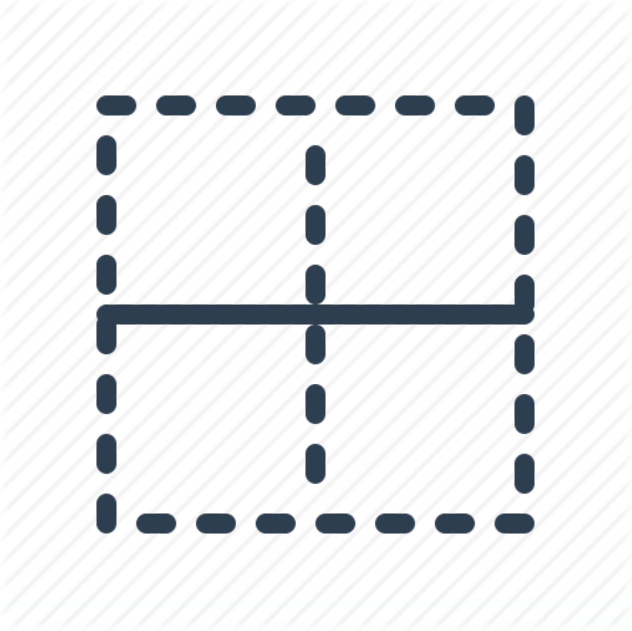
Paolo Monella

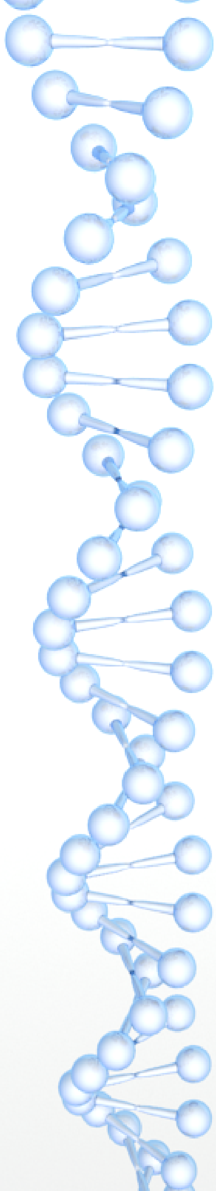
Encoding Pre-modern Writing Systems



SAPIENZA
UNIVERSITÀ DI ROMA

Sunoikisis Digital Classics,
Summer 2022 programme
May 12, 2022

European Research Council
Established by the European Commission



erc

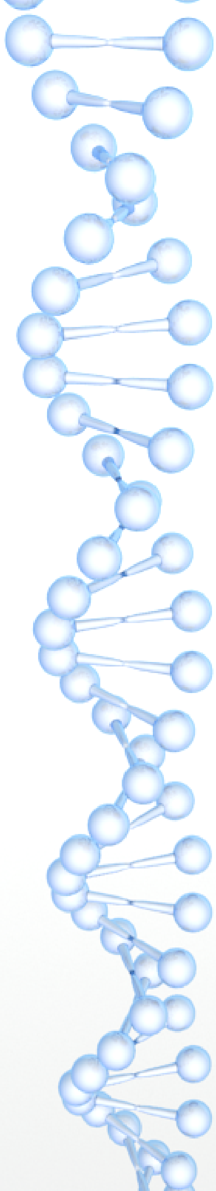
European Research Council

Established by the European Commission

ERC PAGES (AdG 2019 n° 882588)

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 882588)

Outline

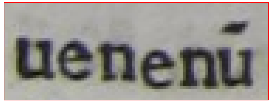


Outline

- **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**

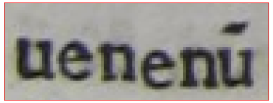
Interoperability: the issue

Interoperability: the issue



uenenú

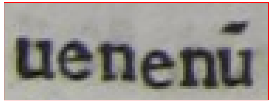
Interoperability: the issue



uenenū

- uenenū

Interoperability: the issue



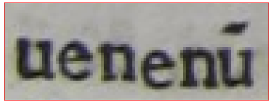
uenenū

- uenenū

Diplomatic

- Manual or HTR
  - Visualization
  - Processing

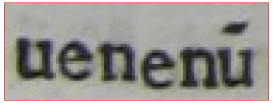
Interoperability: the issue



uenenú

- uenenū

Interoperability: the issue



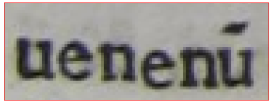
uenenū

- uenenū

- uenenum

Interoperability: the issue

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



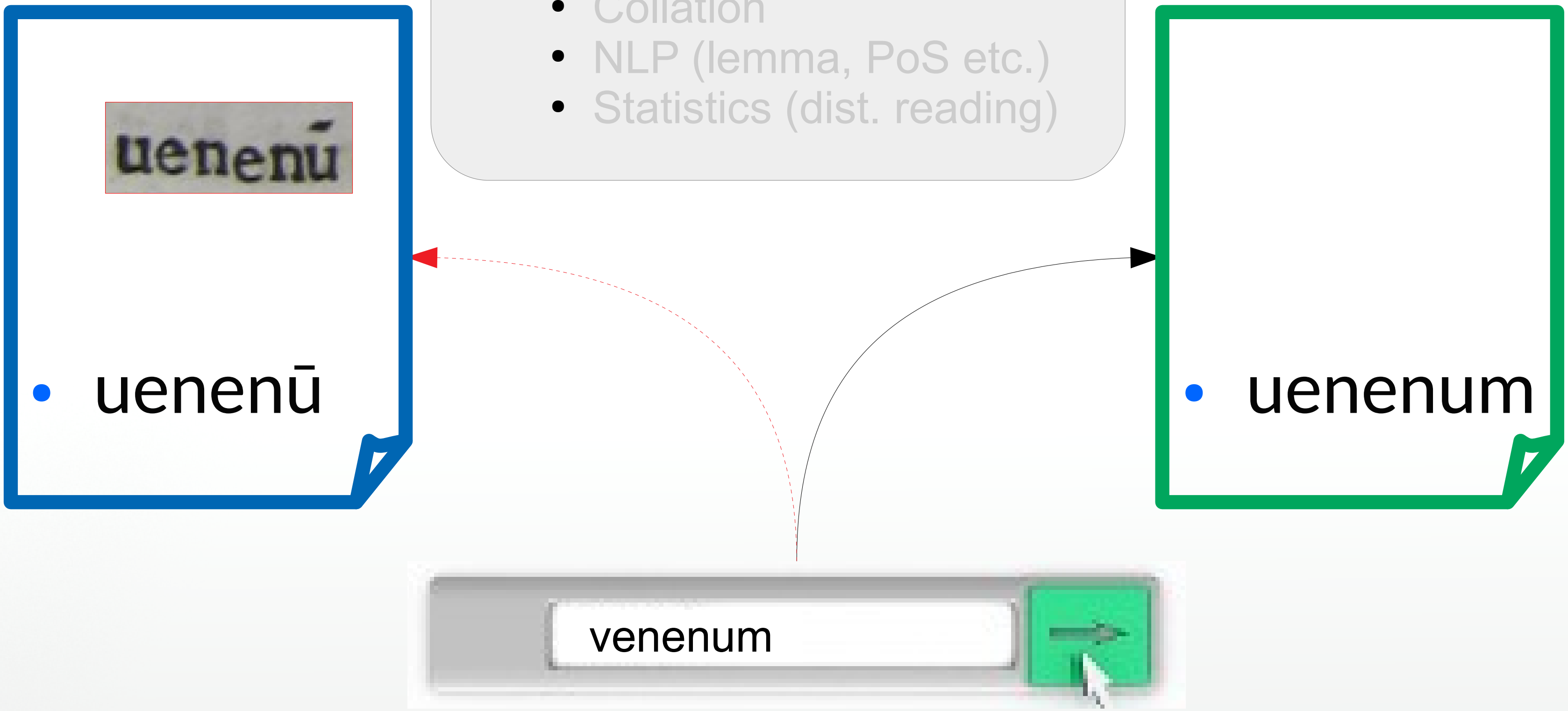
uenenū

- uenenū

- uenenum

Interoperability: the issue

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



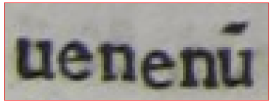
- uenenū

- uenenum

venenum

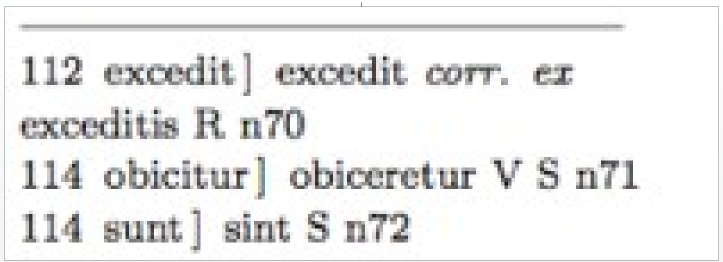
Interoperability: the issue

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

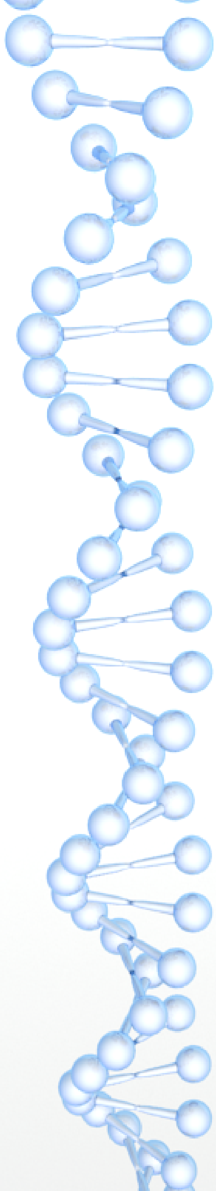


- uenenū

- uenenum

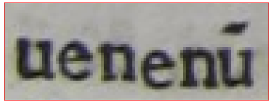


112 excedit] excedit *corr. ex* exceditis R n70
114 obicitur] obiceretur V S n71
114 sunt] sint S n72



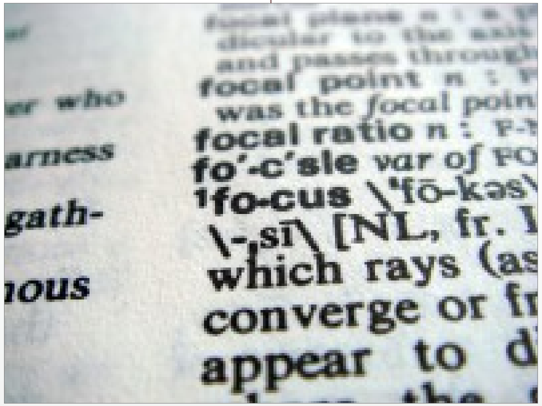
Interoperability: the issue

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

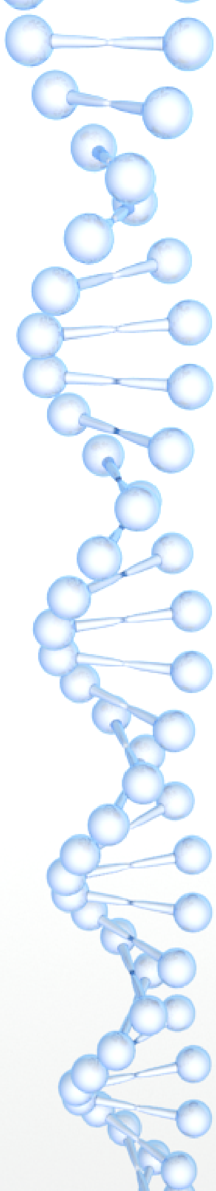


uenenú

- uenenū

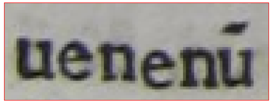


- uenenum



Interoperability: the issue

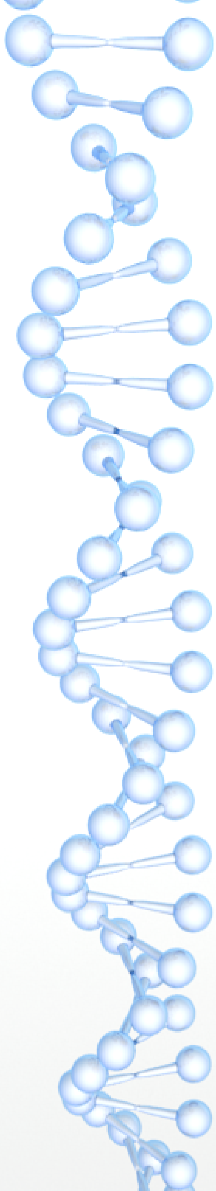
- Processing
  - Search
  - Collation
  - NLP (lemma, PoS etc.)
  - **Statistics (dist. reading)**



uenenú

- uenenū

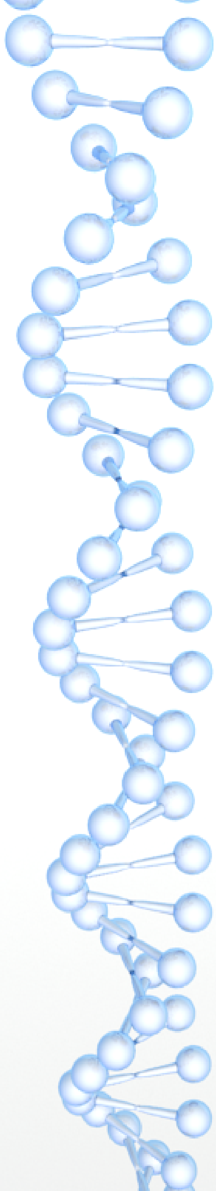
- uenenum



Interoperability: the issue

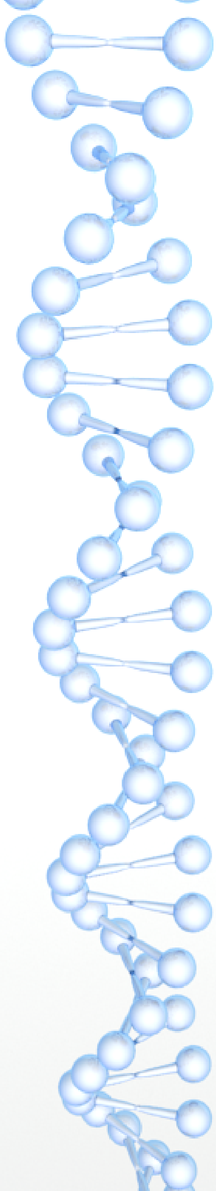
- ***My focus: European Medieval handwriting***
  - ...and early print (imitating handwriting)

Interoperability: current solutions



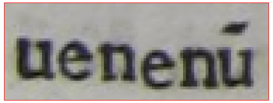
Unicode (TEI's recommendation)

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



Manual normalized transcription

- venenum



uenenū

- uenenū

Normalized

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

Diplomatic

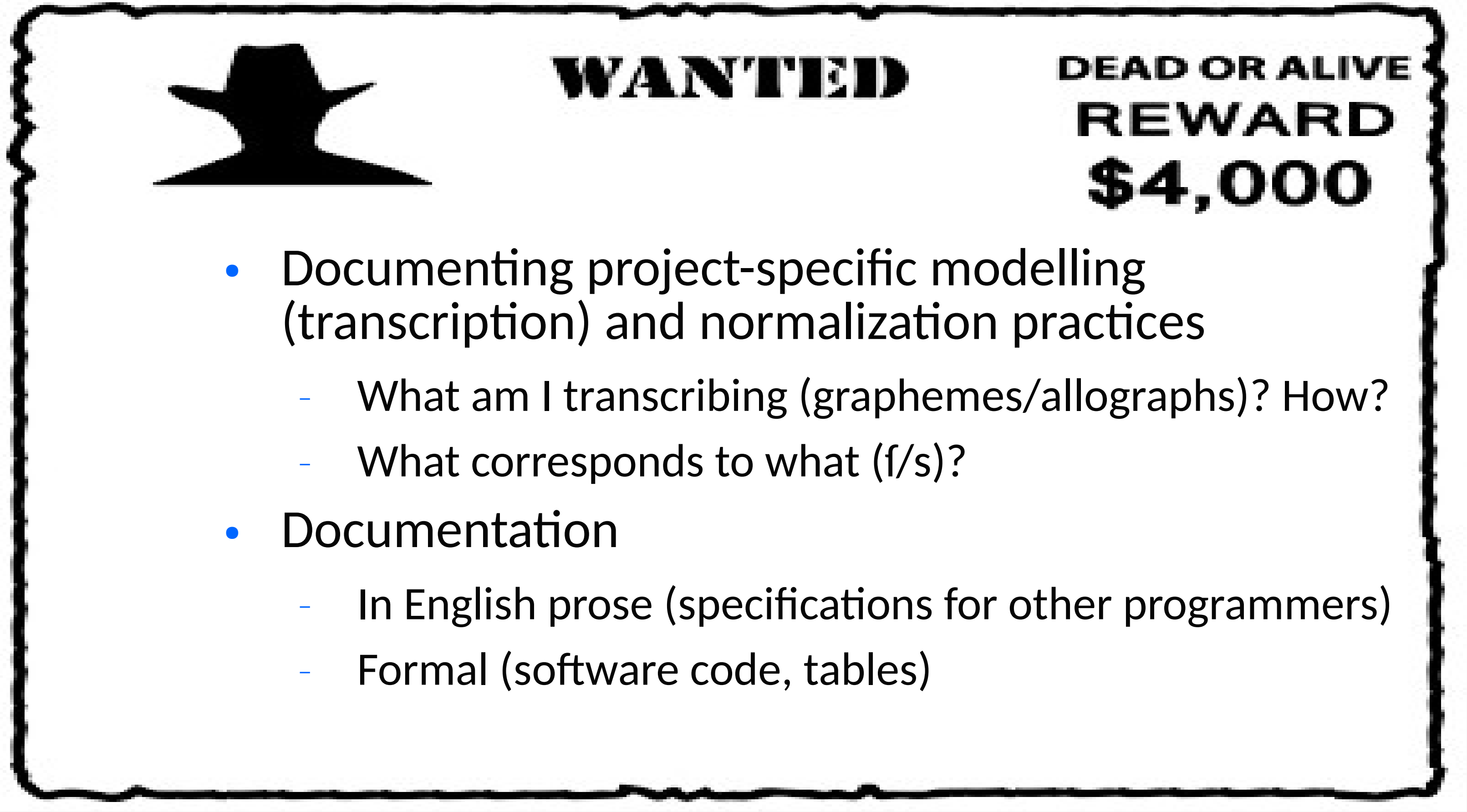
- Visualization
- ~~Processing~~

Project-specific pre-processing

- Collation: pre-processing to normalize the text
  - Or instruct collation software to ignore *some* discrepancies
- Same with search/indexing/NLP/statistics software
- Disposable home-made solutions
  - …for a shared issue

Interoperability through modelling

Interoperability through modelling/doc



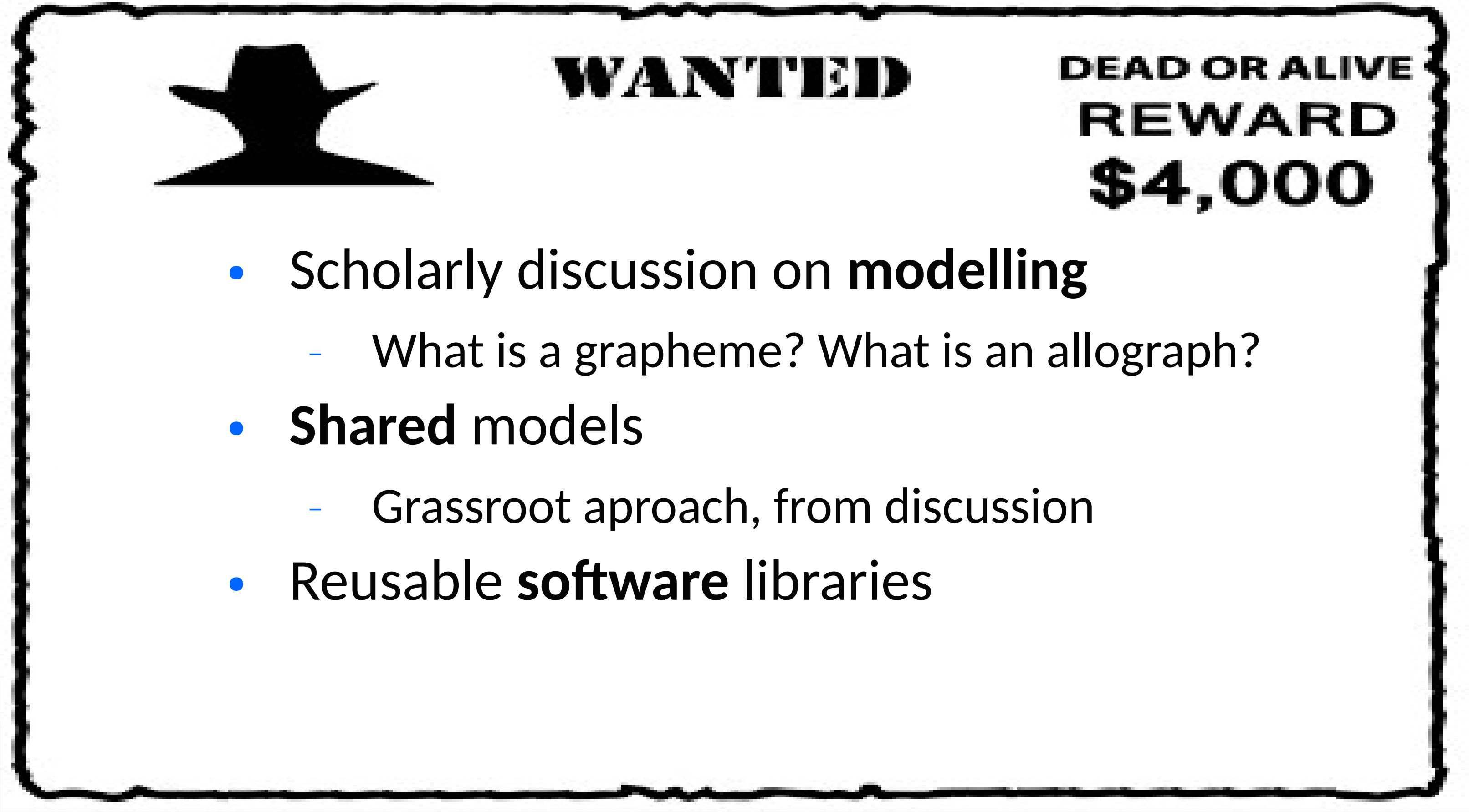
- Documenting project-specific modelling (transcription) and normalization practices
  - What am I transcribing (graphemes/allographs)? How?
  - What corresponds to what (ſ/s)?
- Documentation
  - In English prose (specifications for other programmers)
  - Formal (software code, tables)

Interoperability through modelling/doc

- Documenting project-specific modelling (transcription) and normalization practices
  - What am I transcribing (graphemes/allographs)? How?
  - What corresponds to what (ſ/s)?
- Documentation
  - In English prose (specifications for other programmers)
  - Formal (software code, tables)

Example ahead!

Interoperability through modelling



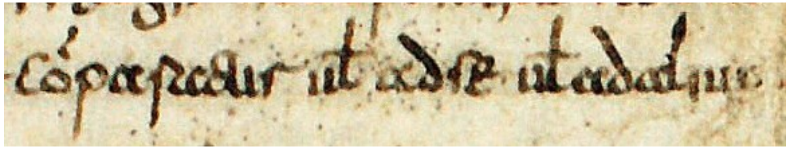
- Scholarly discussion on **modelling**
  - What is a grapheme? What is an allograph?
- **Shared** models
  - Grassroot aproach, from discussion
- Reusable **software** libraries

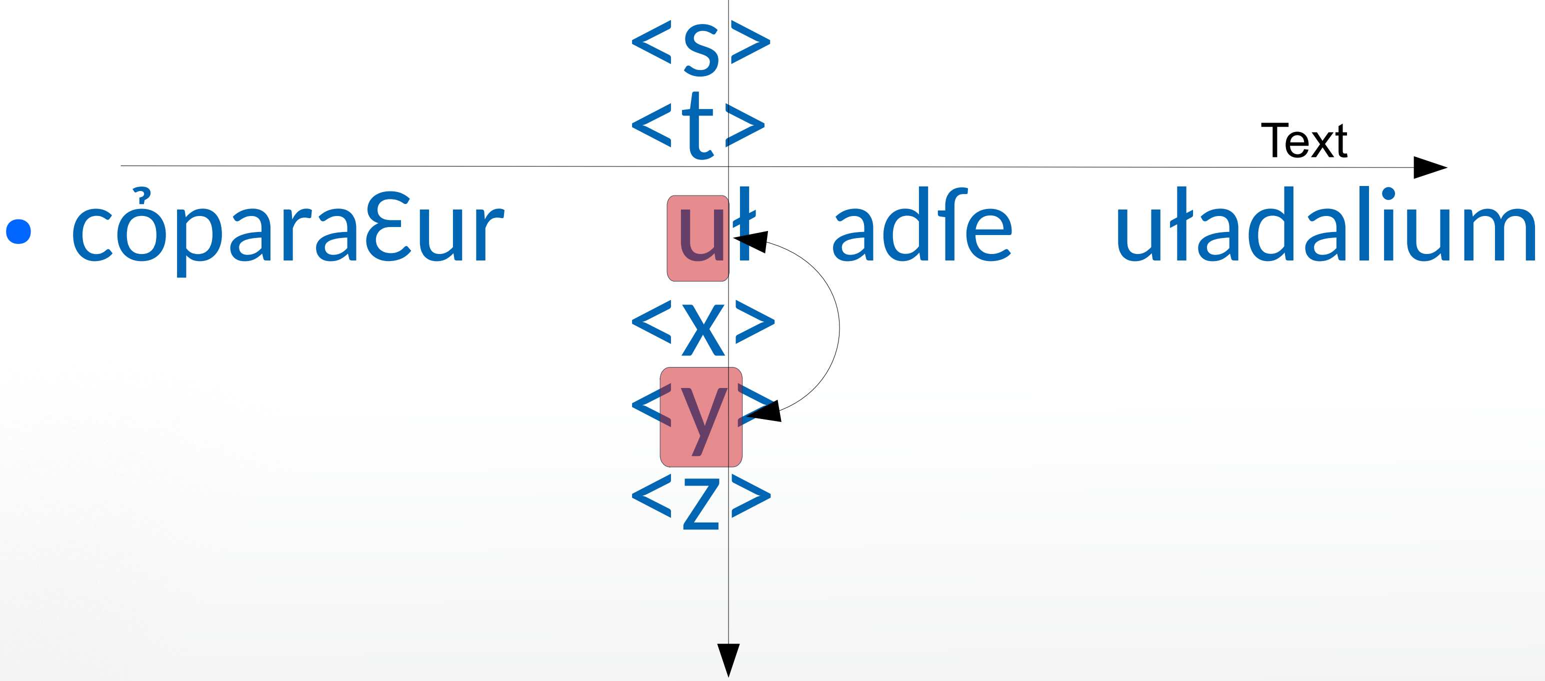
Graphemes/allographs

Graphemes/allographs: the commutation test

Comparatur vel ad se vel ad alium

He is compared to himself or to another





System

<s>

<t>

Text

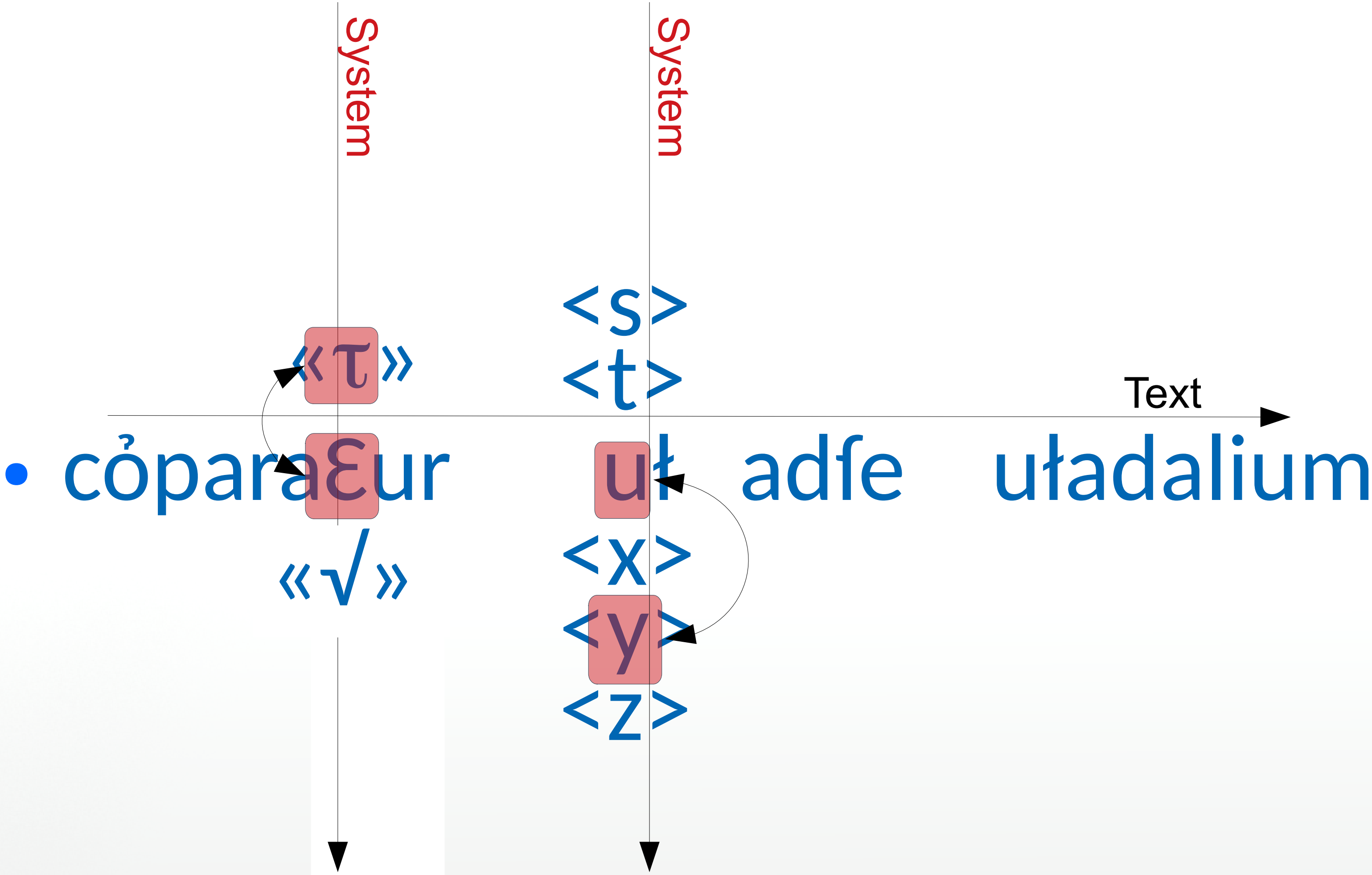
- cỏparaƐur uł adſe uładalium

<x>

<y>

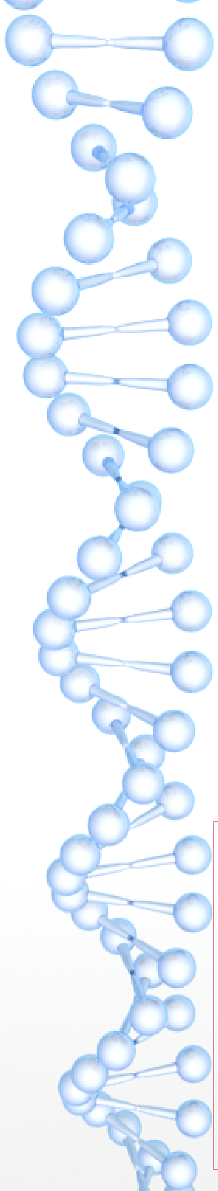
<z>

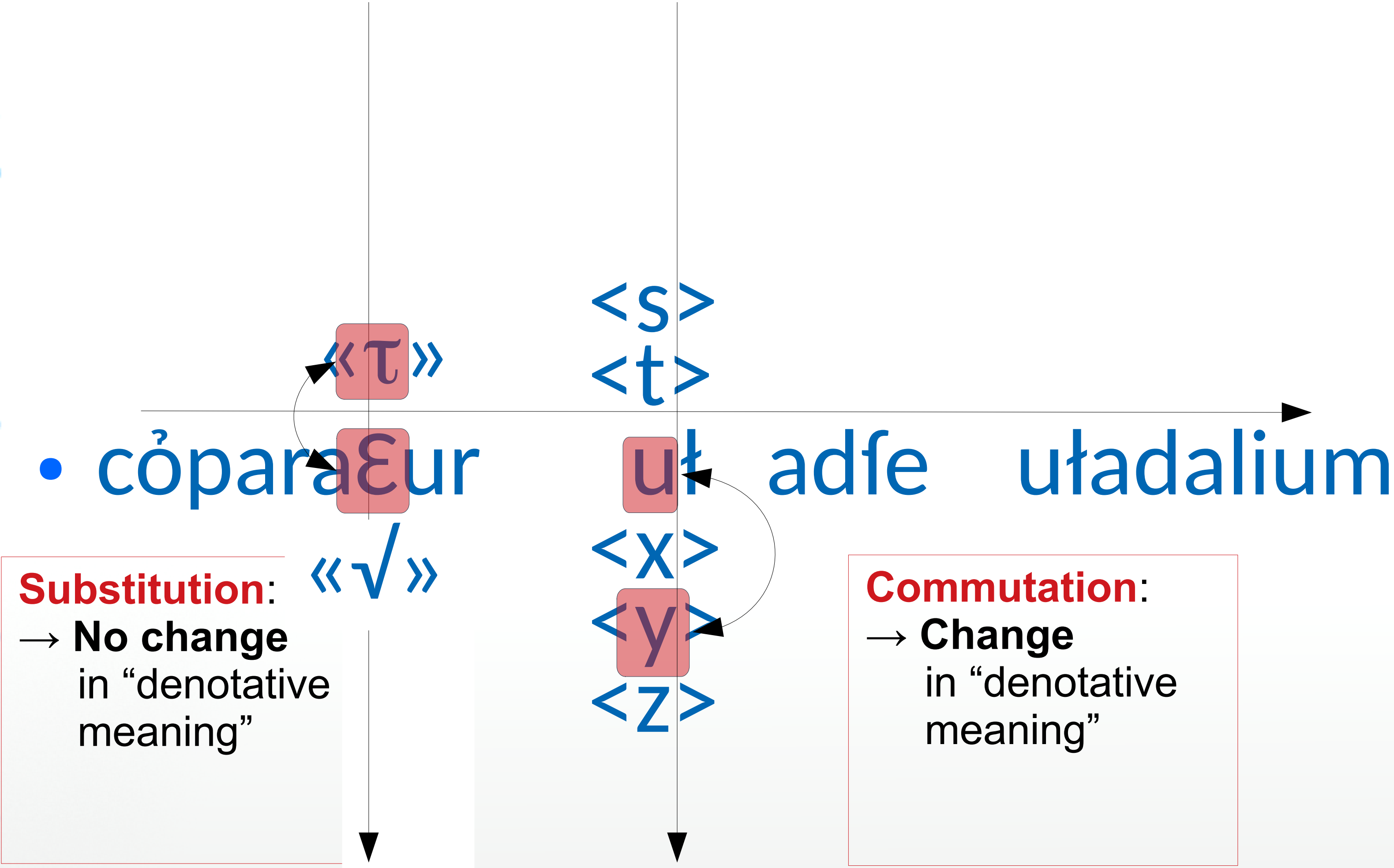
Graphemes/allographs: the commutation test



- cỏparaƐur uł adſe uładalium

Graphemes/allographs: the commutation test





«τ»

<s>

<t>

- cỏparaƐur uł adſe uładalium

«√»

<x>

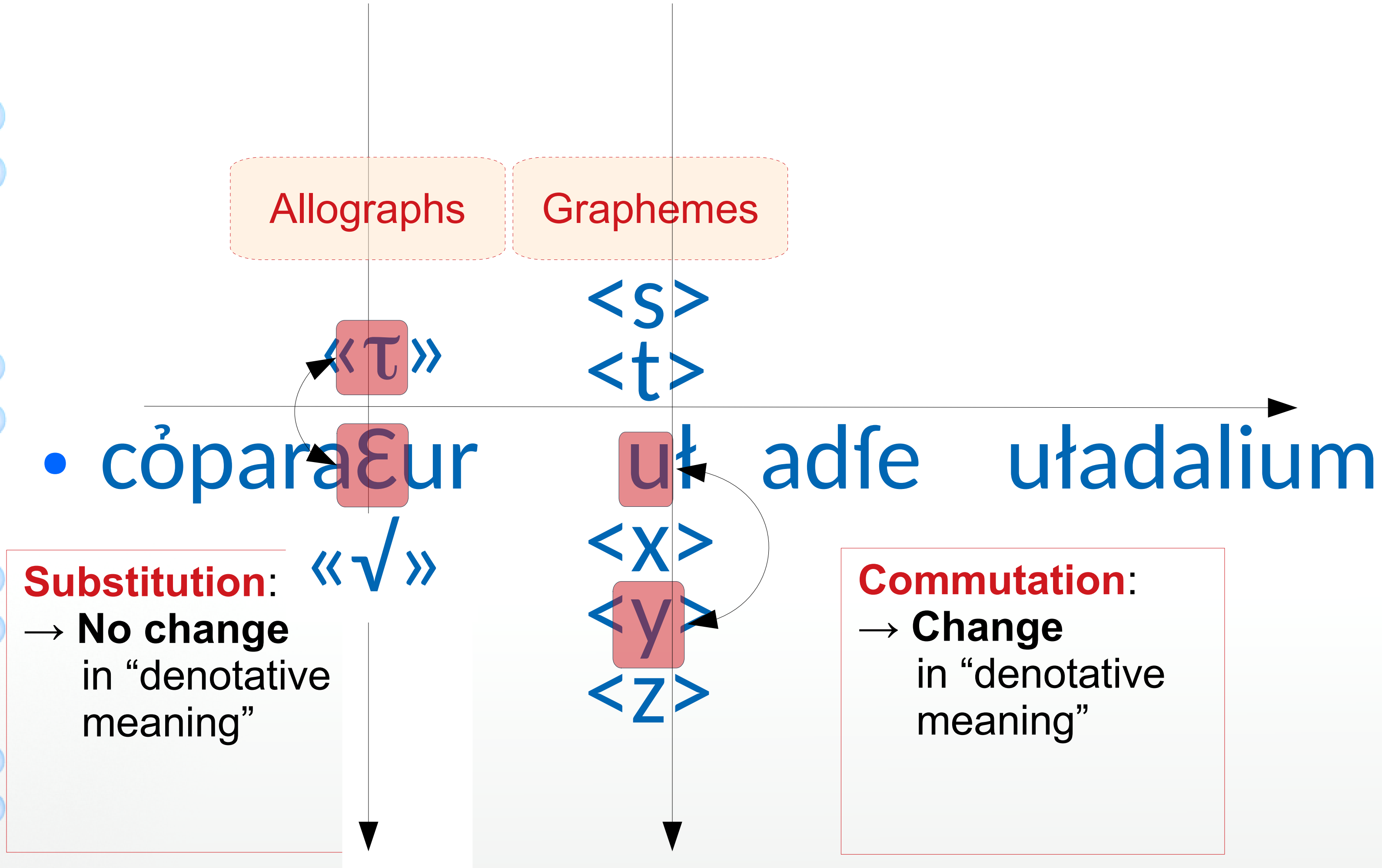
<y>

<z>

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>

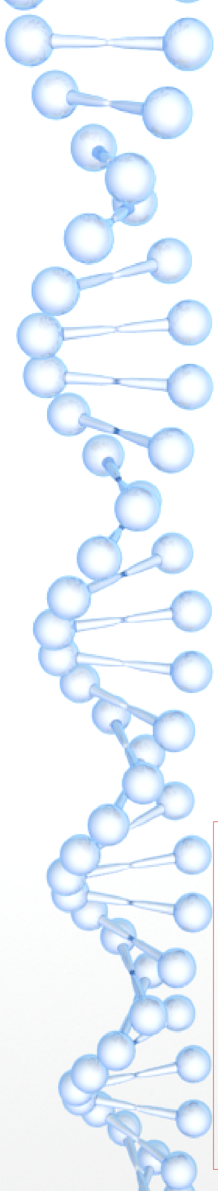
<y>

<z>

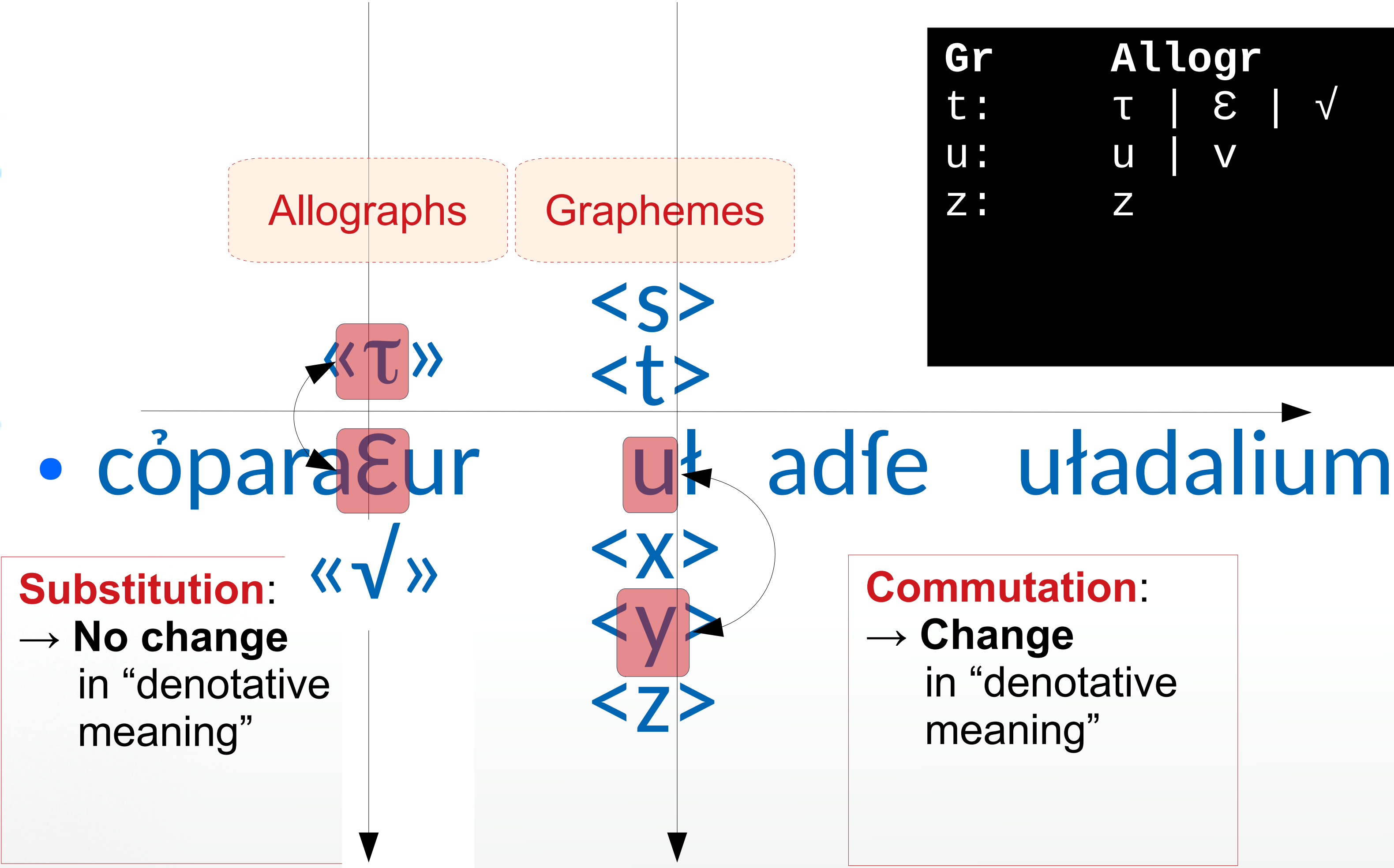
Substitution:
→ **No change**
in “denotative meaning”

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

Graphemes/allographs: the commutation test



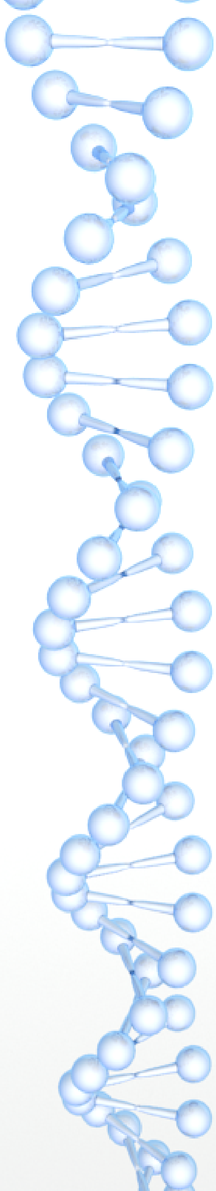
```
Gr       Allogr
t:       τ | ɛ | √
u:       u | v
z:       z
```



- cỏparaɛur uł adſe uładalium

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

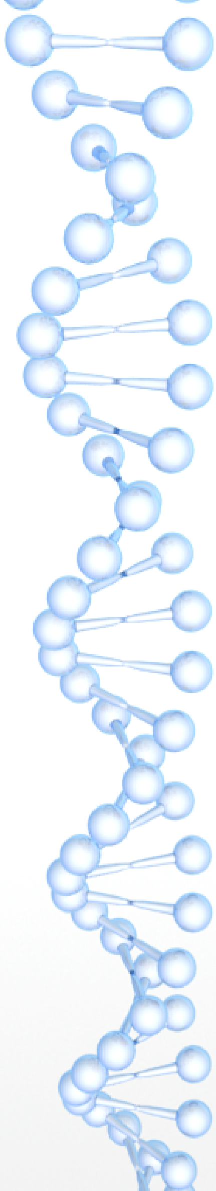
Commutation:
→ **Change**
in “denotative meaning”



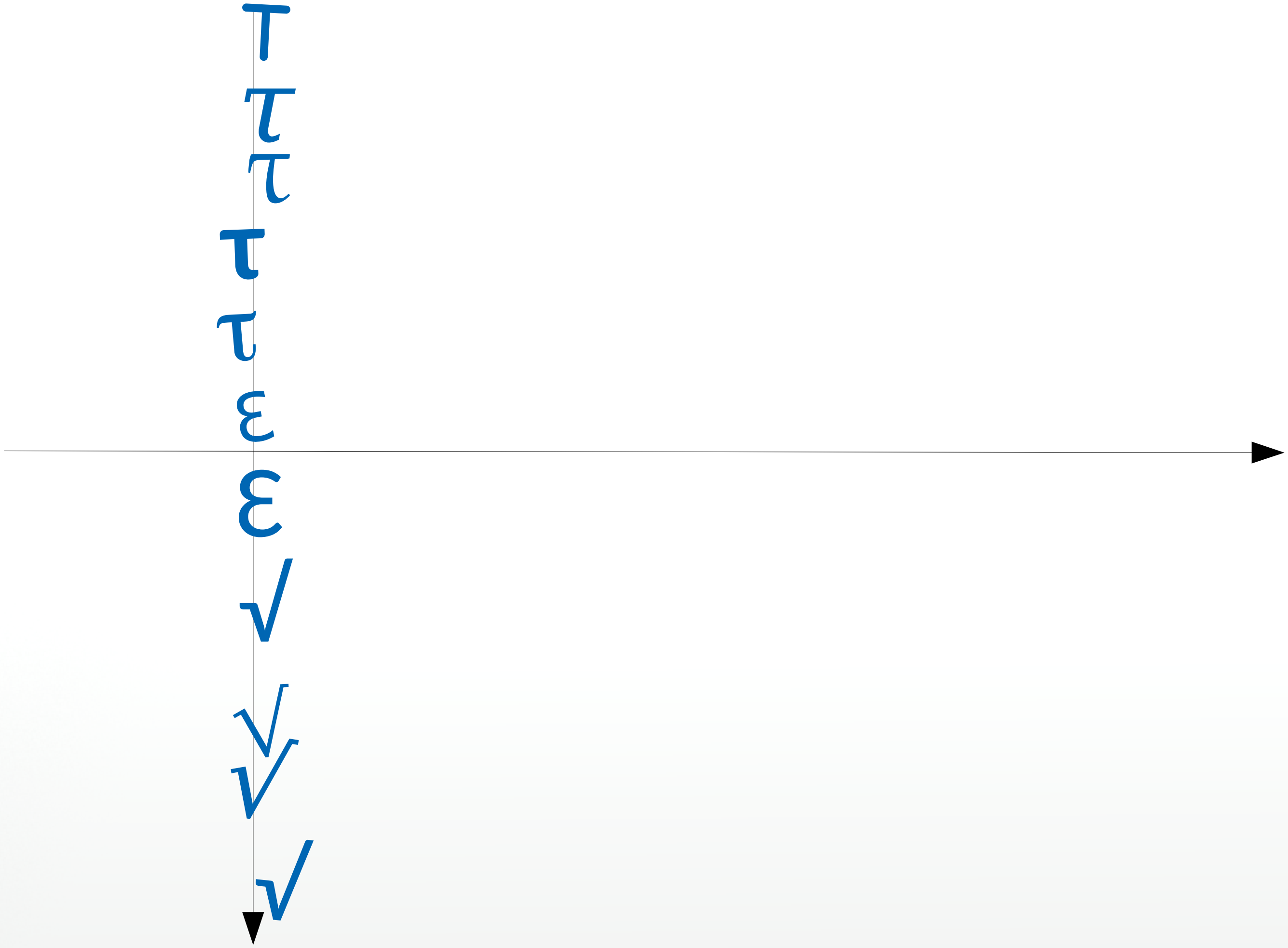
Graphemes / allographs: what to transcribe?

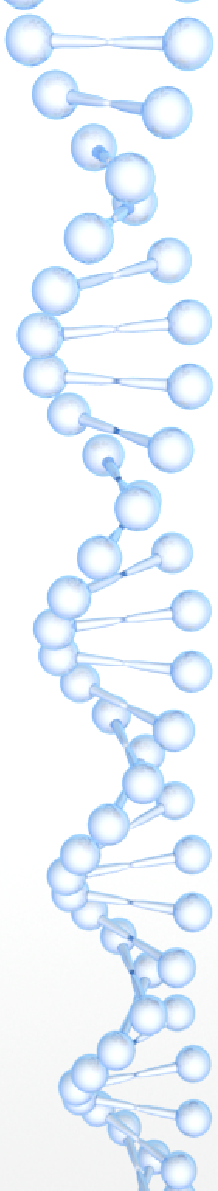
- Whatever our project needs!
  - Based on its scientific interests
  - (...and on time / money)
- But: *declaring* what we are transcribing
  - Based on a formal distinction, e.g., between graphemes / allographs
  - Documenting it
    - English prose
    - Formal (software, tables...)

Can allographs have a distinctive value?

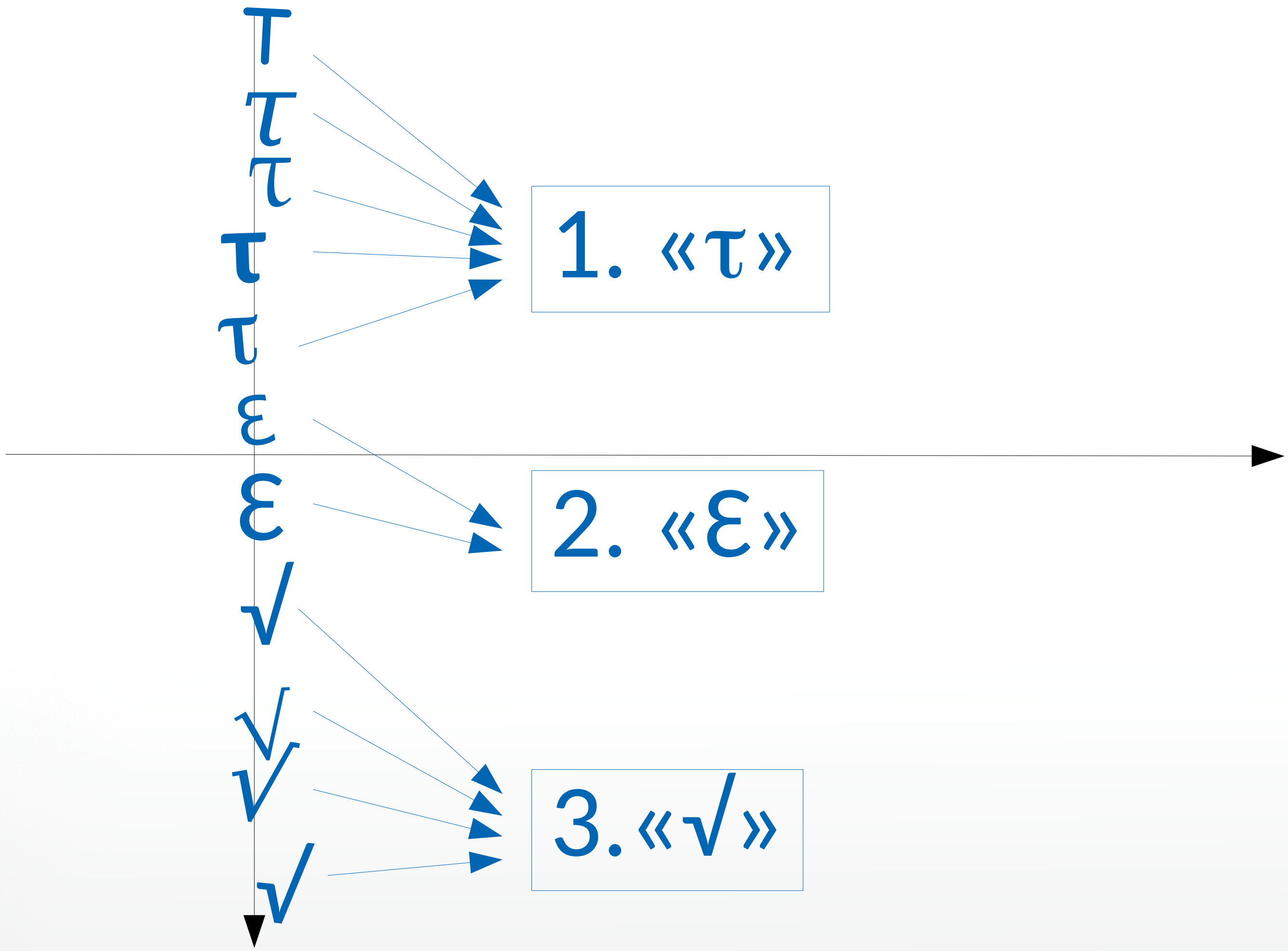


Allographs





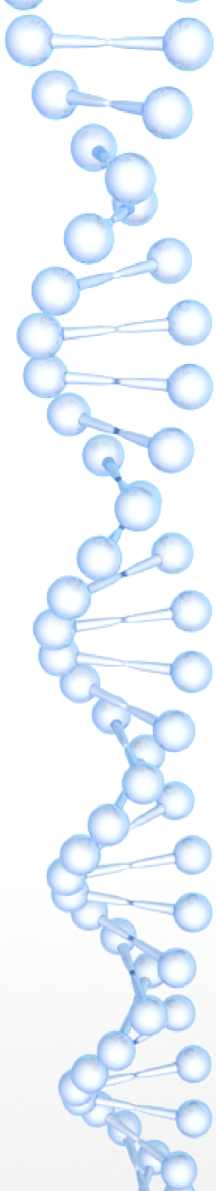
Allographs



1. «τ»

2. «ε»

3.«√»

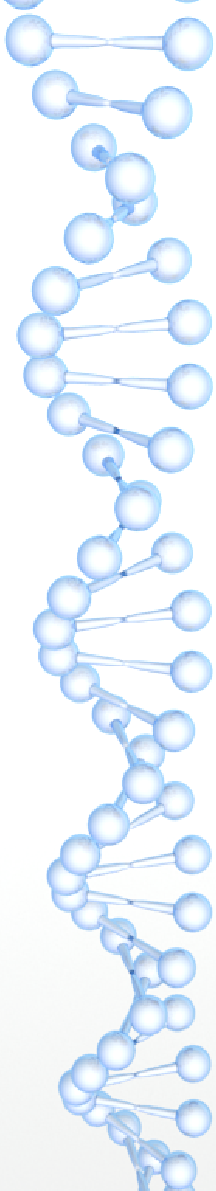


Capitals: allographs or graphemes?

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

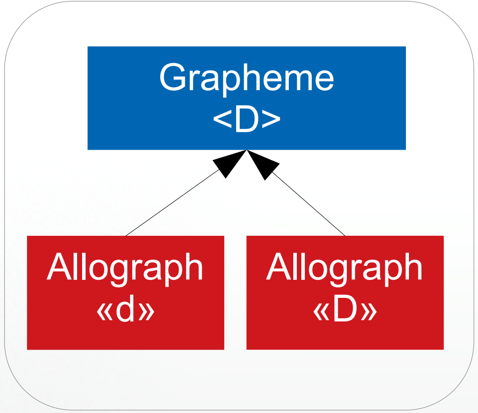
Not for classical/medieval handwriting (see later)



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

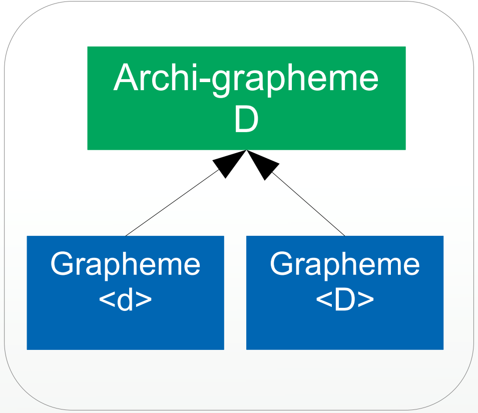


Grapheme <D>

Allograph «d»

Allograph «D»

F. Neuber

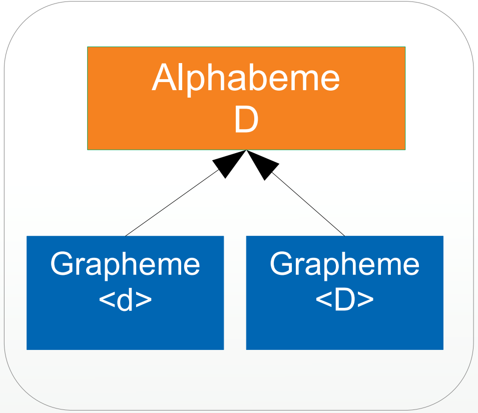


Archi-grapheme D

Grapheme <d>

Grapheme <D>

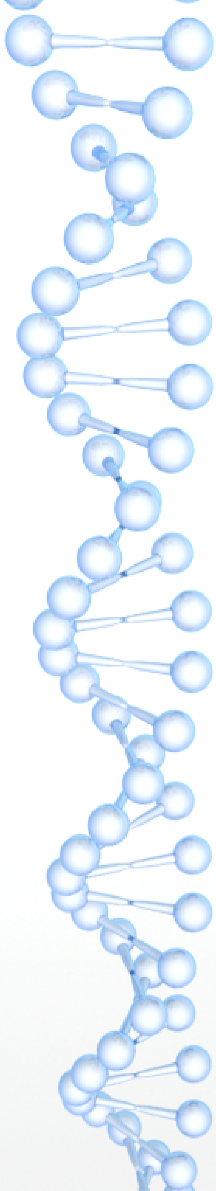
P. Monella



Alphabeme D

Grapheme <d>

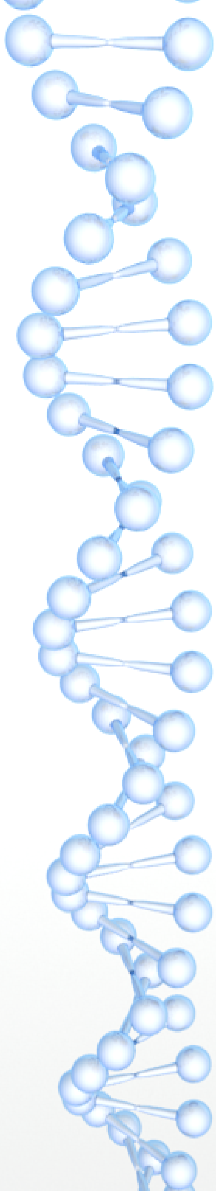
Grapheme <D>



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

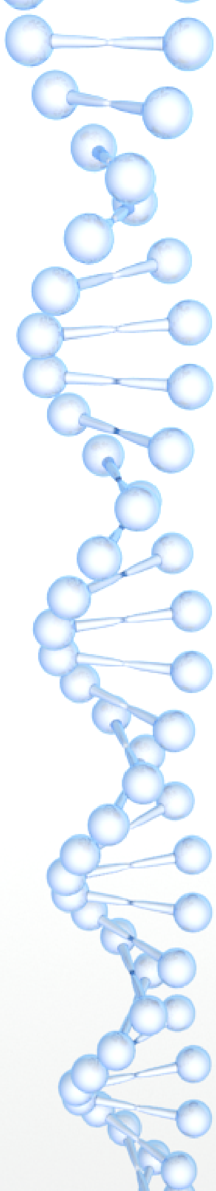
- I go because I have to. Stay here!
  I go because I have to stay here!

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

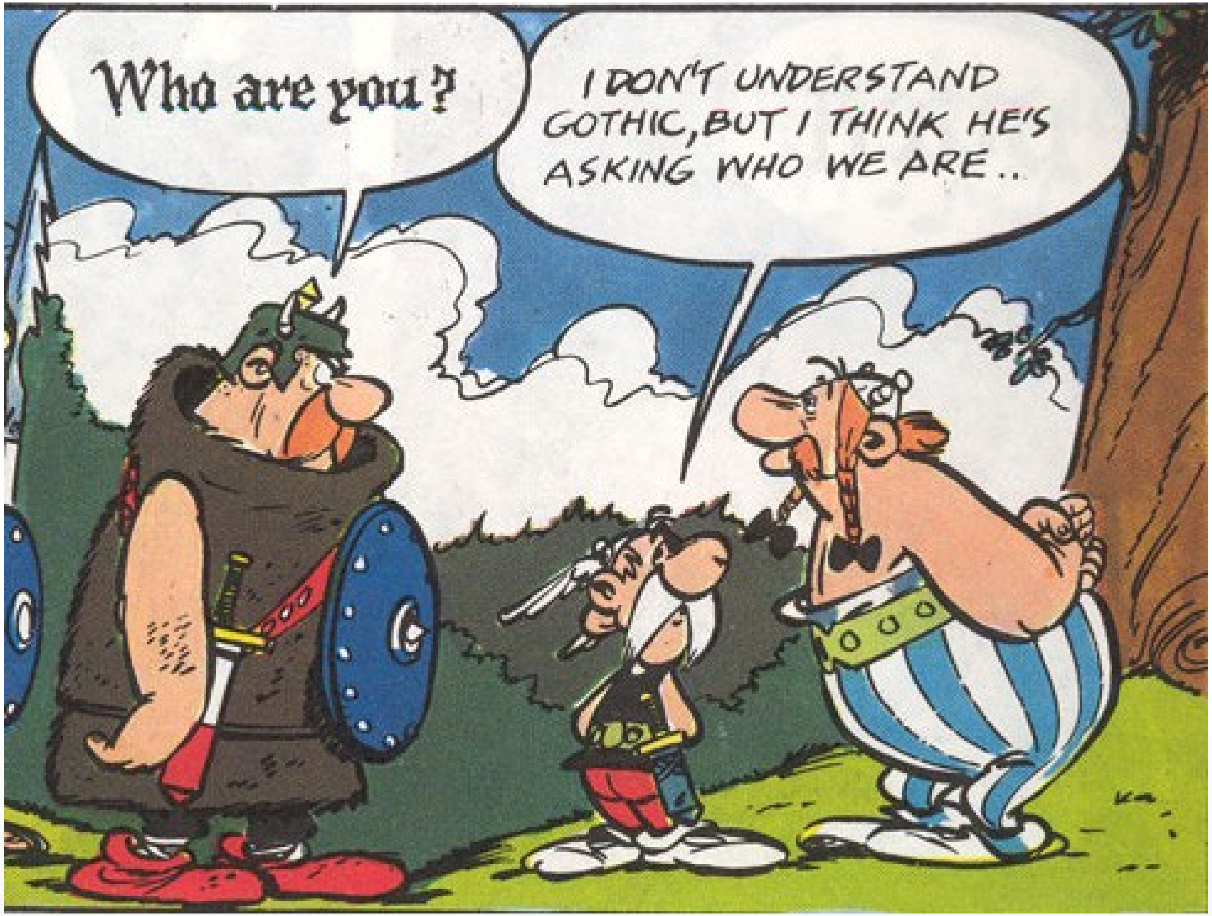
- σαῦρος, ſucceſs, daſs (daß)

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

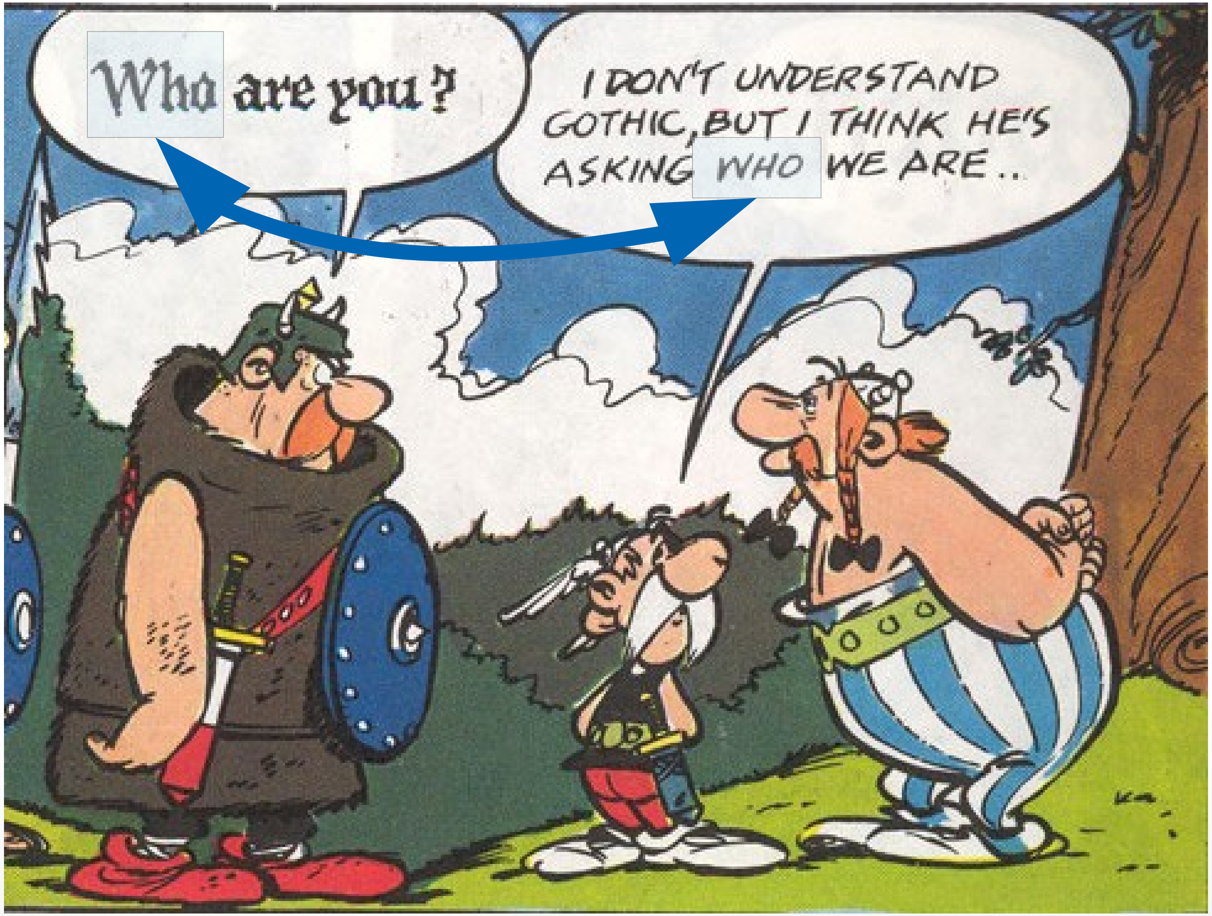
Positional allograph

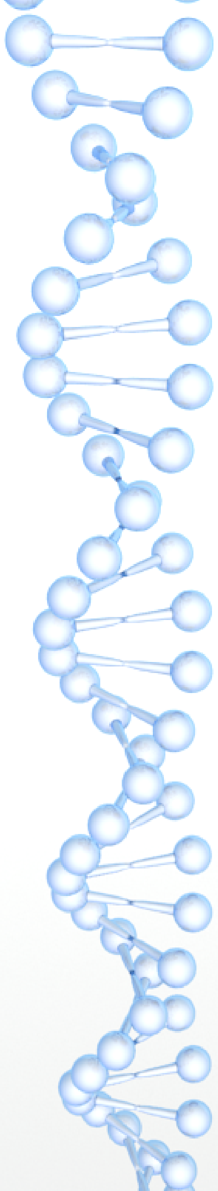
Space

Connotators

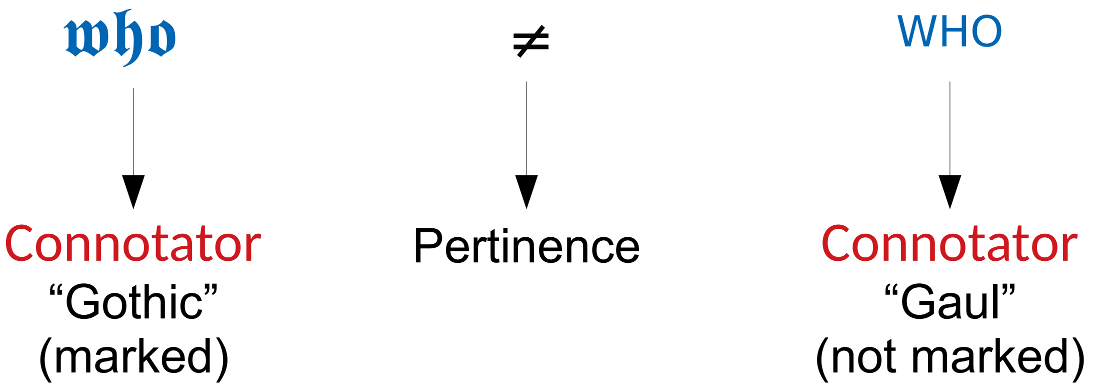

Who are you?
I DON'T UNDERSTAND GOTHIC, BUT I THINK HE'S ASKING WHO WE ARE..

Connotators


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



Connotators



who

Connotator

“Gothic”

(marked)

≠

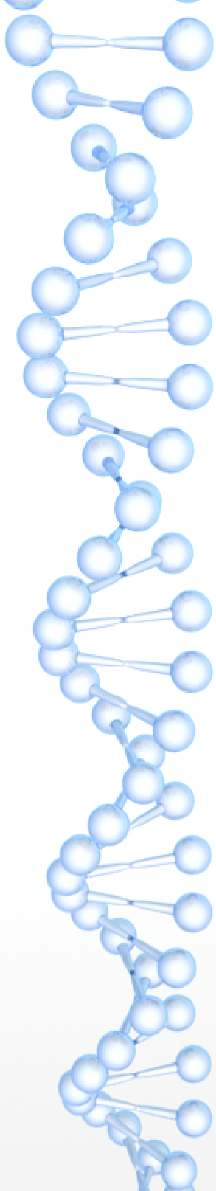
Pertinence

WHO

Connotator

“Gaul”

(not marked)



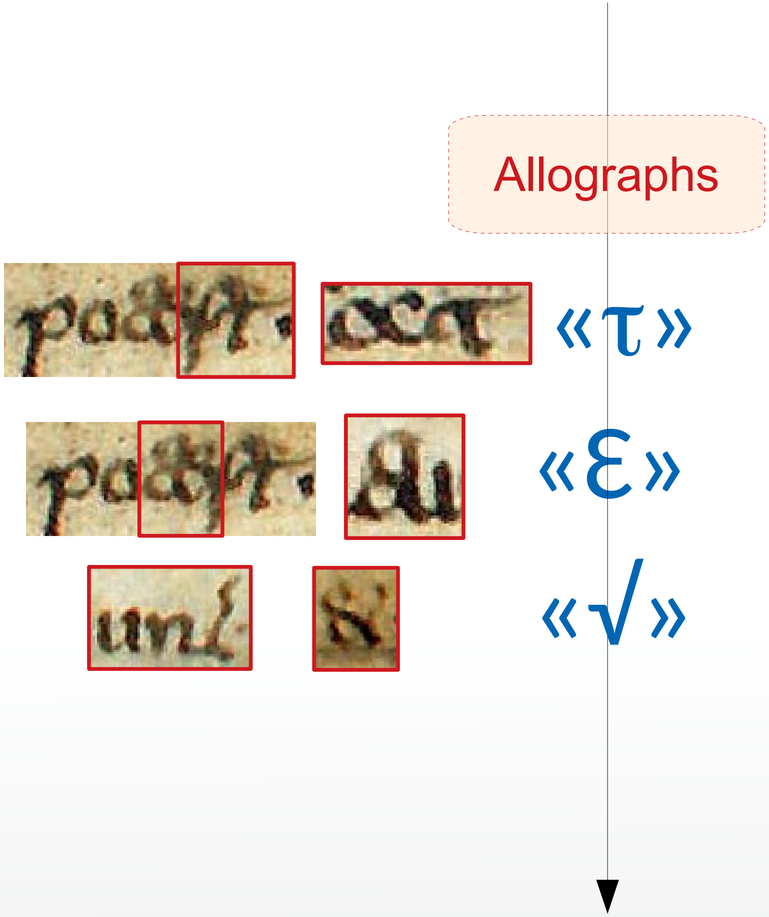
Connotators

Connotators, pertinent for the writer

• I *really* like it	*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, ſ / s)

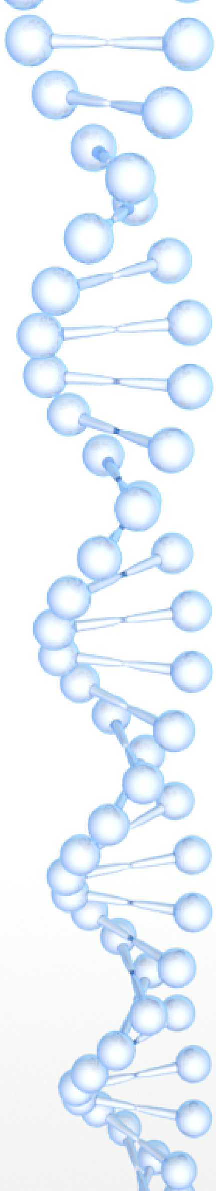


Allographs

«τ»

«ε»

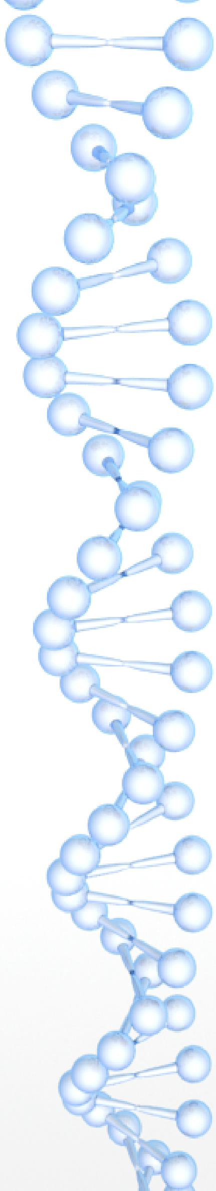
«√»



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?

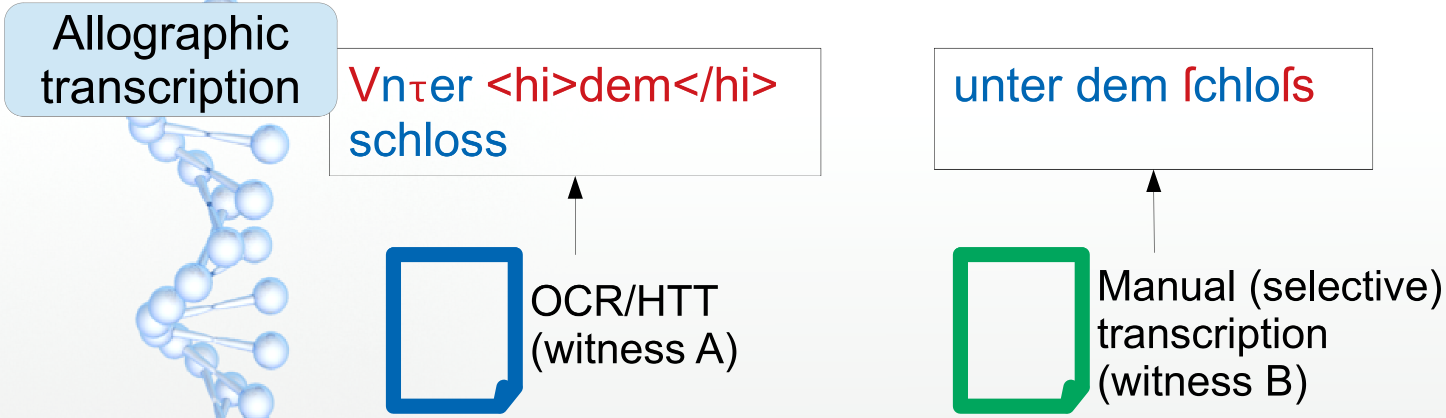


In practice: how can grapheme/allograph modelling make my DSE more interoperable?

OCR/HTT
(witness A)

Manual (selective)
transcription
(witness B)

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



Allographic transcription

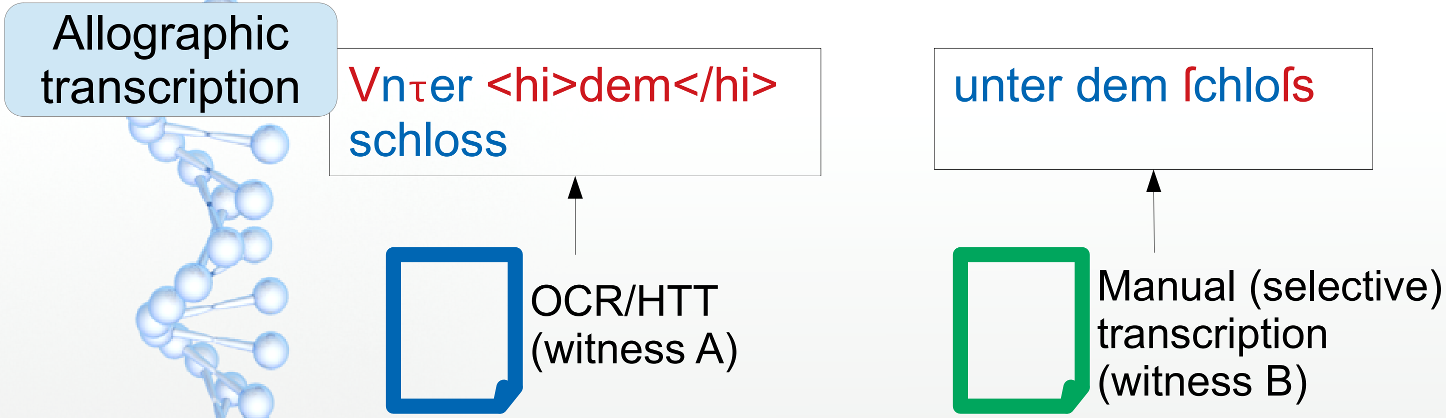
Vnτer <hi>dem</hi> schloss

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

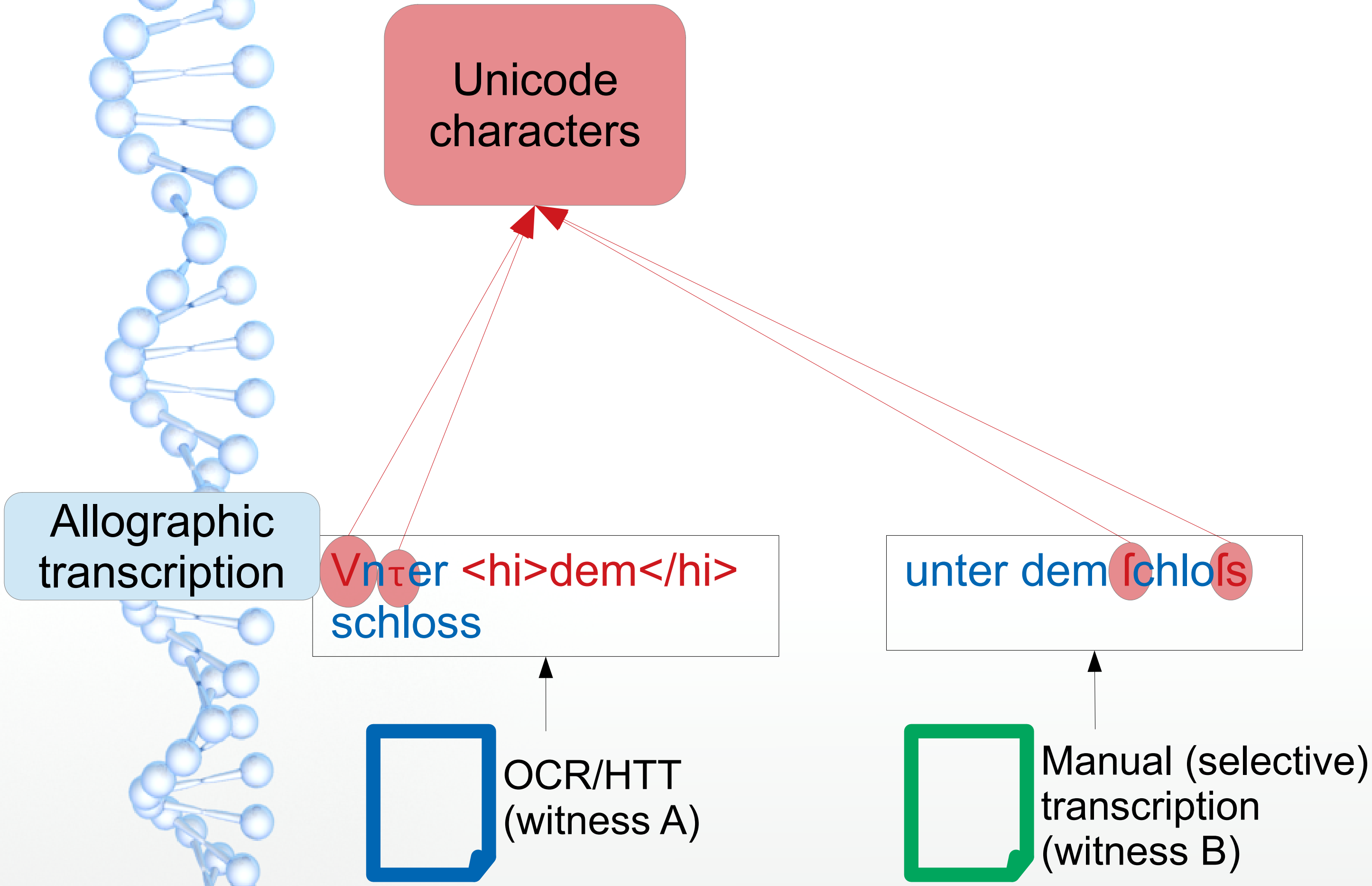
Vnτer <hi>dem</hi> schloss

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

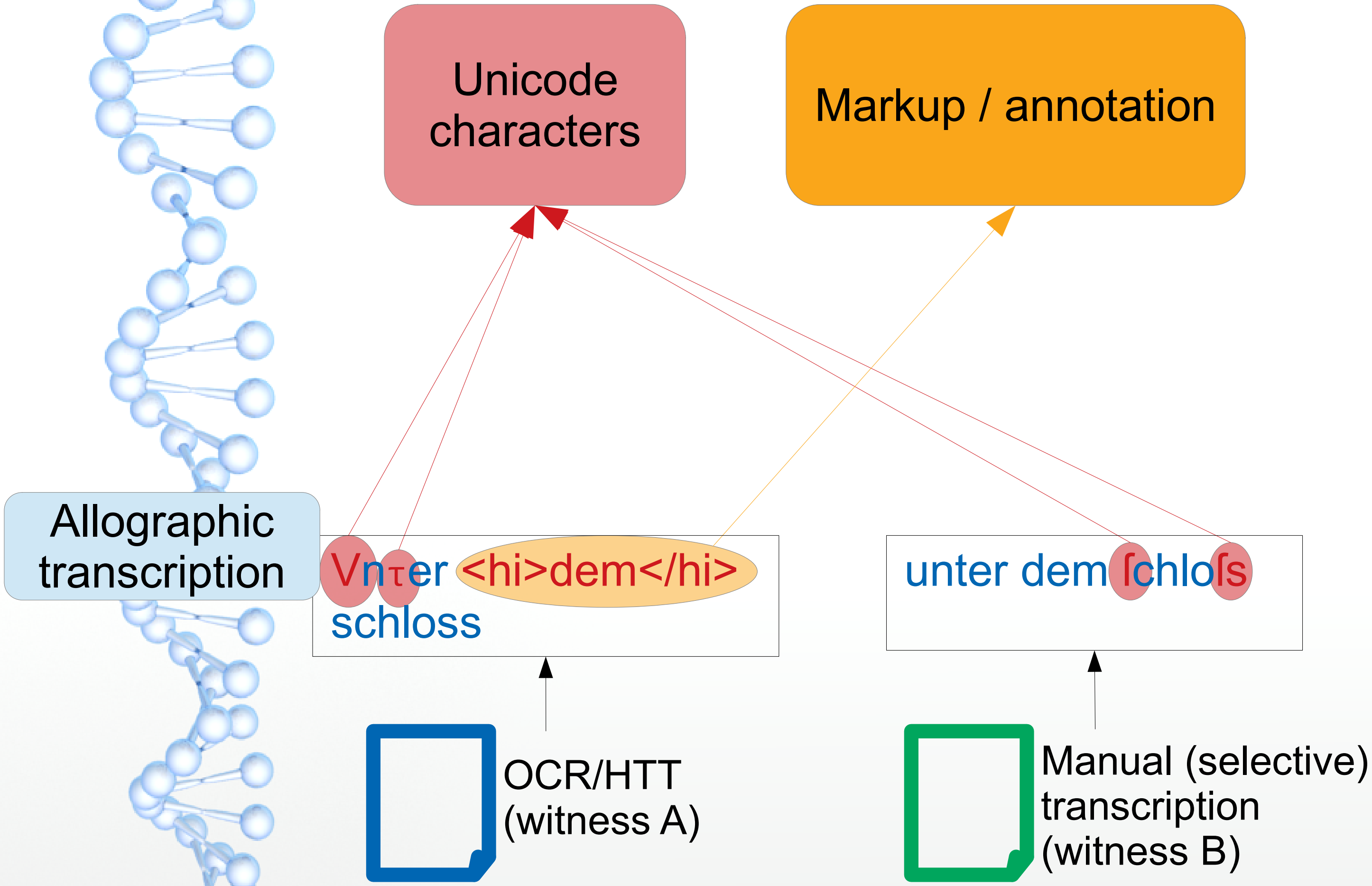
Vnτer <hi>dem</hi>
schloss

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

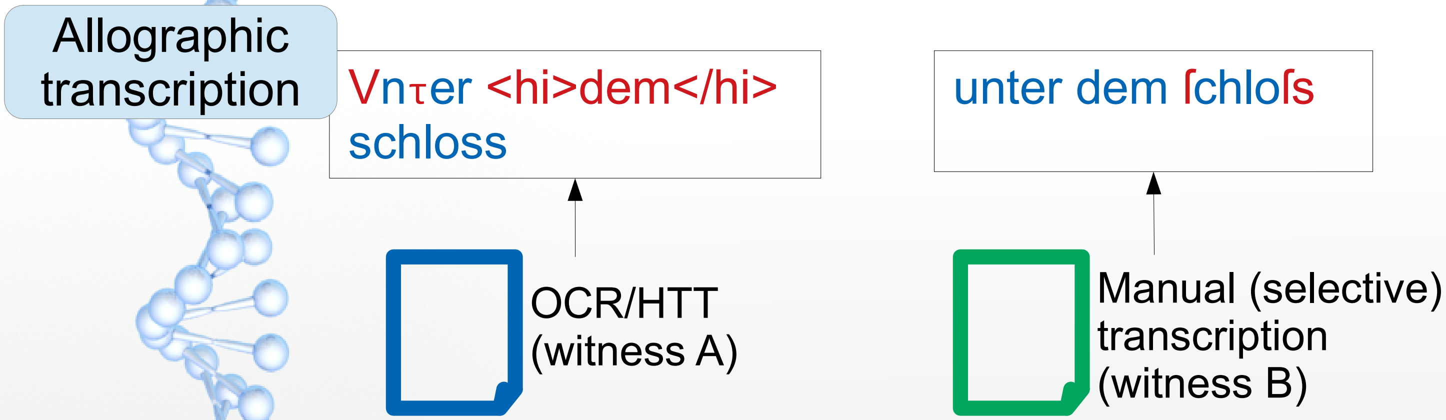
Vnτer <hi>dem</hi> schloss

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

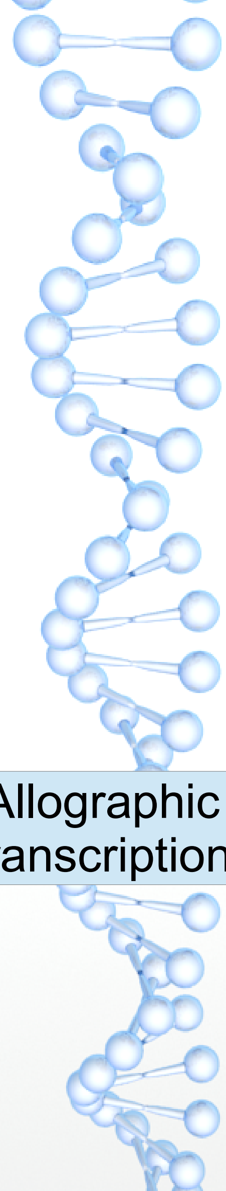
Vnτer <hi>dem</hi> schloss

unter dem ſchloſs

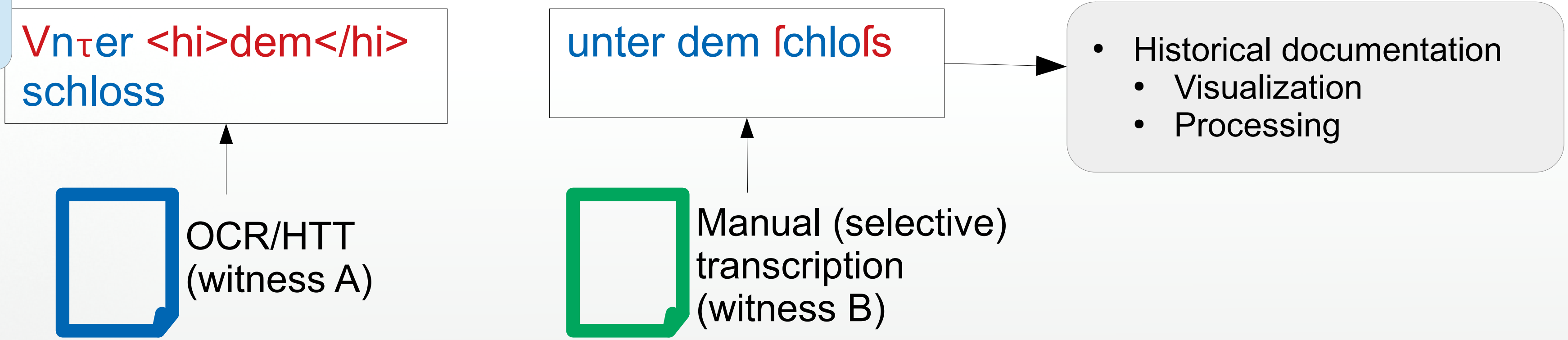
OCR/HTT (witness A)

Manual (selective) transcription (witness B)

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



Allographic transcription



Vnτer <hi>dem</hi> schloss

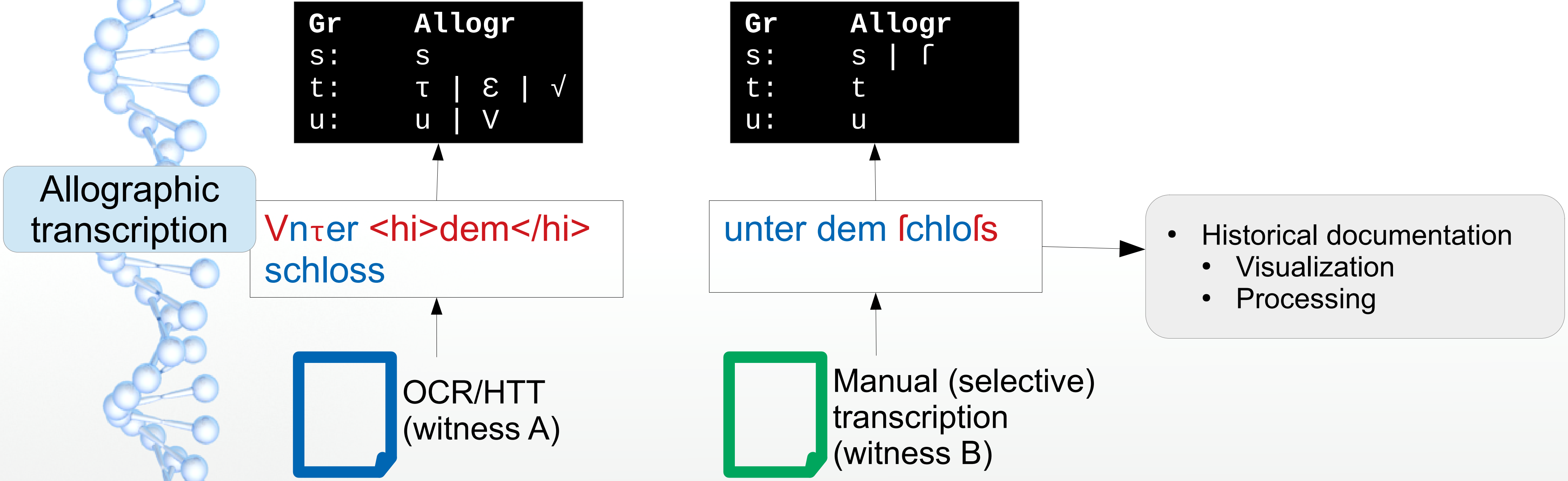
OCR/HTT (witness A)

unter dem ſchloſs

Manual (selective) transcription (witness B)

- Historical documentation
  - Visualization
  - Processing

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



Gr	Allogr
s:	s
t:	τ | ε | √
u:	u | V

Gr	Allogr
s:	s | ſ
t:	t
u:	u

Allographic transcription

Vnτer <hi>dem</hi> schloss

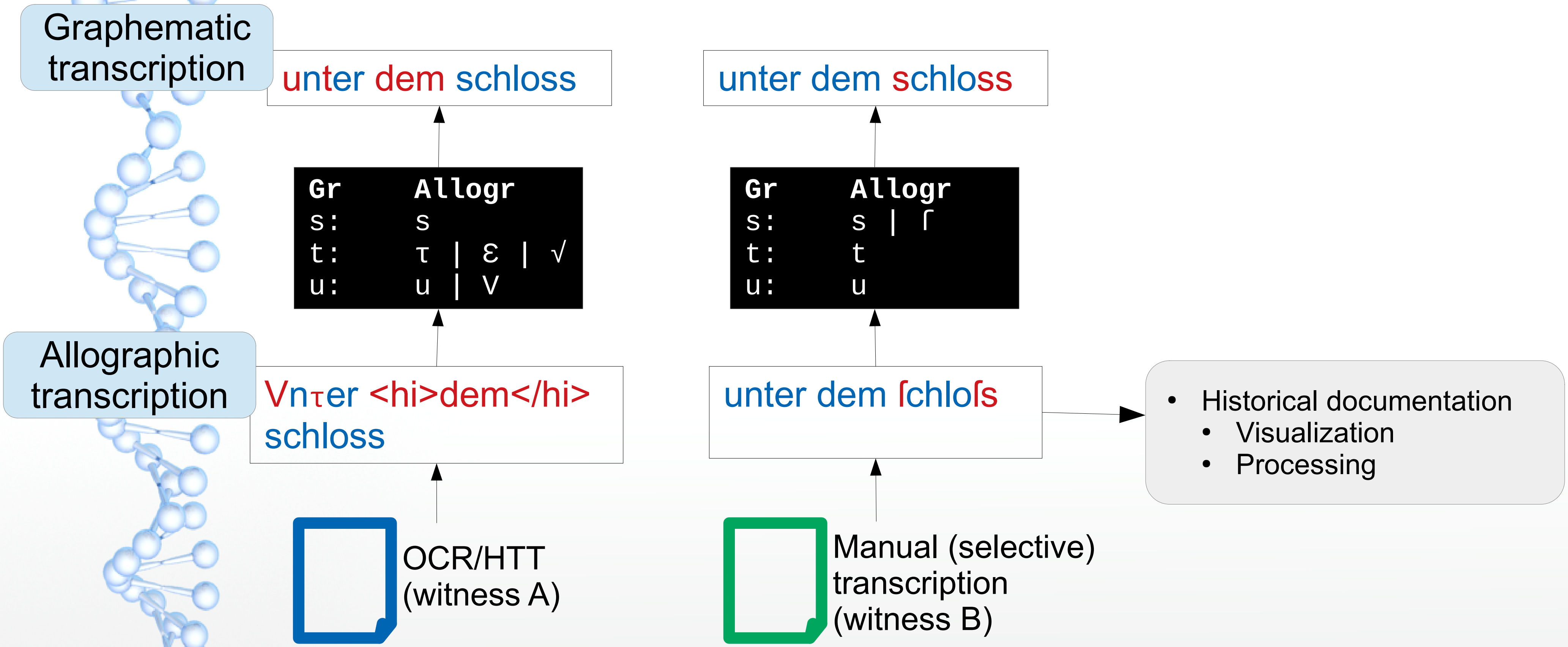
unter dem ſchloſs

OCR/HTT (witness A)

Manual (selective) transcription (witness B)

- Historical documentation
  - Visualization
  - Processing

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



Graphematic transcription

Allographic transcription

unter dem schloss

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

Vnτer <hi>dem</hi> schloss

OCR/HTT (witness A)

unter dem schloss

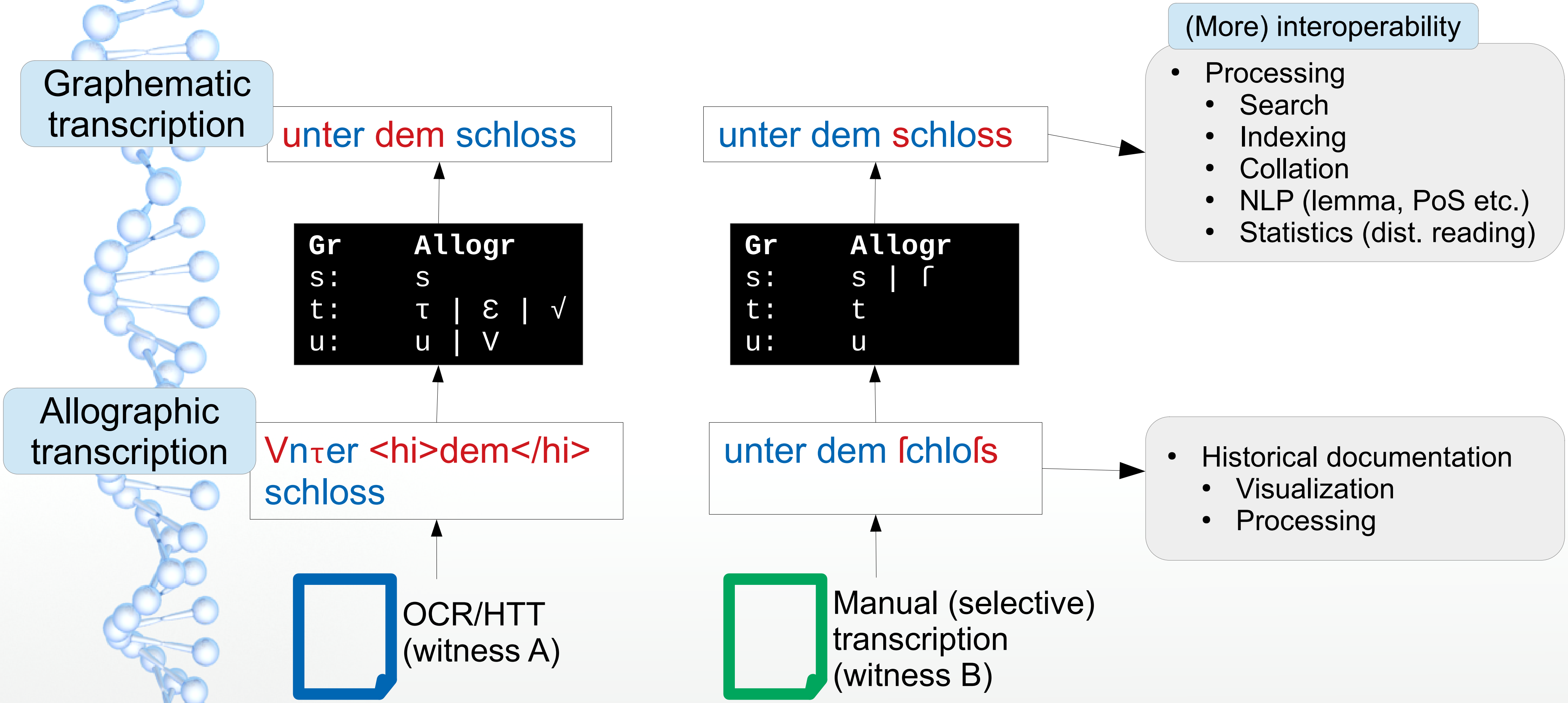
```
Gr      Allogr
s:      s | ſ
t:      t
u:      u
```

unter dem ſchloſs

Manual (selective) transcription (witness B)

- Historical documentation
  - Visualization
  - Processing

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



Graphematic transcription

Allographic transcription

unter dem schloss

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

Vnτer <hi>dem</hi> schloss

OCR/HTT (witness A)

unter dem schloss

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

unter dem ſchloſs

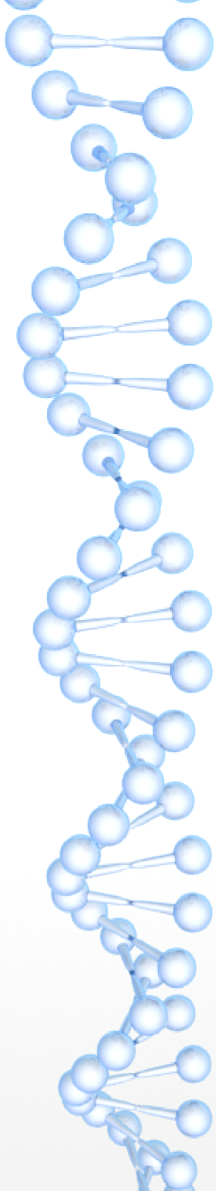
Manual (selective) transcription (witness B)

(More) interoperability

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

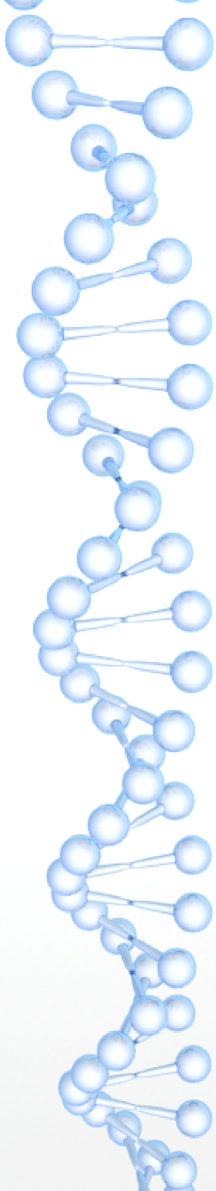
- Historical documentation
  - Visualization
  - Processing

Open issues



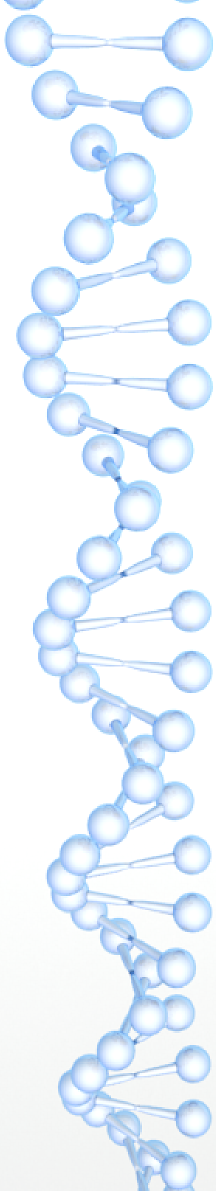
Open issues

- Allographic words
  - Spelling (wife / wyffe)



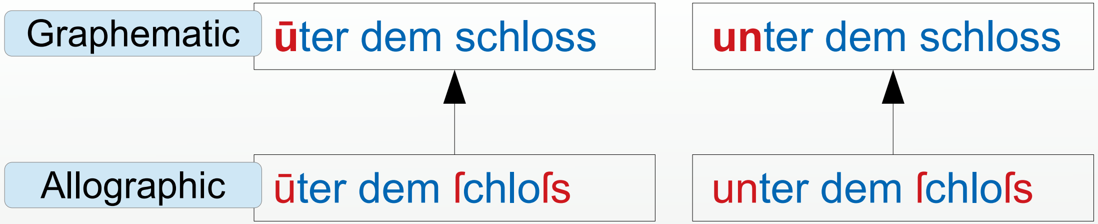
Open issues

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



Open issues

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



Graphematic

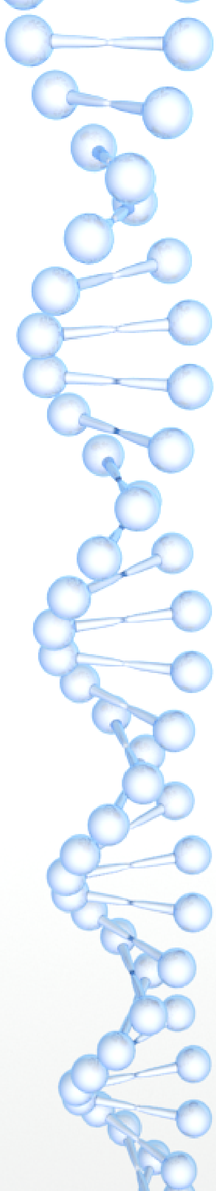
ūter dem schloss

unter dem schloss

Allographic

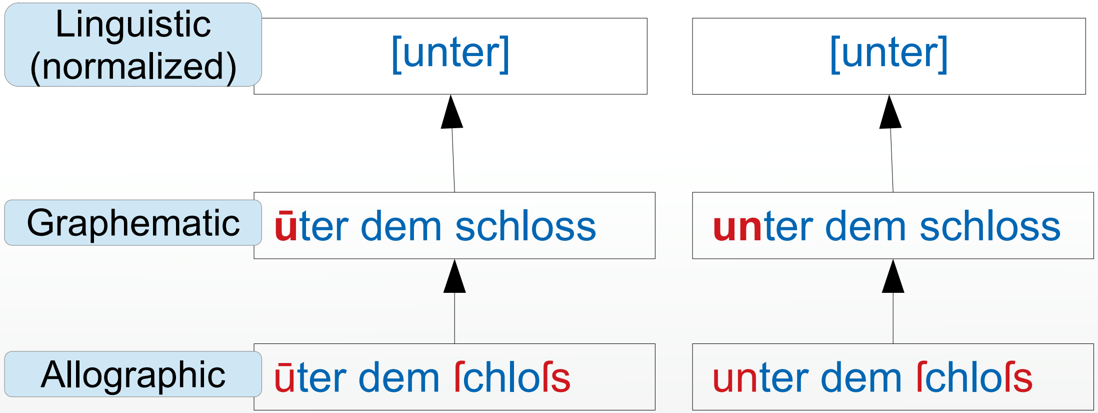
ūter dem ſchloſs

unter dem ſchloſs



Open issues

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



Linguistic (normalized)

[unter]

[unter]

Graphematic

ūter dem schloss

unter dem schloss

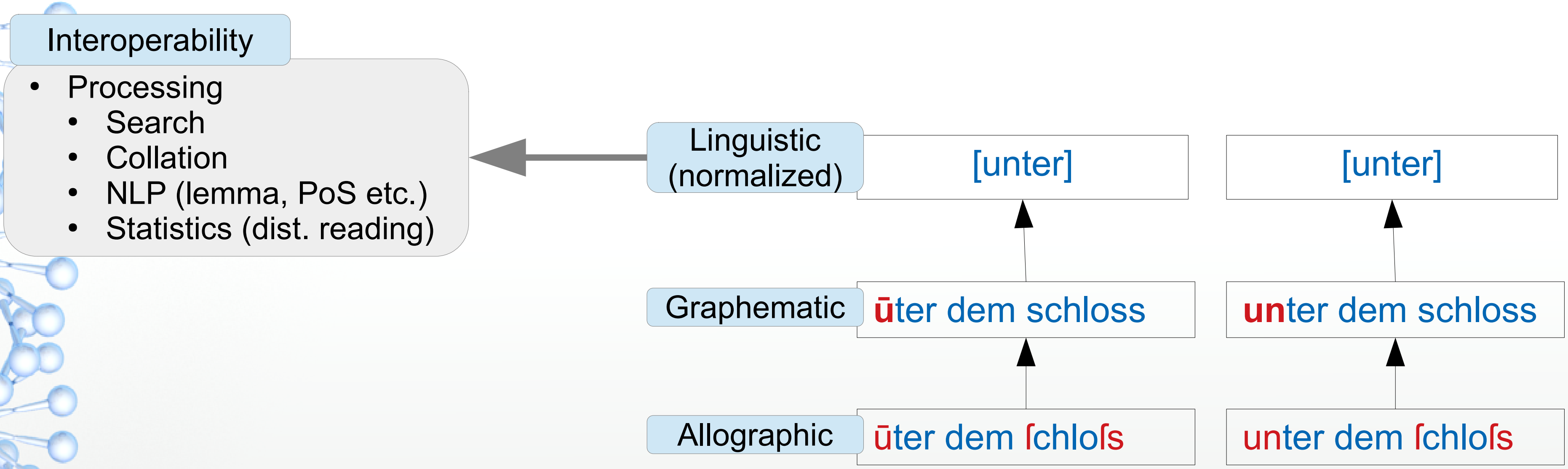
Allographic

ūter dem ſchloſs

unter dem ſchloſs

Open issues

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



Interoperability

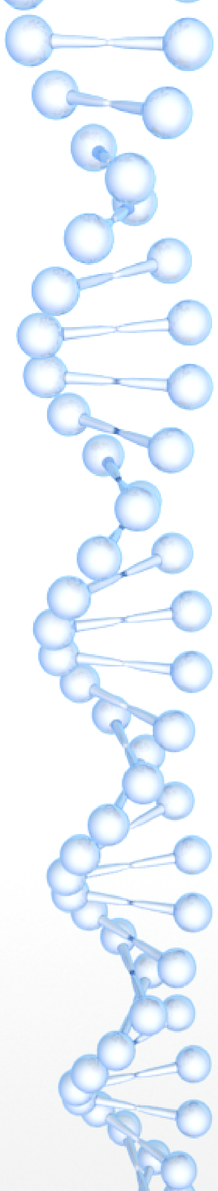
- Processing
  - Search
  - Collation
  - NLP (lemma, PoS etc.)
  - Statistics (dist. reading)

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**