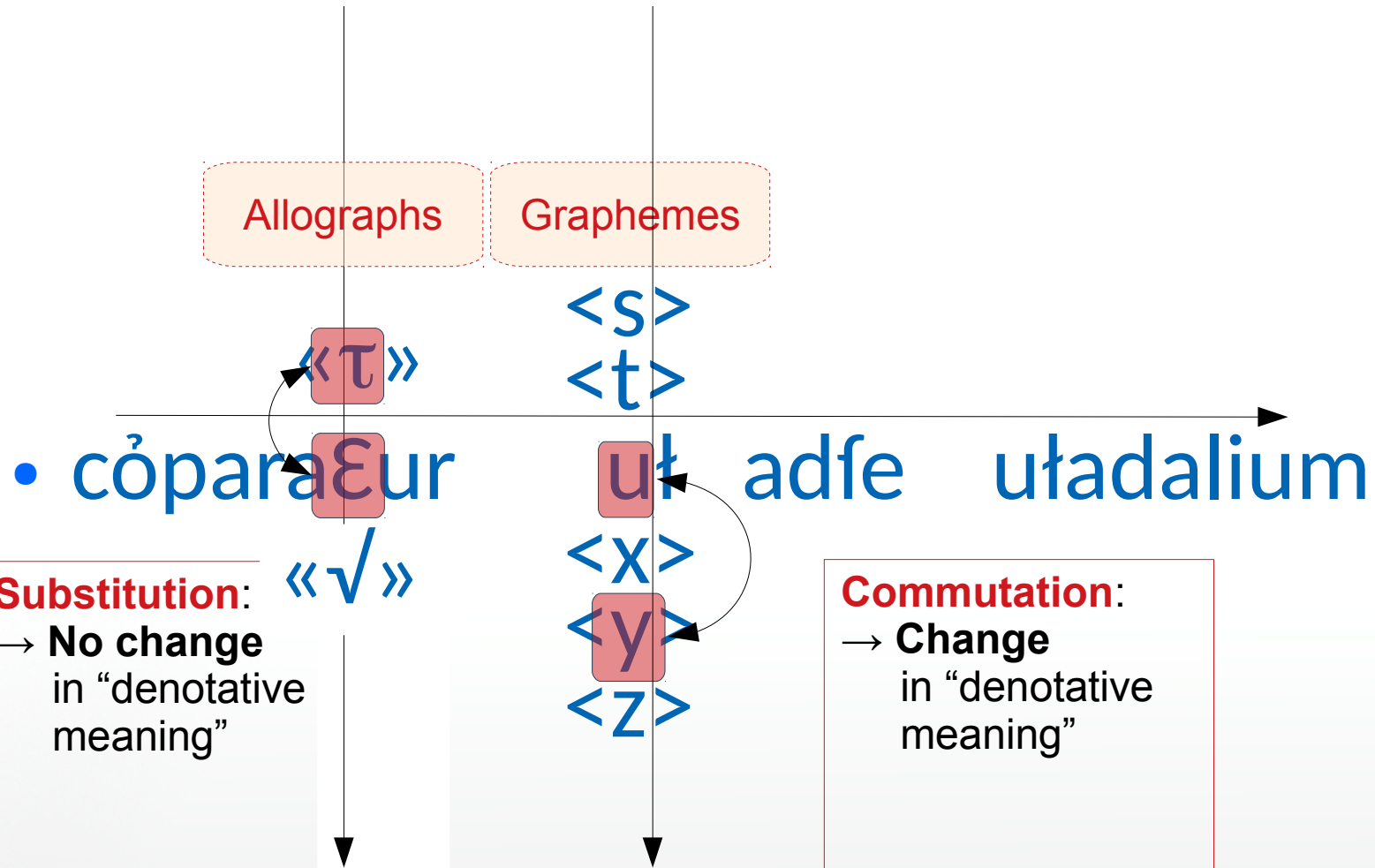
adfe

uładaliu

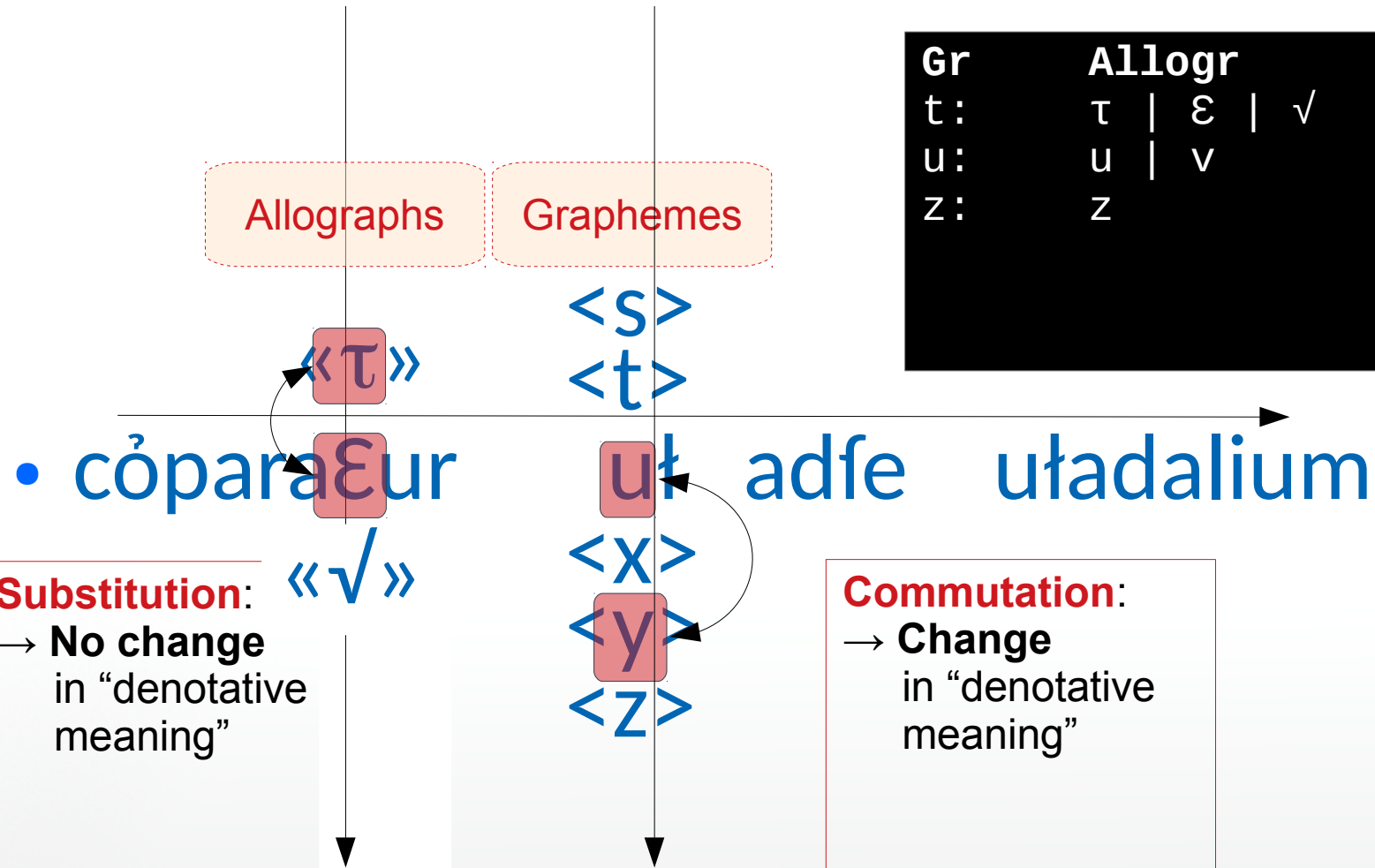
Commutation:

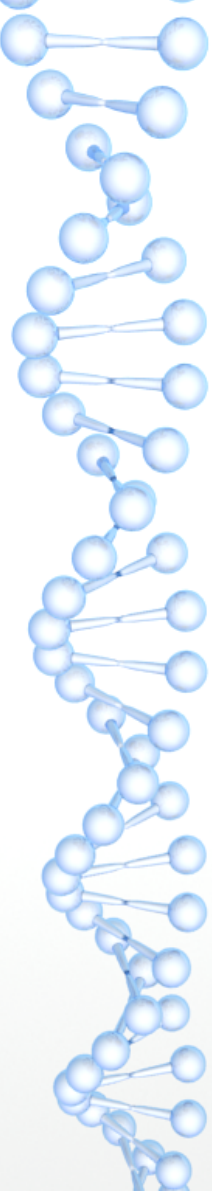
→ Change
in “denotative
meaning”

Graphemes/allographs: the commutation test



Graphemes/allographs: the commutation test

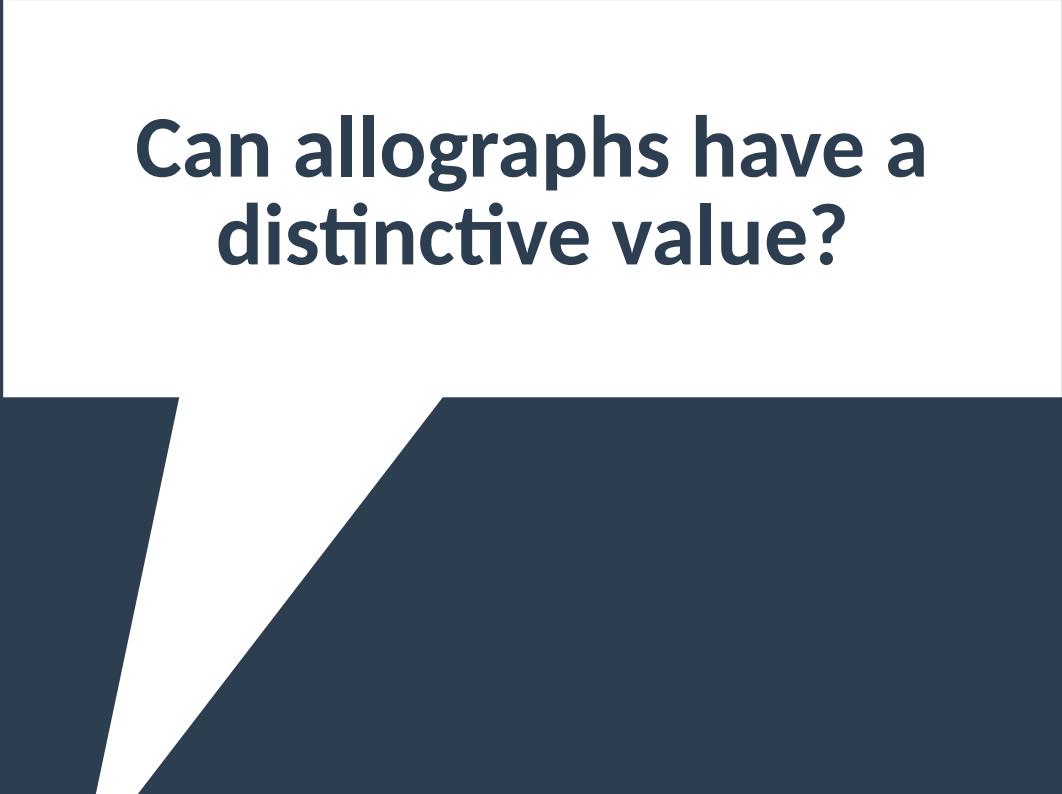




Graphemes / allographs: what to transcribe?

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

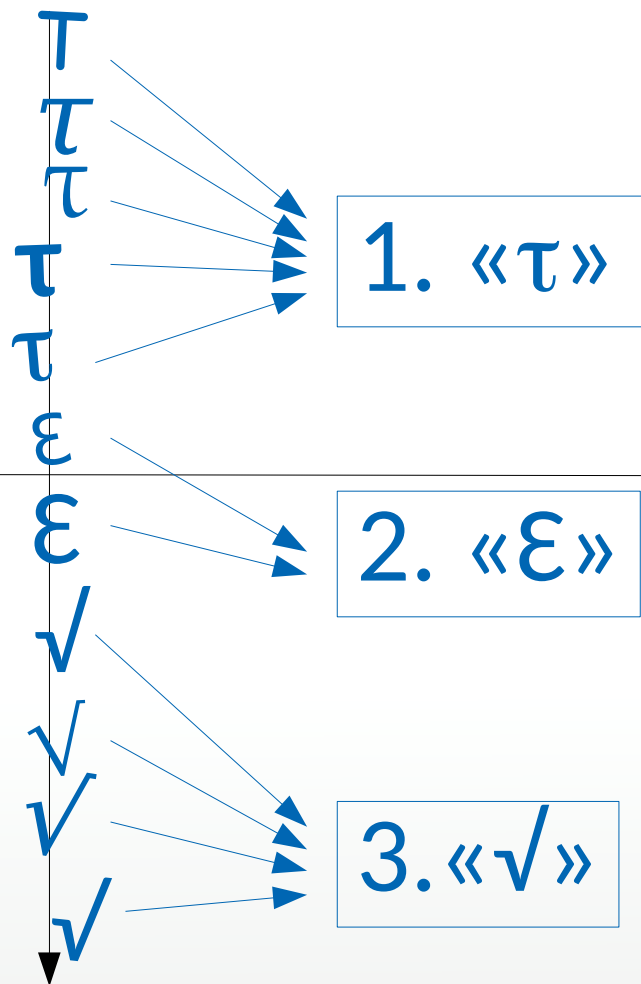


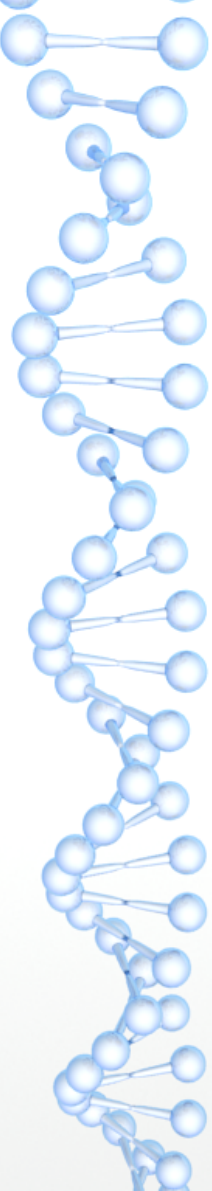
A white speech bubble with a dark blue background. The bubble has a rectangular body and a triangular tail pointing downwards and to the left.

**Can allographs have a
distinctive value?**

A vertical stack of symbols: τ , τ , τ , τ , ϵ , ϵ , $\sqrt{}$, $\sqrt{}$, $\sqrt{}$, and a downward arrow.

Allographs





Capitals: allographs or graphemes?

- Cool (CA) is a cool town
- Smith is a good smith
- ODD files are odd files

Geographical name

Proper name

Acronym

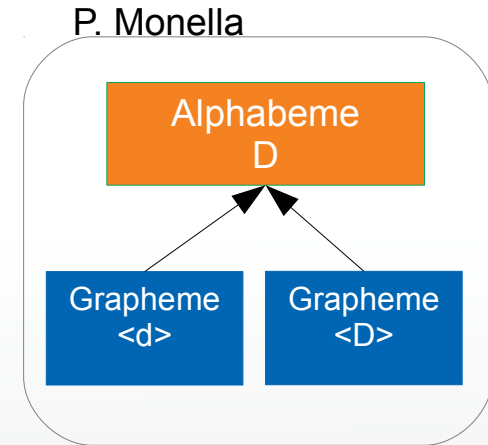
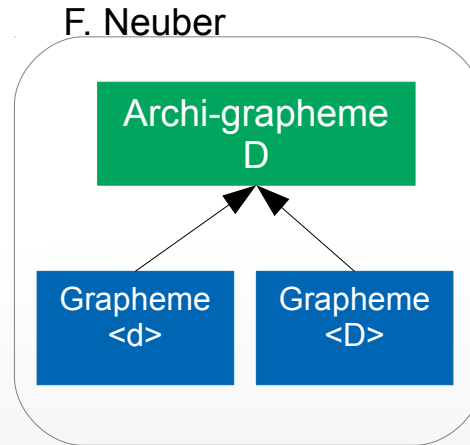
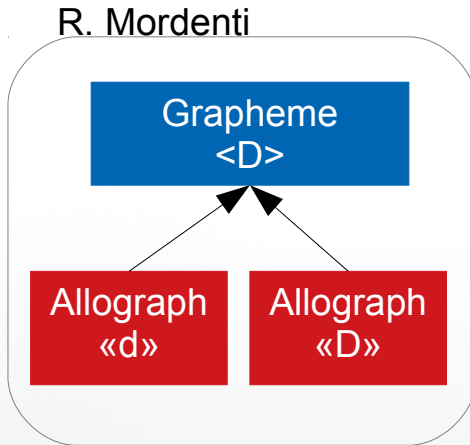


OK for contemporary
Western writing systems

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*



Sentence segmentation: distinctive value for meaning of the whole text

- I go because I have to. **S**tay here!
I go because I have to **s**tay 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, fuccefs, dafs (daß)



Word segmentation: distinctive value for meaning of the whole text

- σαῦρος, fuccefs, dafs (daß)

Paulus suftinet me

(Paolo holds me up)

Paulus fus tinet me

(Paolo the pig holds me)

Positional
allograph



Word segmentation: distinctive value for meaning of the whole text

- σαῦρος, fuccefs, dafs (daß)

Paulus suftinet me

(Paolo holds me up)

Paulus fus tinet me

(Paolo the pig holds me)

Positional
allograph

Space



Connotators



Connotators



Connotators

who



Connotator
“Gothic”
(marked)

≠

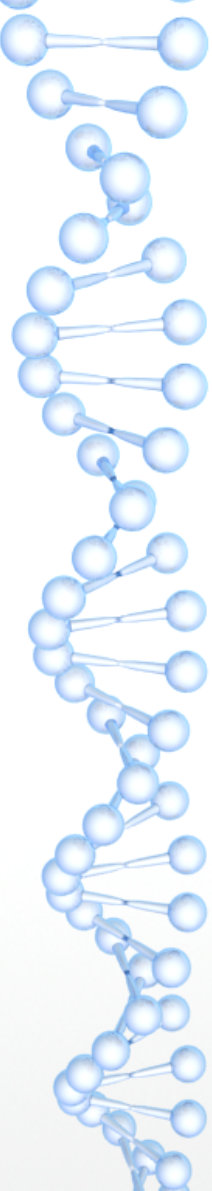


Pertinence

WHO



Connotator
“Gaul”
(not marked)



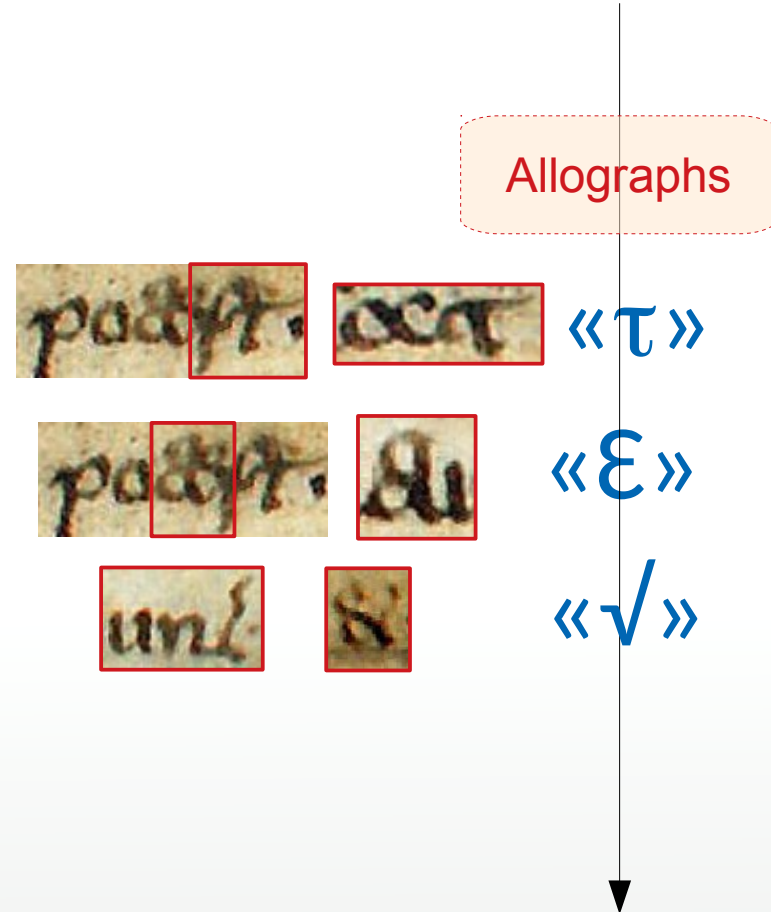
Connotators

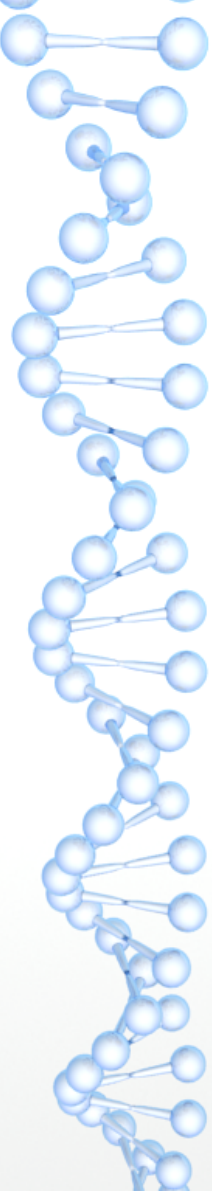
Connotators, pertinent for the writer

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

(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, f / s)





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?



OCR/HTT
(witness A)



Manual (selective)
transcription
(witness B)

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



Allographic
transcription

Vn_τer <hi>dem</hi>
schloss



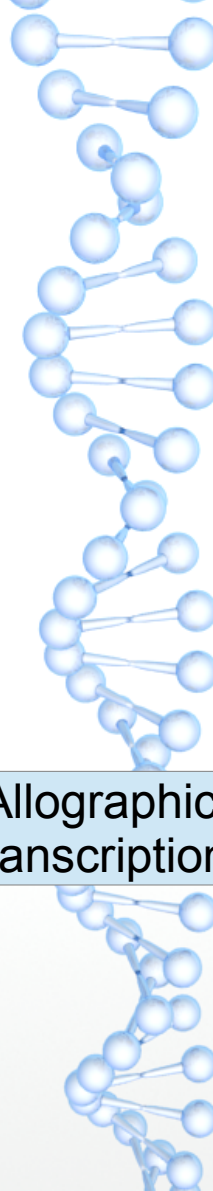
OCR/HTT
(witness A)

unter dem fchlofs



Manual (selective)
transcription
(witness B)

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



Allographic transcription

Vn_τer <hi>dem</hi>
schloss



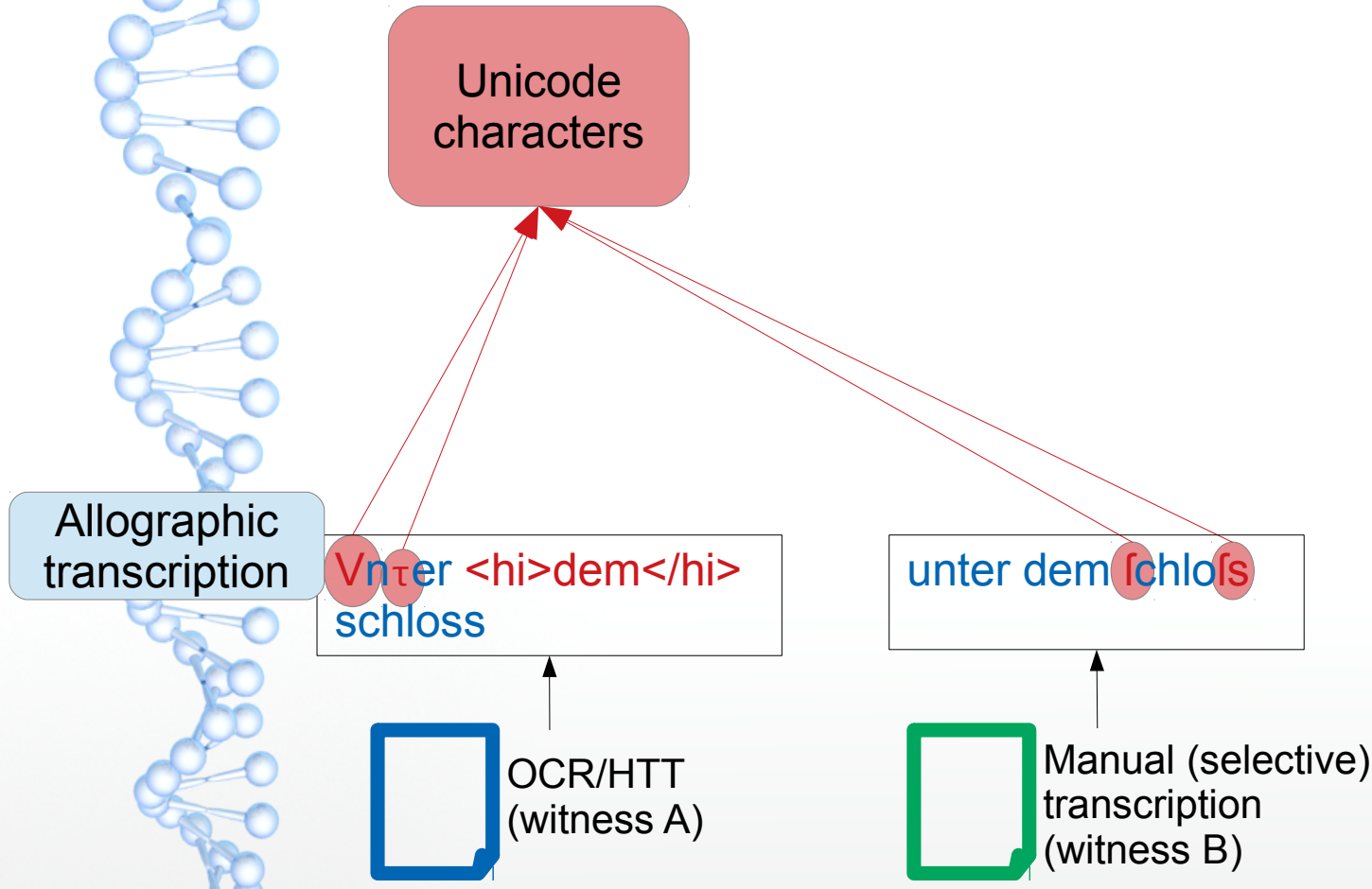
OCR/HTT
(witness A)

unter dem fchlofs

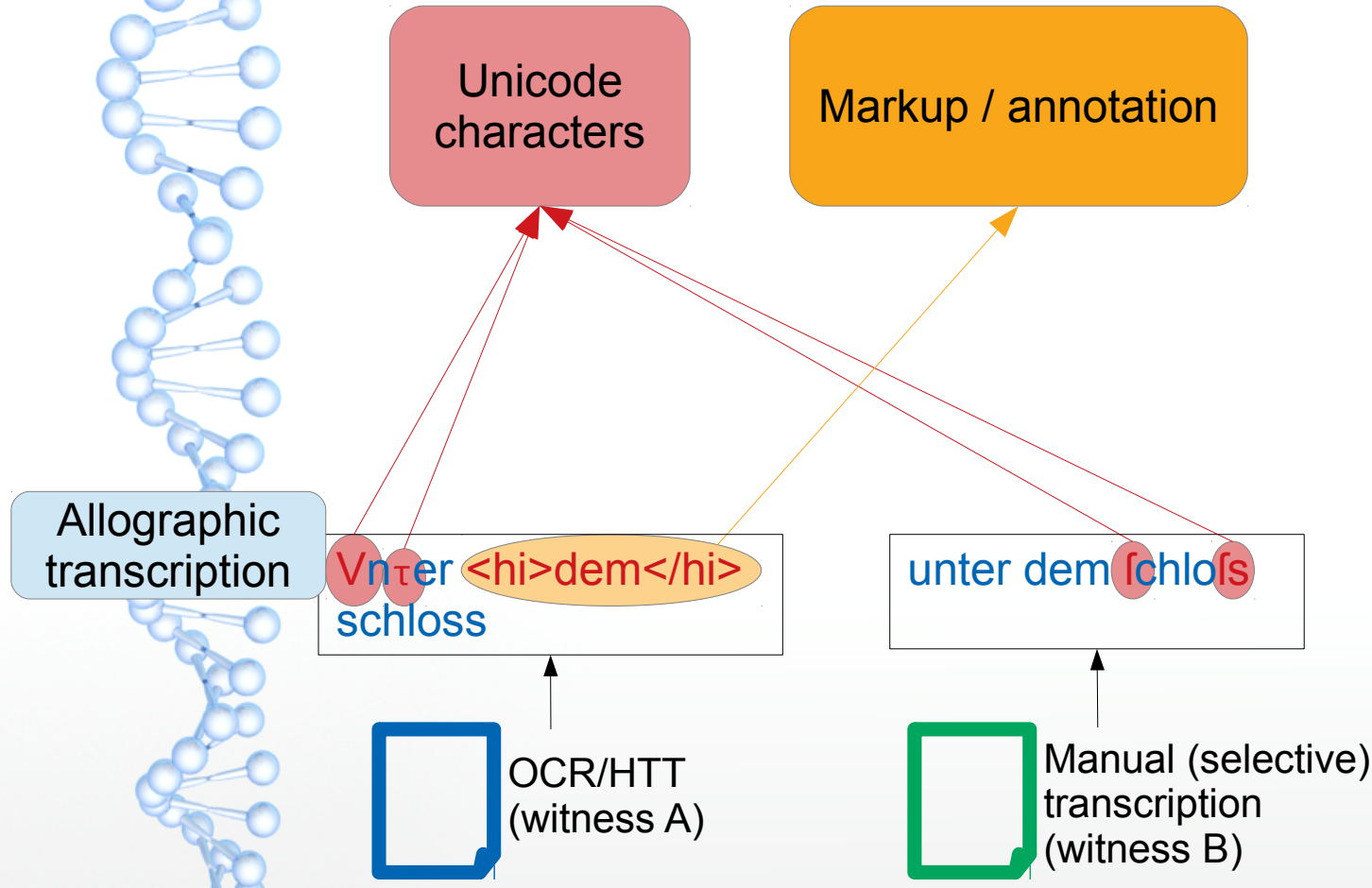


Manual (selective)
transcription
(witness B)

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



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



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

Allographic transcription

Vn_τer <hi>dem</hi>
schloss



OCR/HTT
(witness A)

unter dem fchlofs



Manual (selective)
transcription
(witness B)

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

Allographic transcription

Vn_τer <hi>dem</hi>
schloss



OCR/HTT
(witness A)

unter dem fchlofs



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

Vn^τer <hi>dem</hi>
schloss



OCR/HTT
(witness A)

Gr	Allogr
s:	s ſ
t:	t
u:	u

unter dem fchlofs



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

Gr	Allogr
s:	s
t:	τ ε √
u:	u v

Allographic transcription

Vn_τer <hi>dem</hi>
schloss

OCR/HTT
(witness A)

unter dem schloss

Gr	Allogr
s:	s r
t:	t
u:	u

unter dem fchlofs

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

Gr	Allogr
s:	s
t:	τ ε √
u:	u v

Allographic transcription

Vn_τer <hi>dem</hi>
schloss

OCR/HTT
(witness A)

unter dem schloss

Gr	Allogr
s:	s r
t:	t
u:	u

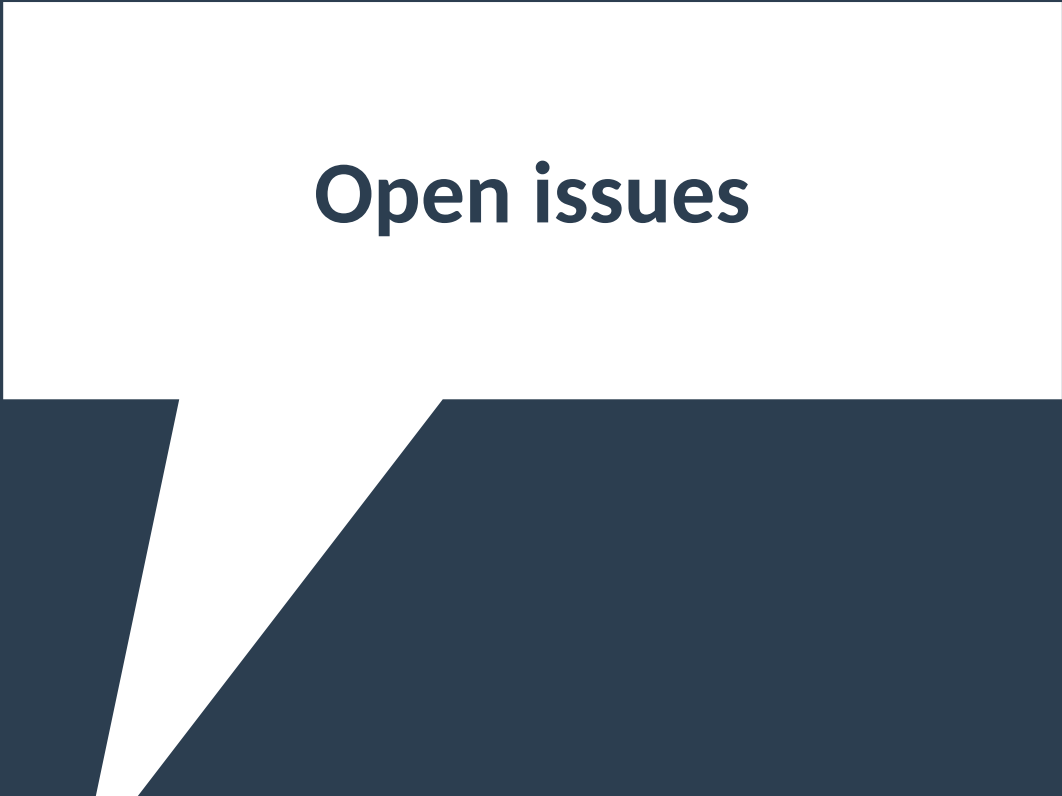
unter dem fchlofs

Manual (selective)
transcription
(witness B)

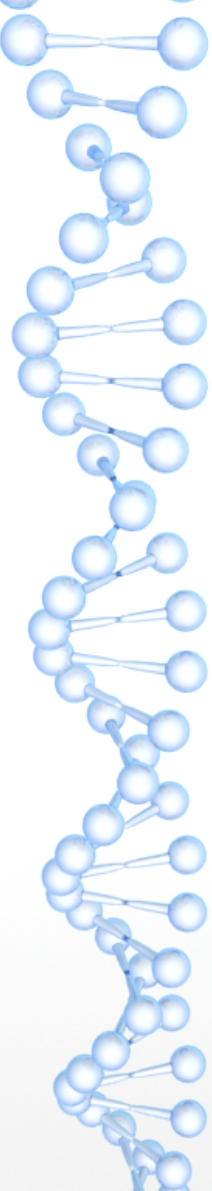
(More) interoperability

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

- Historical documentation
- Visualization
- Processing
 - (Erkenntnispotentiale)

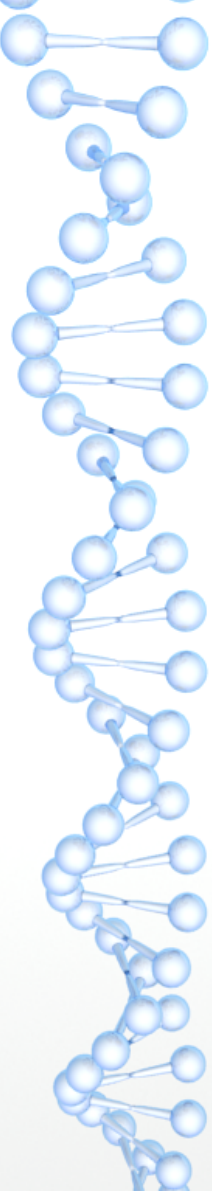
A white speech bubble with a dark blue background. The bubble has a rectangular body and a triangular tail pointing downwards and to the left. The text "Open issues" is centered within the rectangular body.

Open issues



Open issues

- Allographic words
 - Spelling (wife / wyffe)

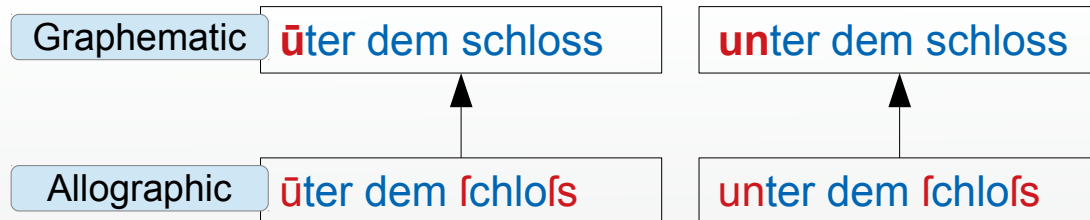


Open issues

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

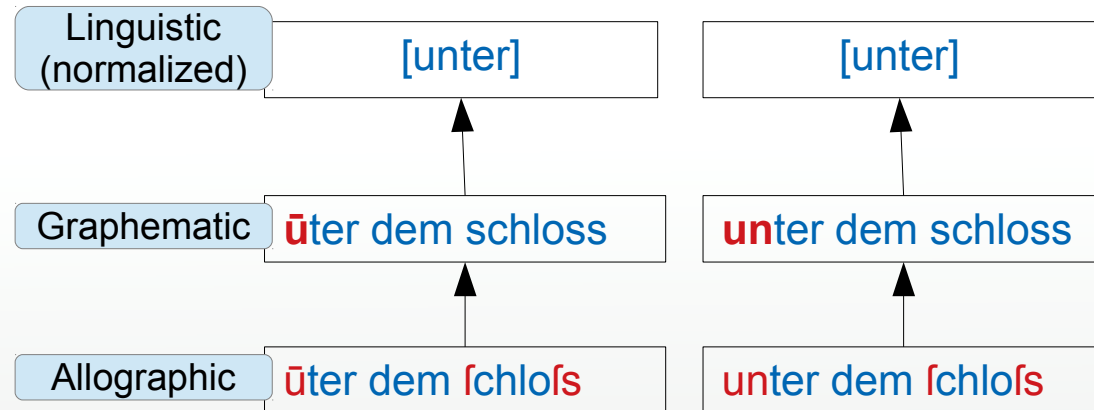
Open issues

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



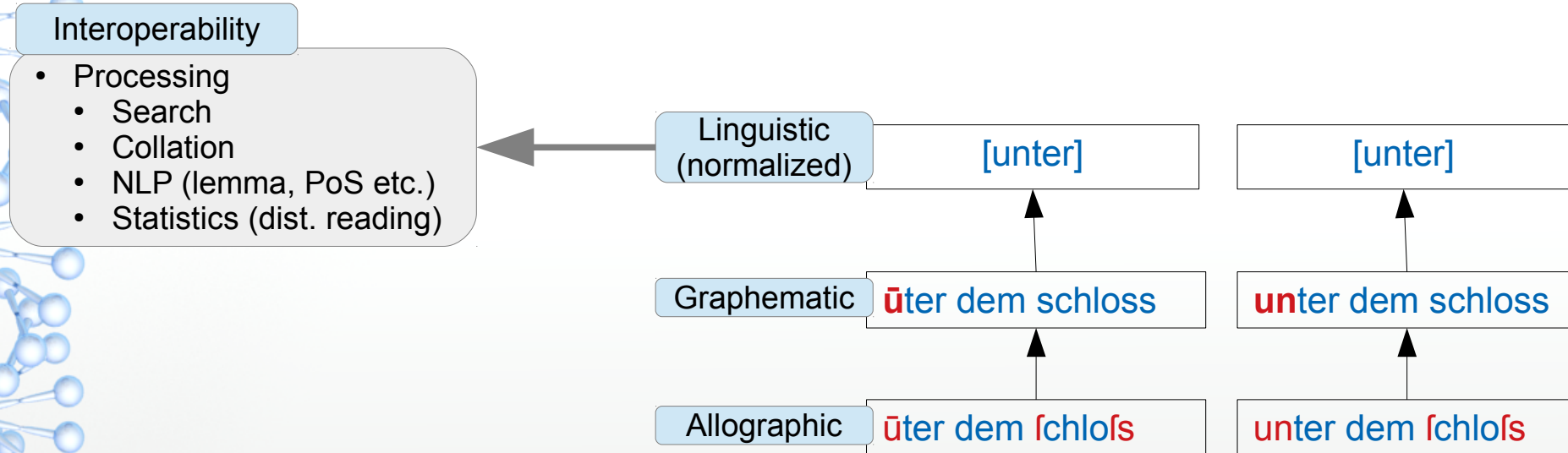
Open issues

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



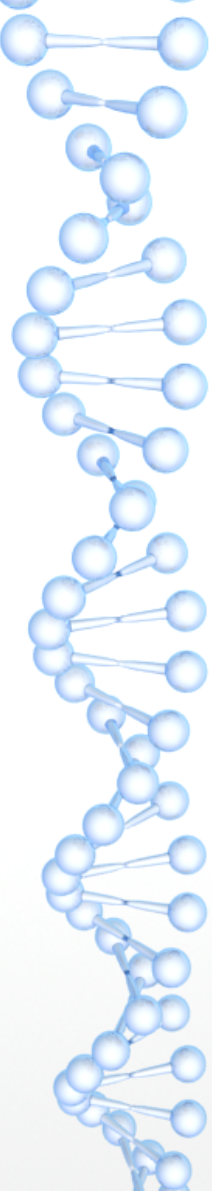
Open issues

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





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**