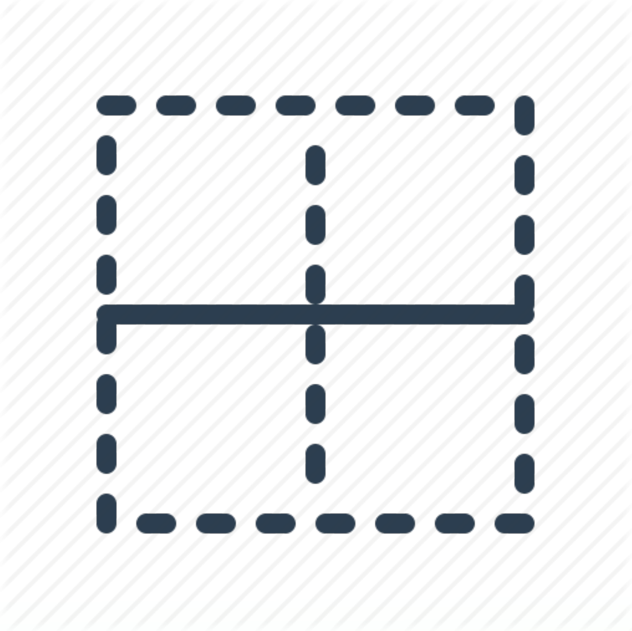
Paolo Monella

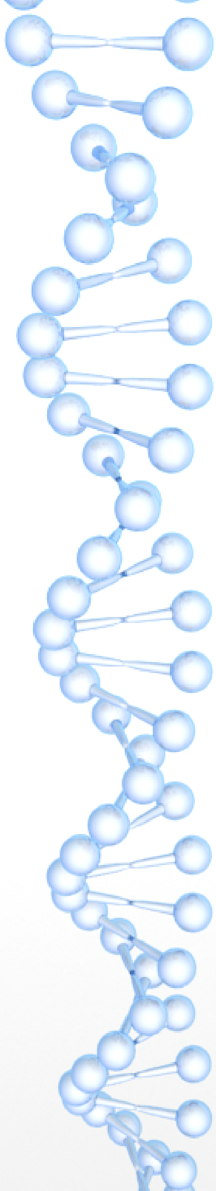
Encoding Pre-modern Writing Systems



SAPIENZA
UNIVERSITÀ DI ROMA

Sunoikisis Digital Classics,
Summer 2022 programme
May 12, 2022

European Research Council
Established by the European Commission



erc

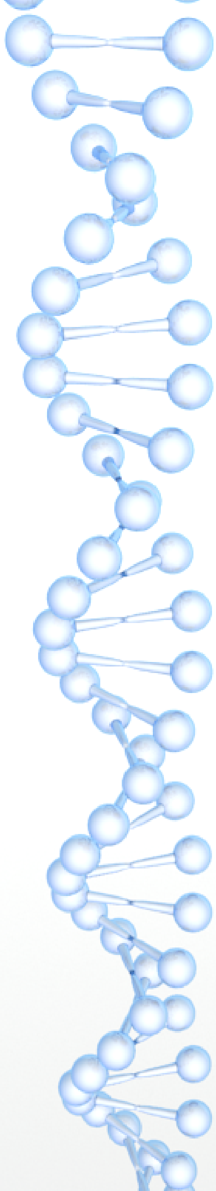
European Research Council

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

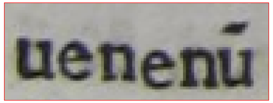


Outline

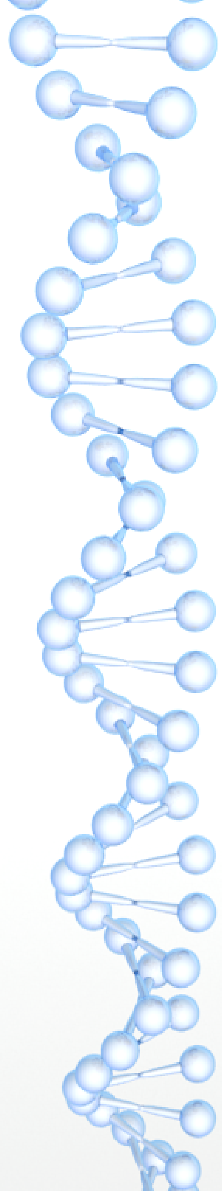
- **Interoperability**
  of digital scholarly editions (DSEs)
  based on diplomatic transcriptions
  - The issue
  - Current solutions
  - Interoperability through modelling
- Orlandi’s **table of sign**
- **Graphemes/allographs**

Interoperability: the issue

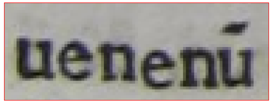
Interoperability: the issue



uenenú

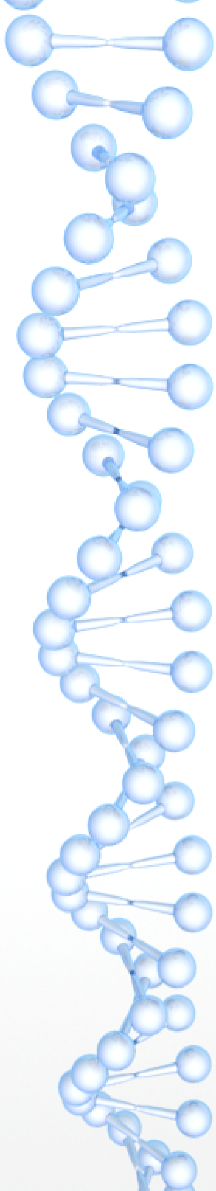


Interoperability: the issue

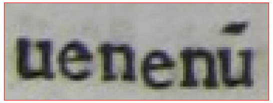


uenenū

- uenenū



Interoperability: the issue



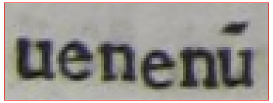
uenenū

- uenenū

Diplomatic

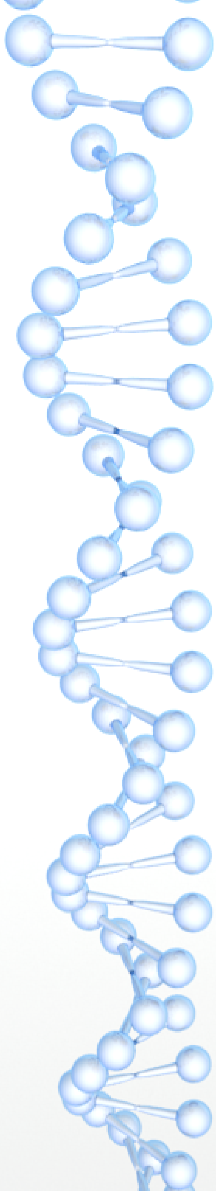
- Manual or HTR
  - Visualization
  - Processing

Interoperability: the issue

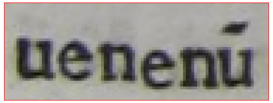


uenenú

- uenenū



Interoperability: the issue



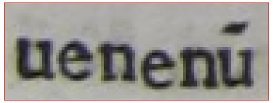
uenenū

- uenenū

- uenenum

Interoperability: the issue

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



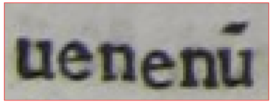
uenenū

- uenenū

- uenenum

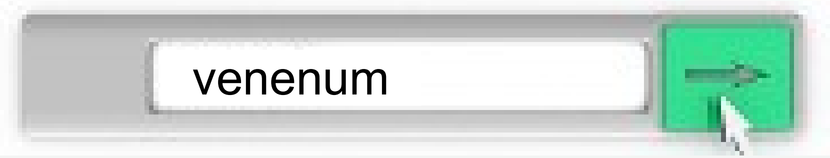
Interoperability: the issue

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



- uenenū

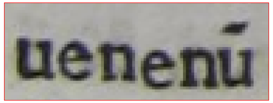
- uenenum



venenum

Interoperability: the issue

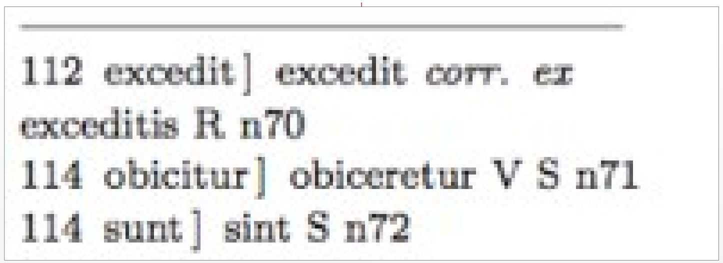
- Processing
  - Search
  - **Collation**
  - NLP (lemma, PoS etc.)
  - Statistics (dist. reading)



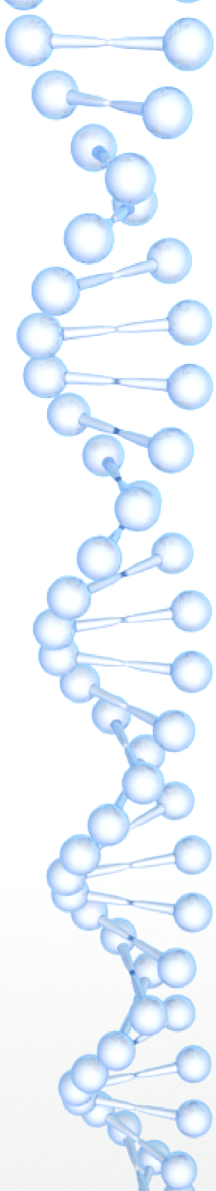
uenenú

- uenenū

- uenenum

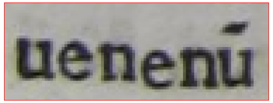


112 excedit] excedit *corr. ex* exceditis R n70
114 obicitur] obiceretur V S n71
114 sunt] sint S n72



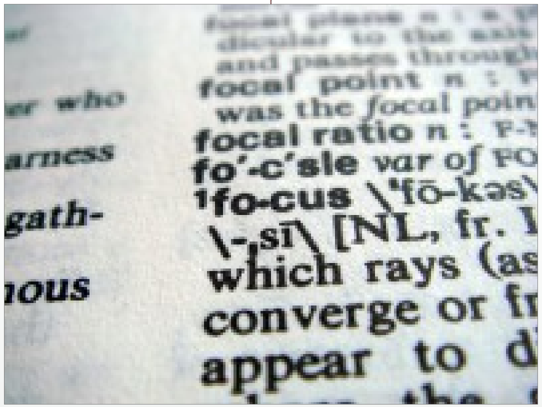
Interoperability: the issue

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



uenenú

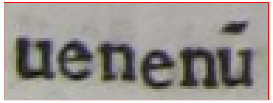
- uenenū



- uenenum

Interoperability: the issue

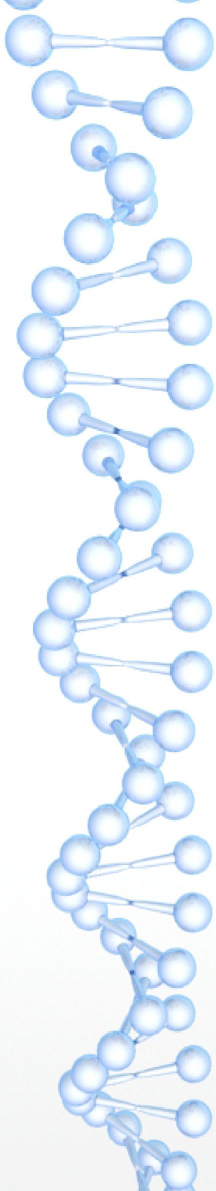
- Processing
  - Search
  - Collation
  - NLP (lemma, PoS etc.)
  - **Statistics (dist. reading)**



uenenú

- uenenū

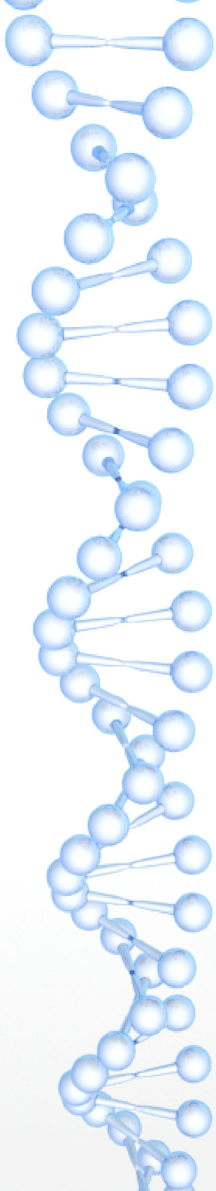
- uenenum



Interoperability: the issue

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

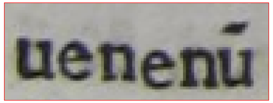
Interoperability: current solutions



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>)?
  - 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>?

Manual normalized transcription

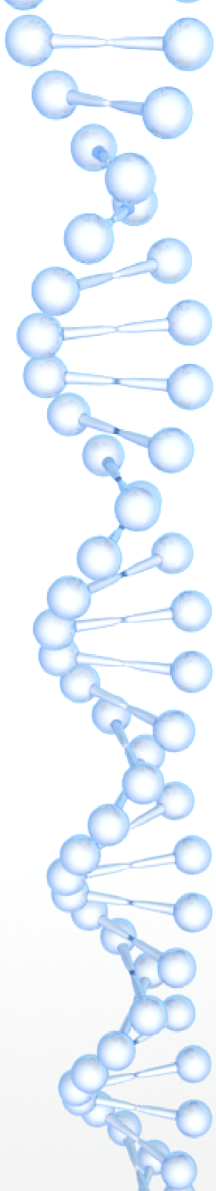


uenenú

- uenenū

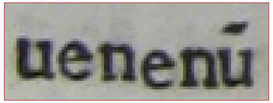
Diplomatic

- Visualization
- ~~Processing~~



Manual normalized transcription

- venenum



uenenú

- uenenū

Normalized

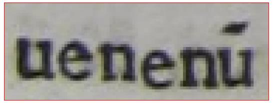
- Processing
  - Search
  - Indexing
  - Collation
  - NLP (lemma, PoS etc.)
  - Statistics (distant reading)...

Diplomatic

- Visualization
- ~~Processing~~

Manual normalized transcription

- venenum



uenenú

- uenenū

Not generated by computer: repeated manual transcription

Normalized

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

Diplomatic

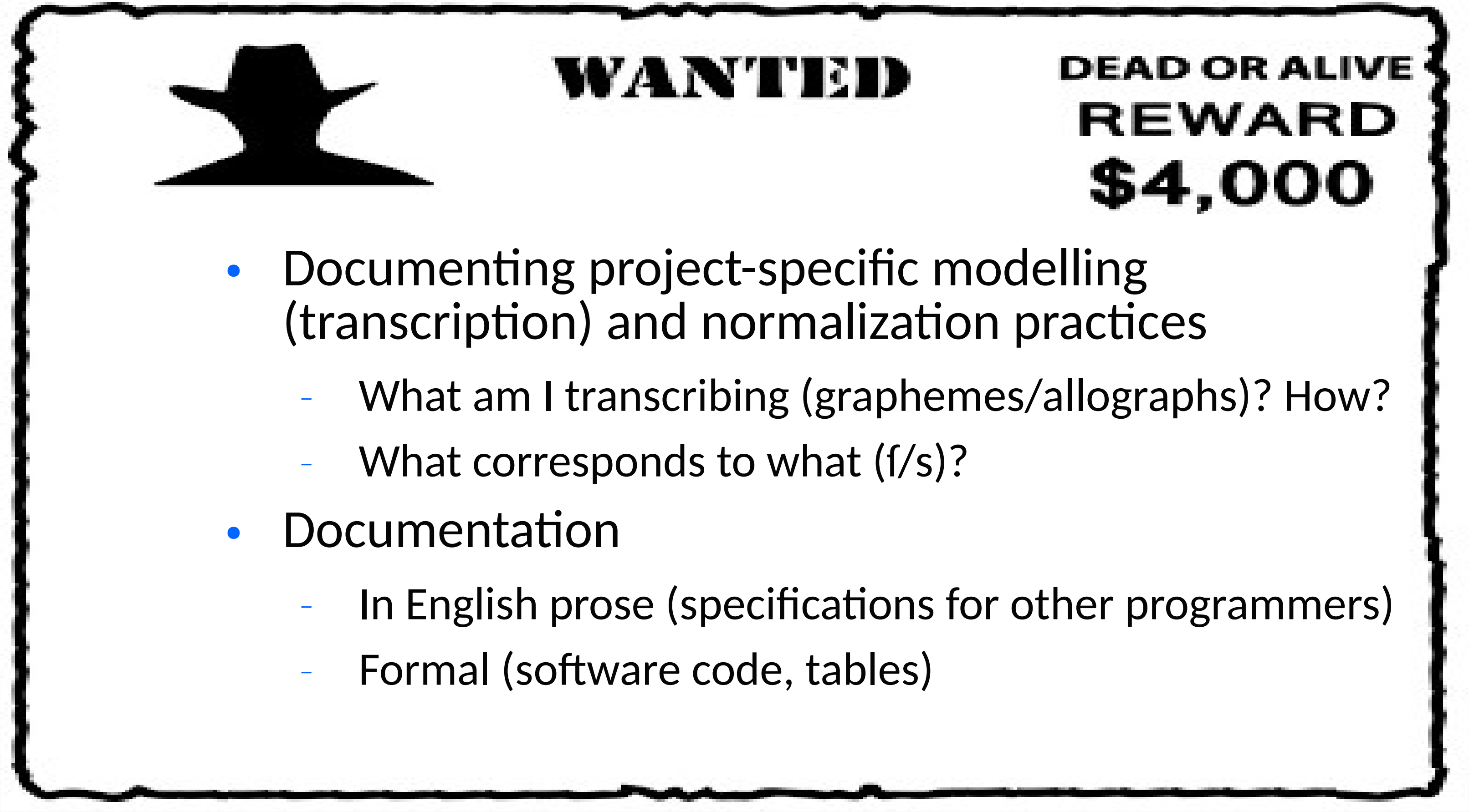
- Visualization
- ~~Processing~~

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

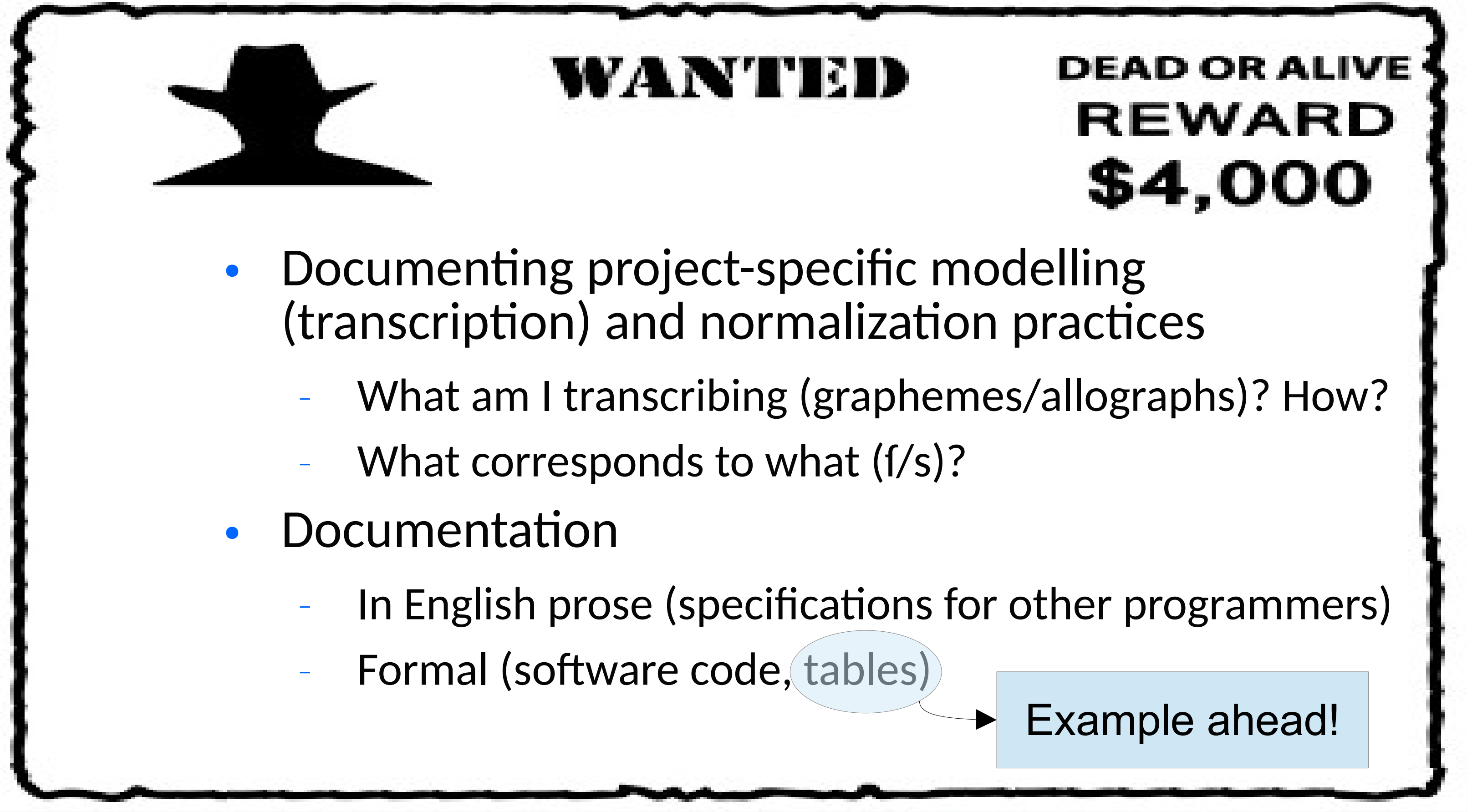
Interoperability through modelling

Interoperability through modelling



- 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)

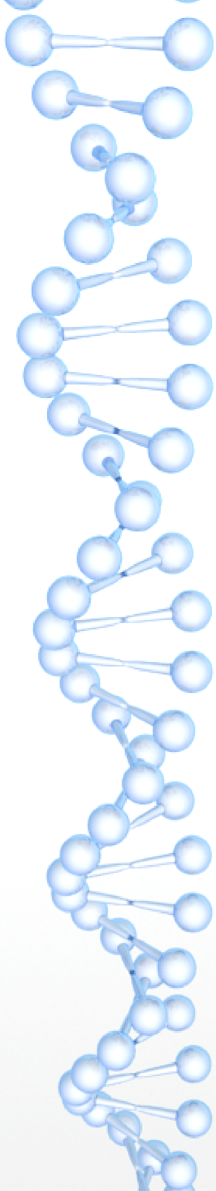
Interoperability through modelling



- 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!

Orlandi’s table of signs

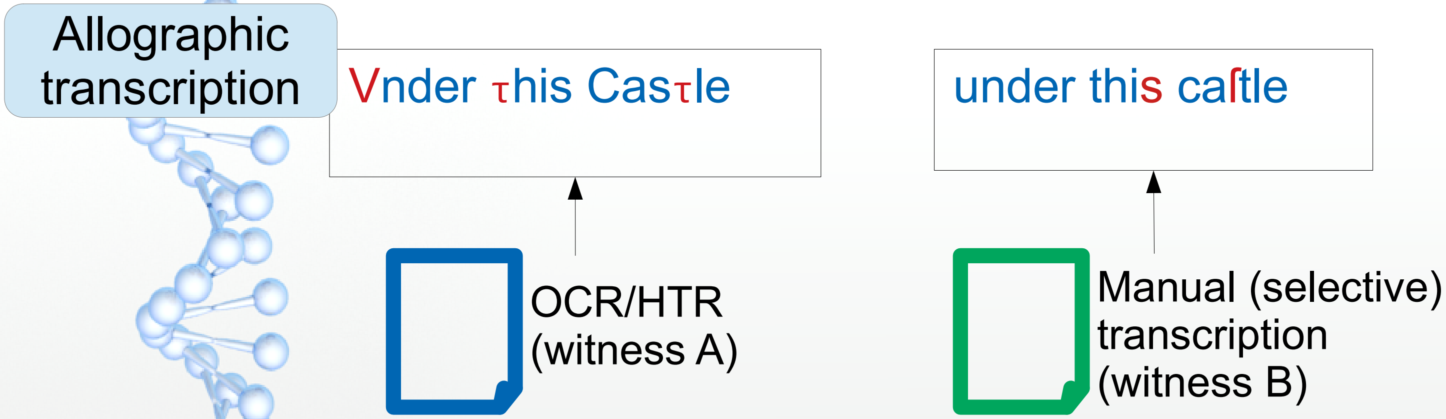


Orlandi’s table of signs

🟦 OCR/HTR
(witness A)

🟩 Manual (selective)
transcription
(witness B)

Orlandi’s table of signs



Allographic transcription

Vnder τhis Casτle

under this caſtle

OCR/HTR (witness A)

Manual (selective) transcription (witness B)

Orlandi’s table of signs

Allographic transcription

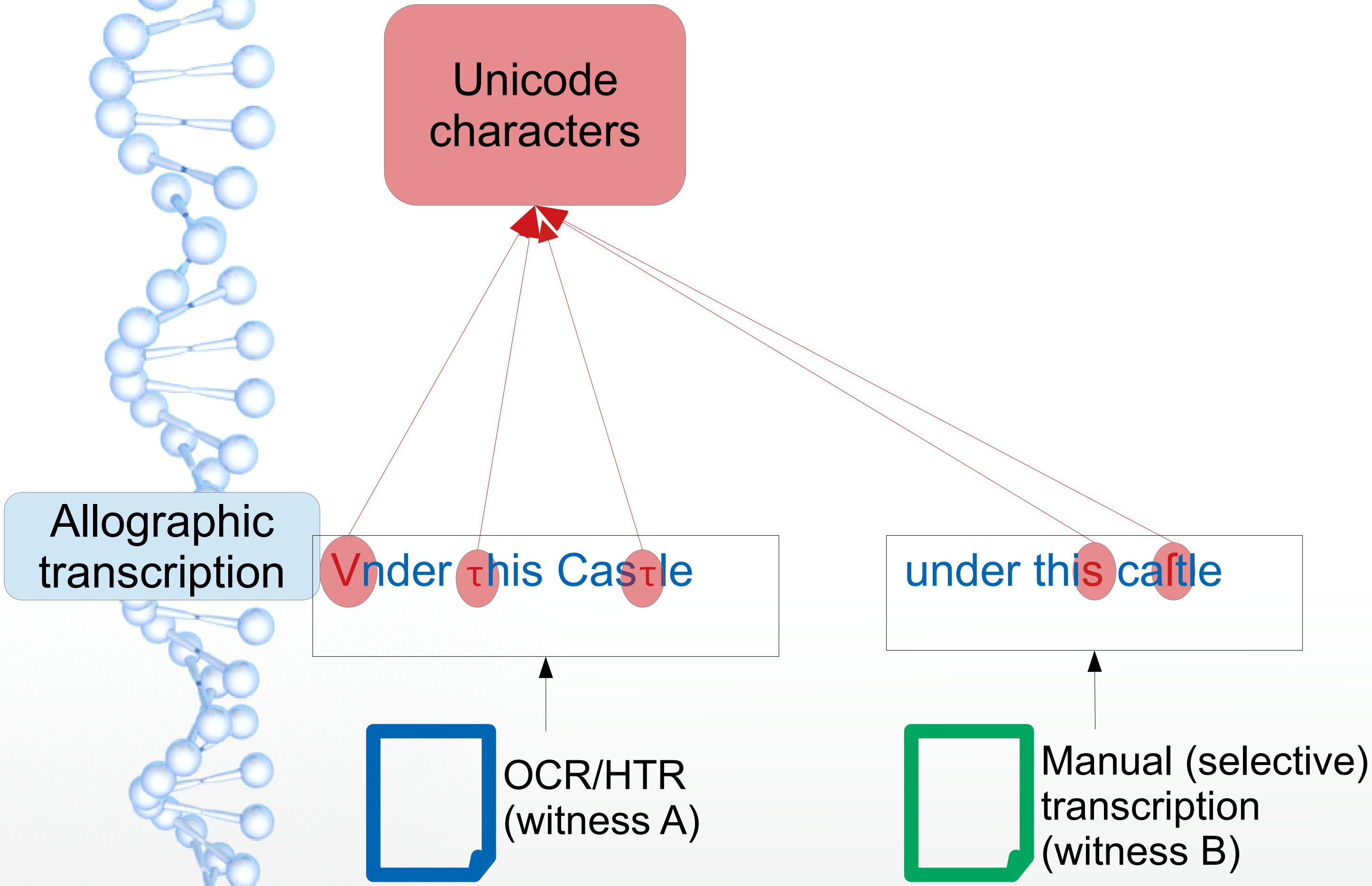
Vnder τhis Casτle

under this caſtle

OCR/HTR (witness A)

Manual (selective) transcription (witness B)

Orlandi’s table of signs



Unicode characters

Allographic transcription

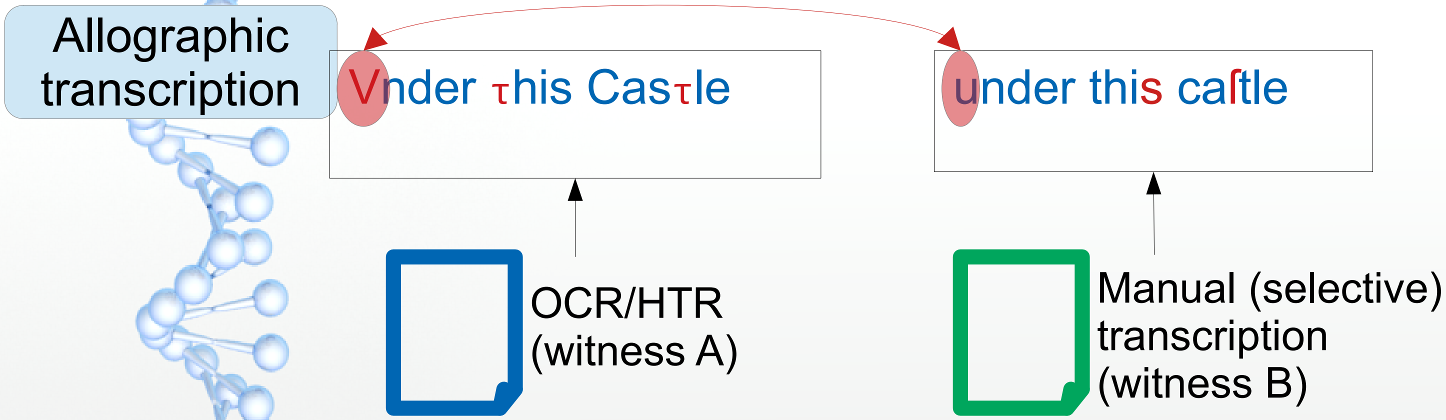
Vnder τhis Casτle

under this caſtle

OCR/HTR (witness A)

Manual (selective) transcription (witness B)

Orlandi’s table of signs



Allographic transcription

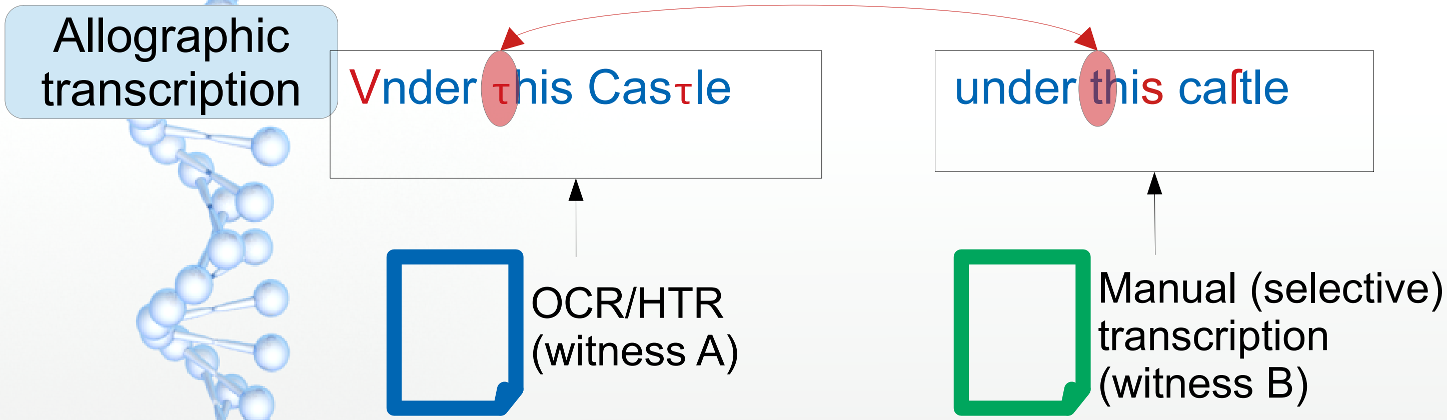
Vnder τhis Casτle

under this caſtle

OCR/HTR (witness A)

Manual (selective) transcription (witness B)

Orlandi’s table of signs



Allographic transcription

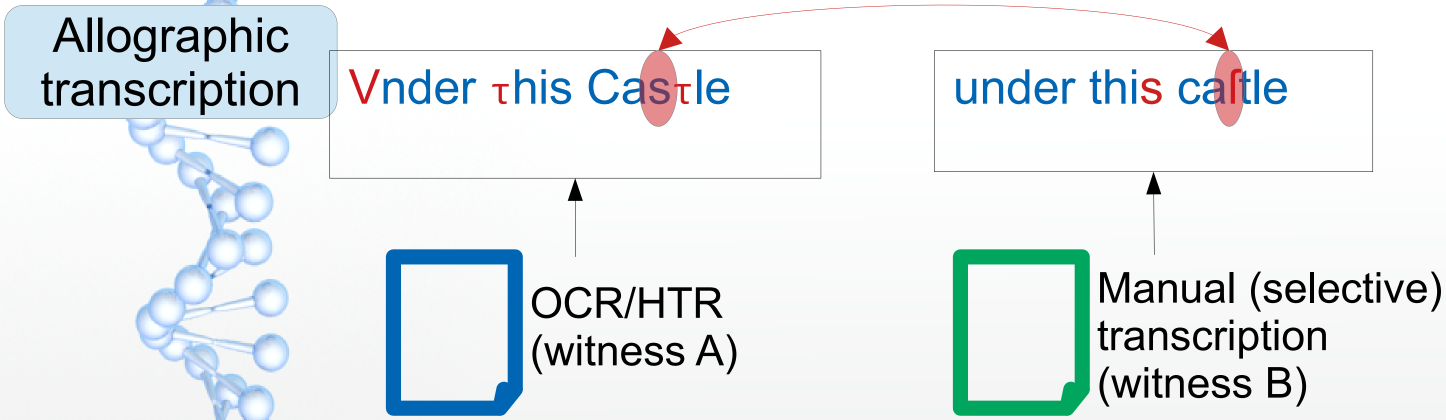
Vnder τhis Casτle

under this caſtle

OCR/HTR (witness A)

Manual (selective) transcription (witness B)

Orlandi’s table of signs



Allographic transcription

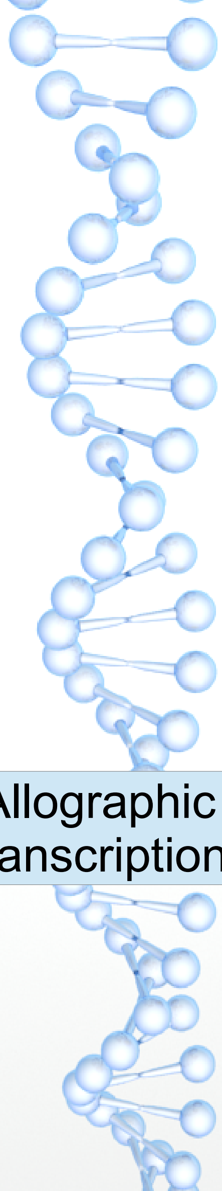
Vnder τhis Casτle

under this caſtle

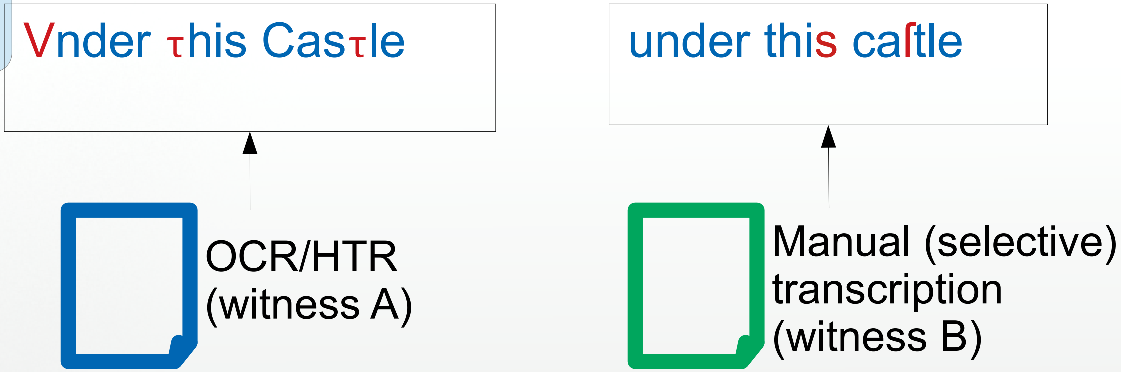
OCR/HTR (witness A)

Manual (selective) transcription (witness B)

Orlandi’s table of signs



Allographic transcription



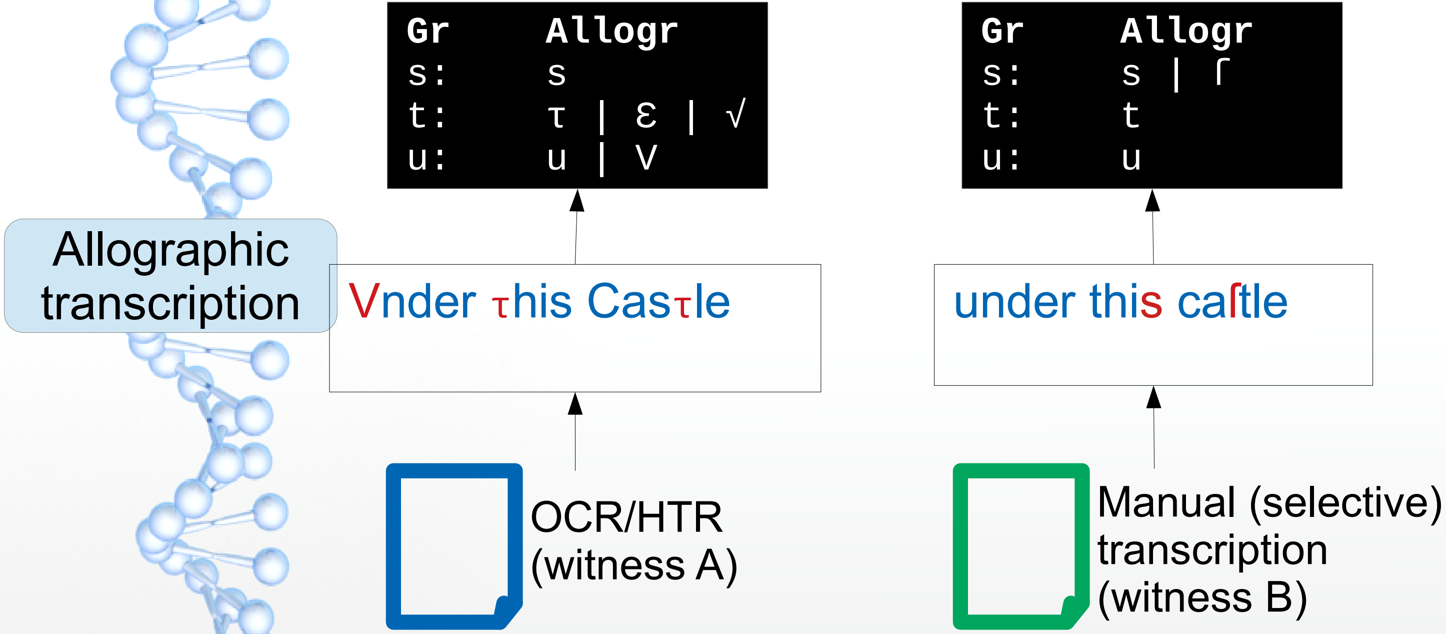
Vnder τhis Casτle

under this caſtle

OCR/HTR (witness A)

Manual (selective) transcription (witness B)

Orlandi’s table of signs



Allographic transcription

Gr	Allogr
s:	s
t:	τ | Ɛ | √
u:	u | V

Vnder τhis Casτle

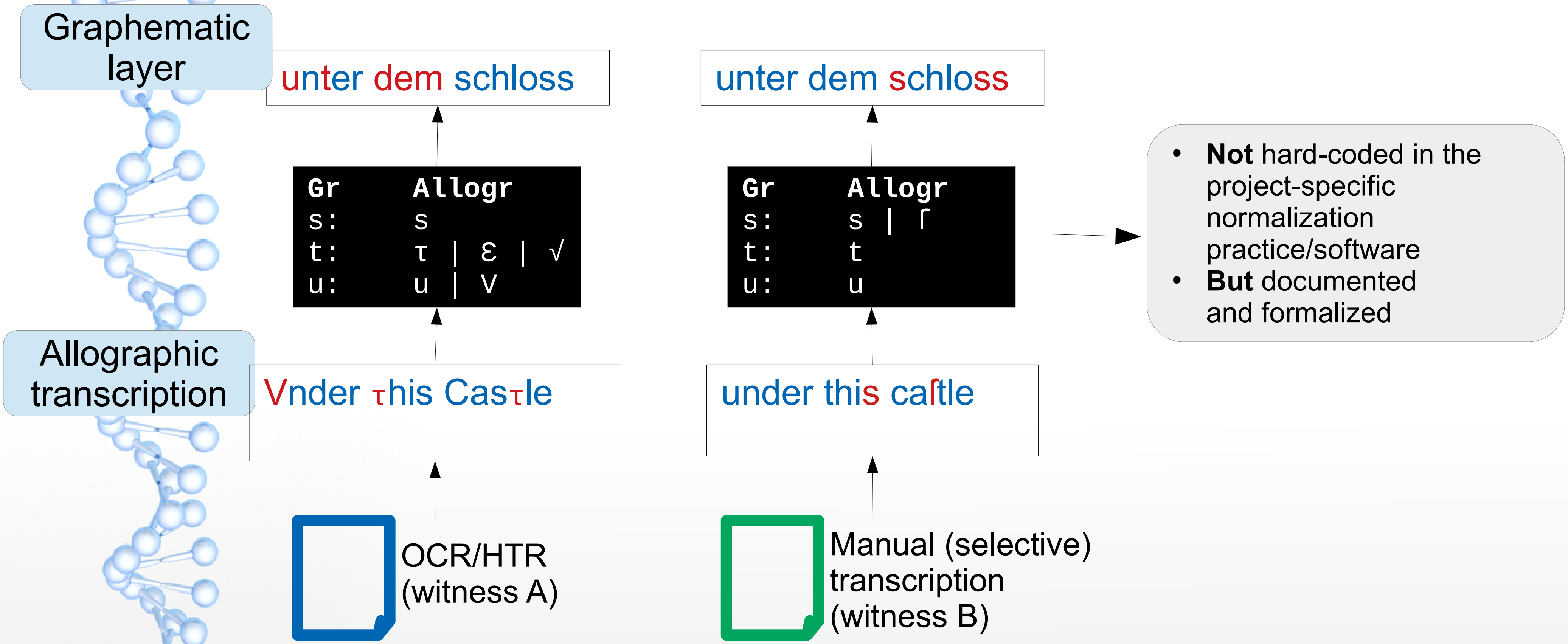
OCR/HTR (witness A)

Gr	Allogr
s:	s | ſ
t:	t
u:	u

under this caſtle

Manual (selective) transcription (witness B)

Orlandi’s table of signs



Graphematic layer

unter dem schloss

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

Vnder τhis Casτle

OCR/HTR (witness A)

unter dem schloss

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

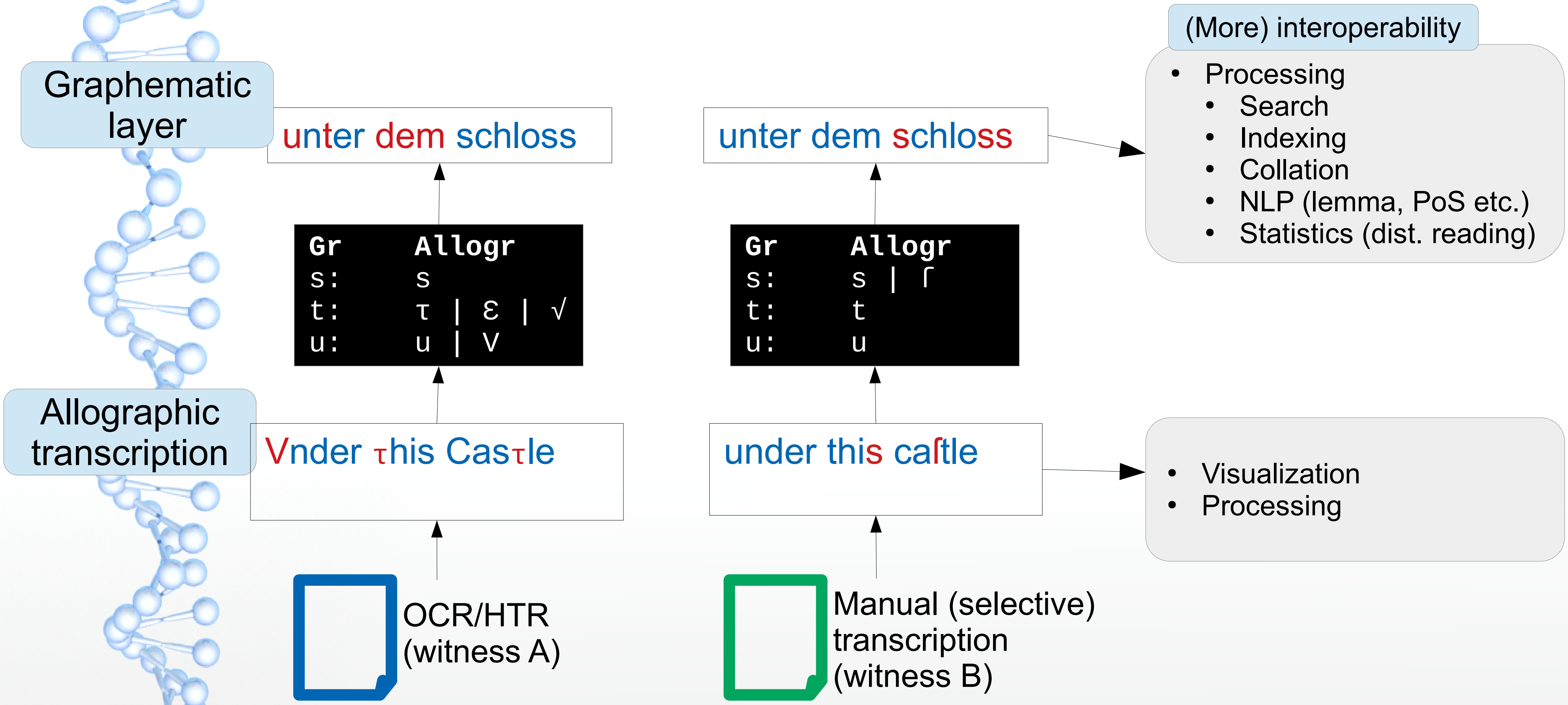
under this caſtle

Manual (selective) transcription (witness B)

Allographic transcription

- **Not** hard-coded in the project-specific normalization practice/software
- **But** documented and formalized

Orlandi’s table of signs



Graphematic layer

Allographic transcription

unter dem schloss

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

Vnder τhis Casτle

OCR/HTR (witness A)

unter dem schloss

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

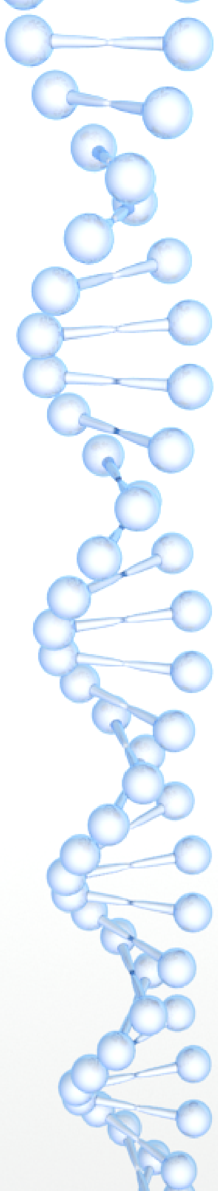
under this caſtle

Manual (selective) transcription (witness B)

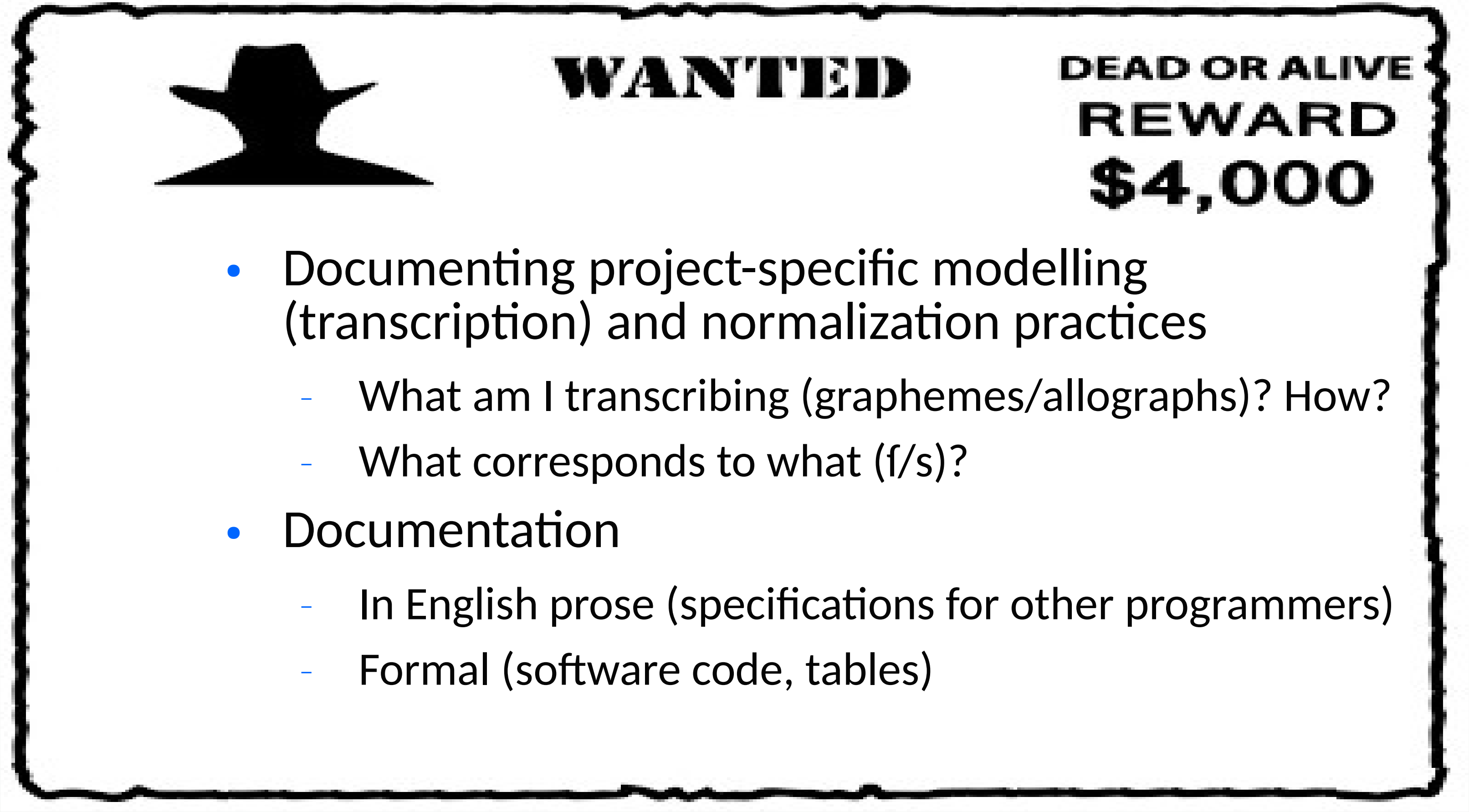
(More) interoperability

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

- Visualization
- Processing

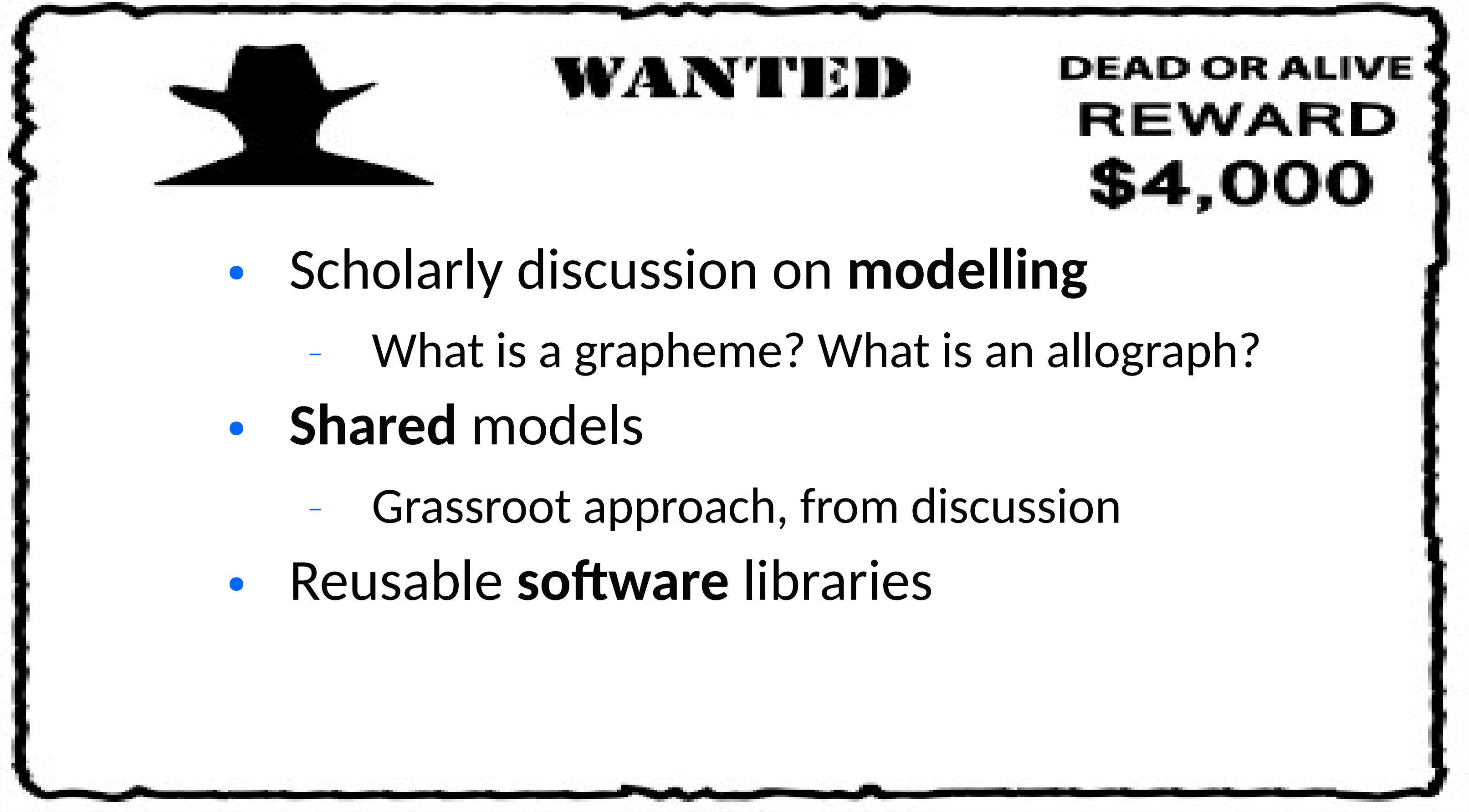


Interoperability through modelling



- 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

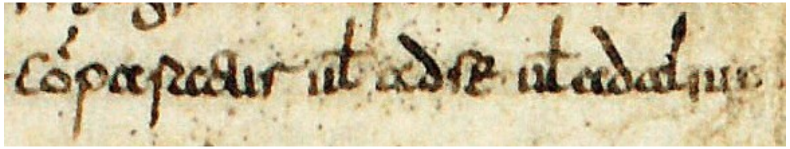


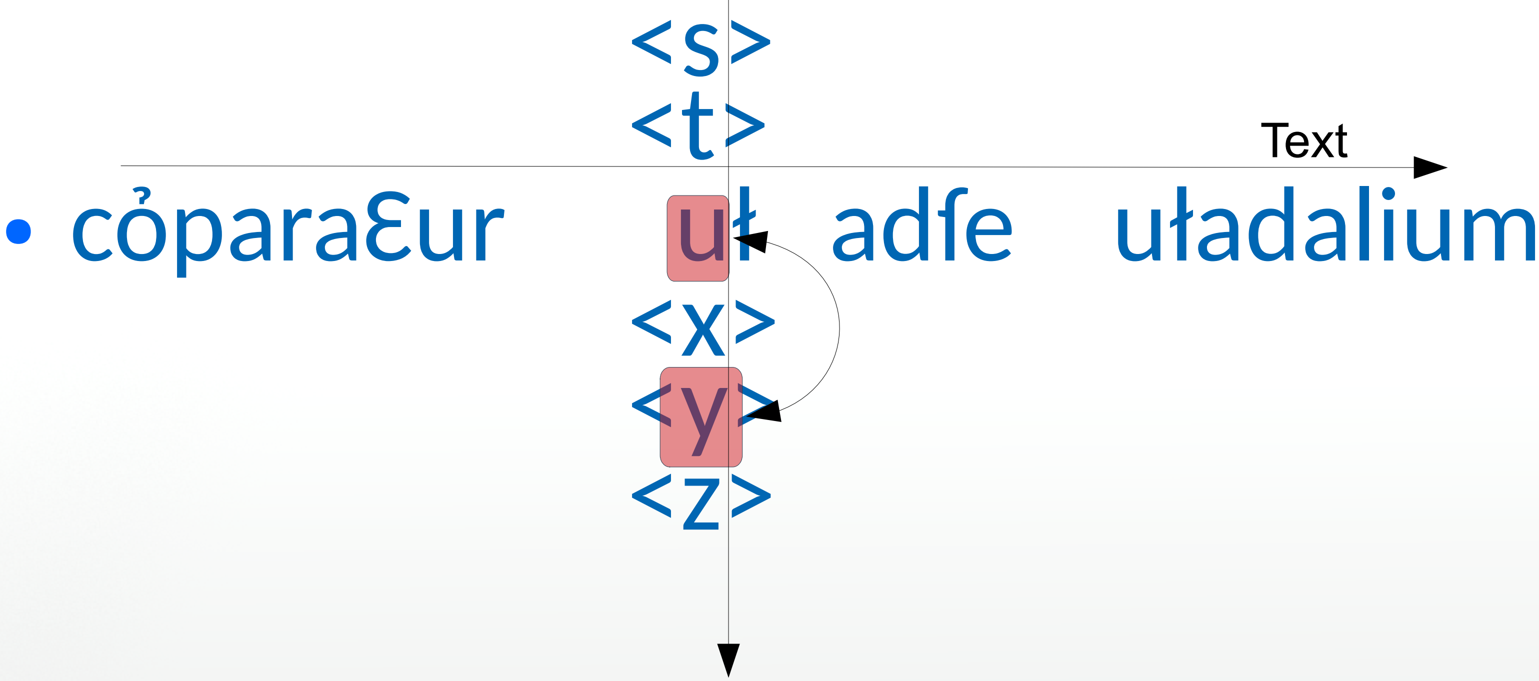
- Scholarly discussion on **modelling**
  - What is a grapheme? What is an allograph?
- **Shared** models
  - Grassroot approach, from discussion
- Reusable **software** libraries

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

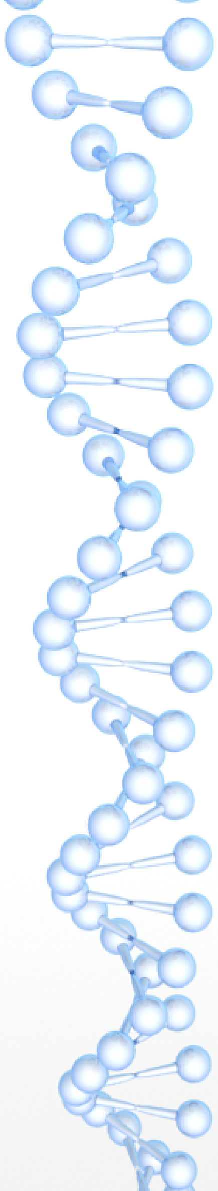
- cỏparaƐur uł adſe uładalium

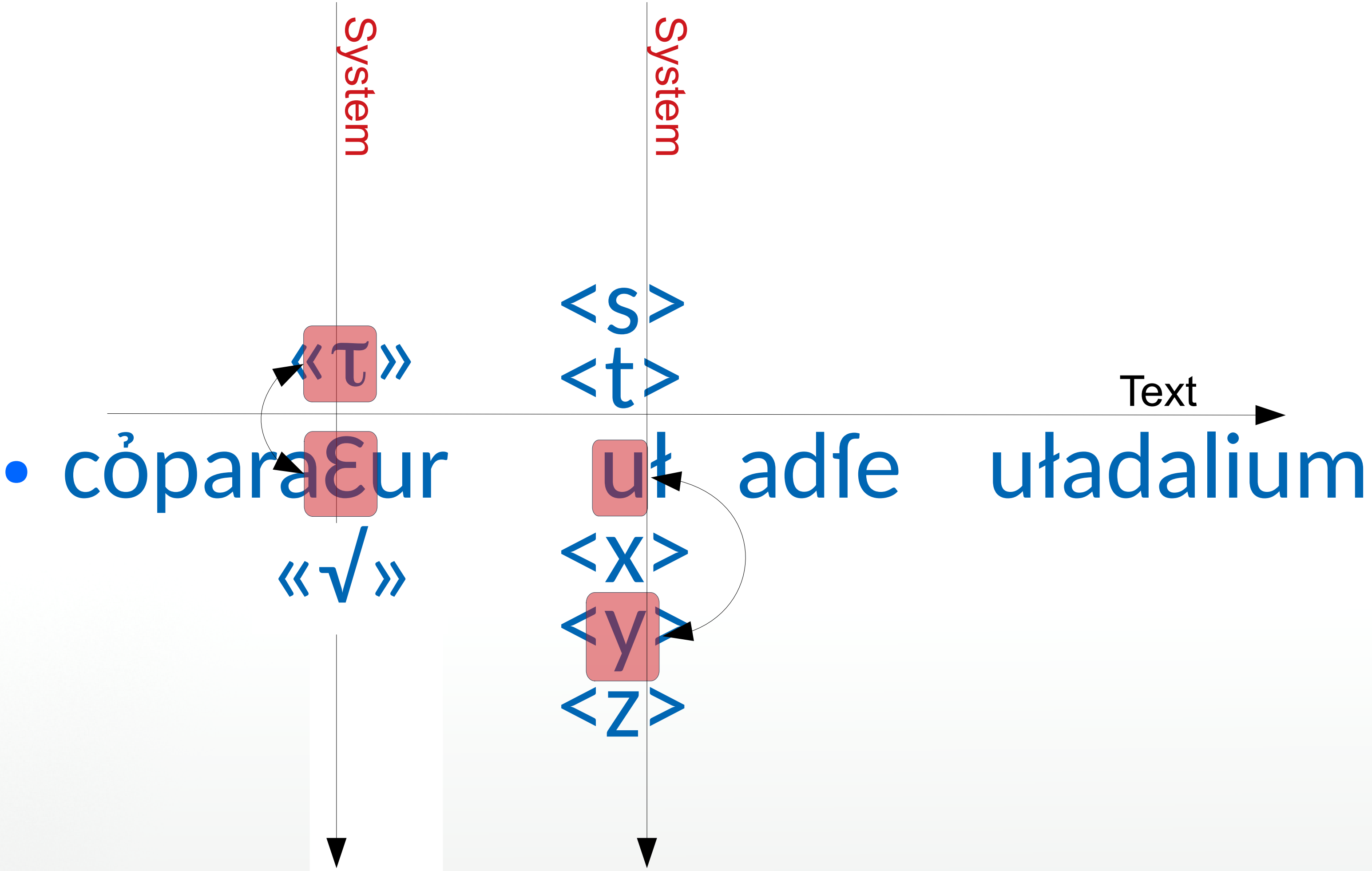
<x>

<y>

<z>

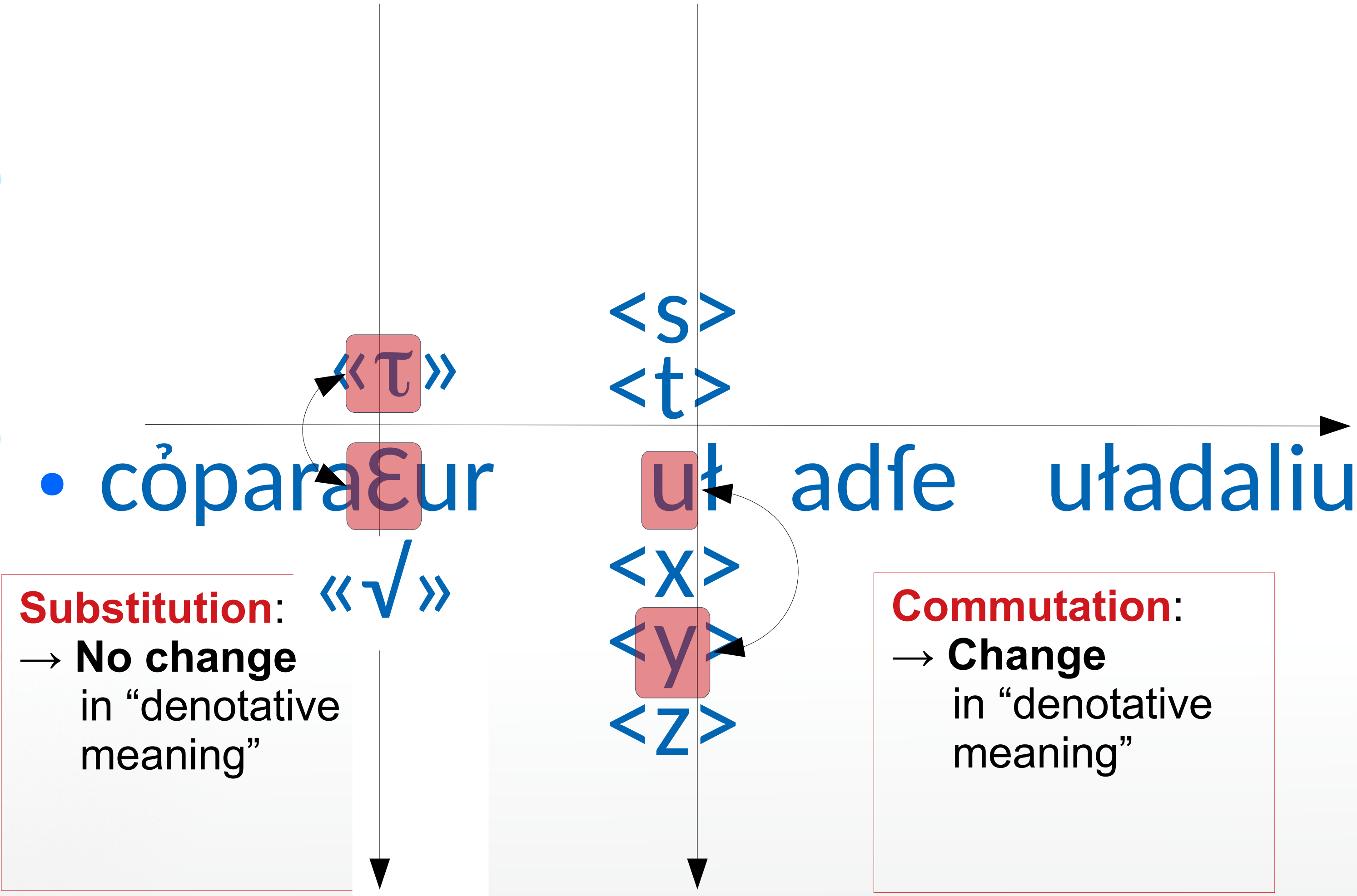
Graphemes/allographs: the commutation test





- cỏparaƐur uł adſe uładalium

Graphemes/allographs: the commutation test

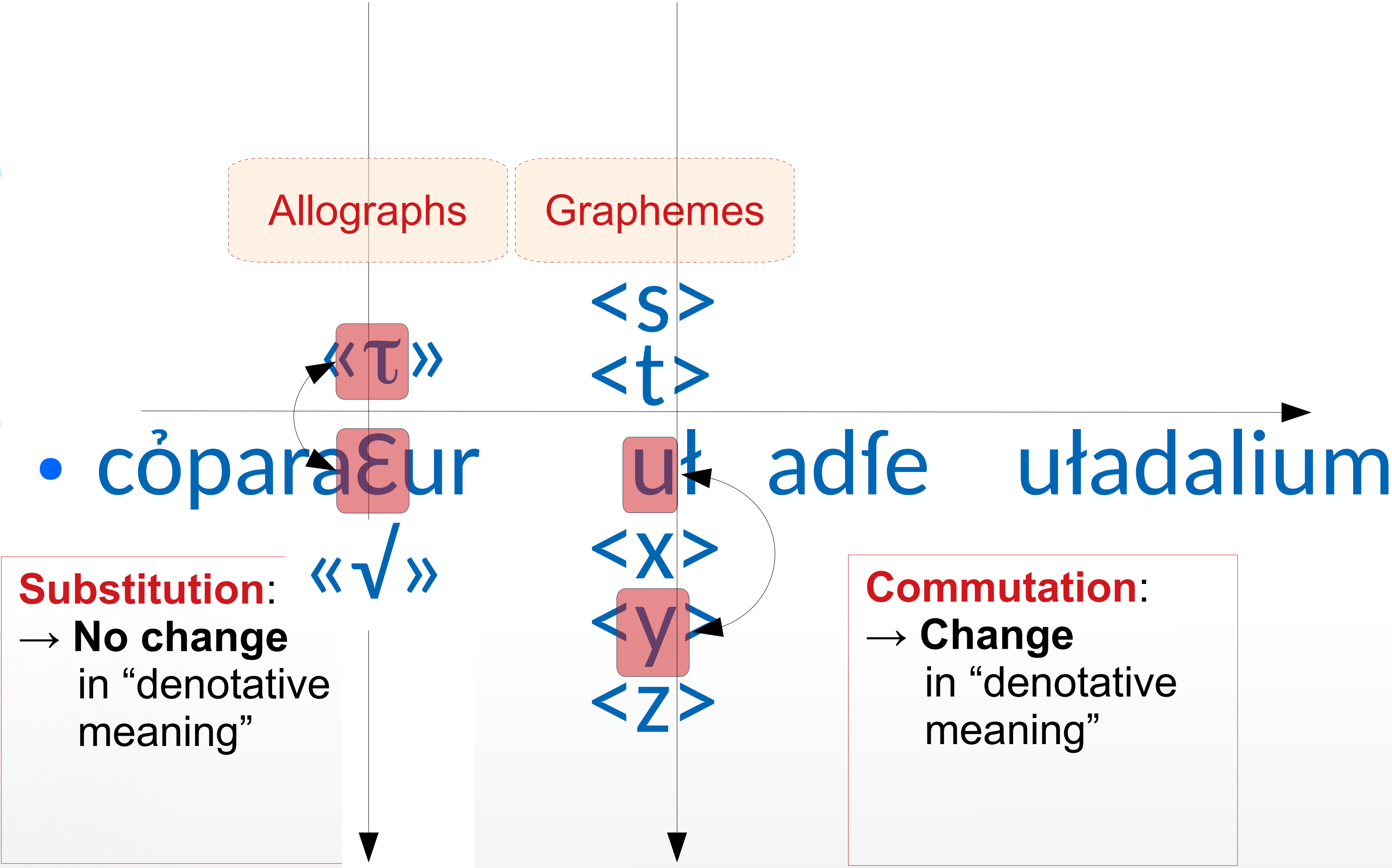


- cỏparaƐur uł adſe uładalium

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>

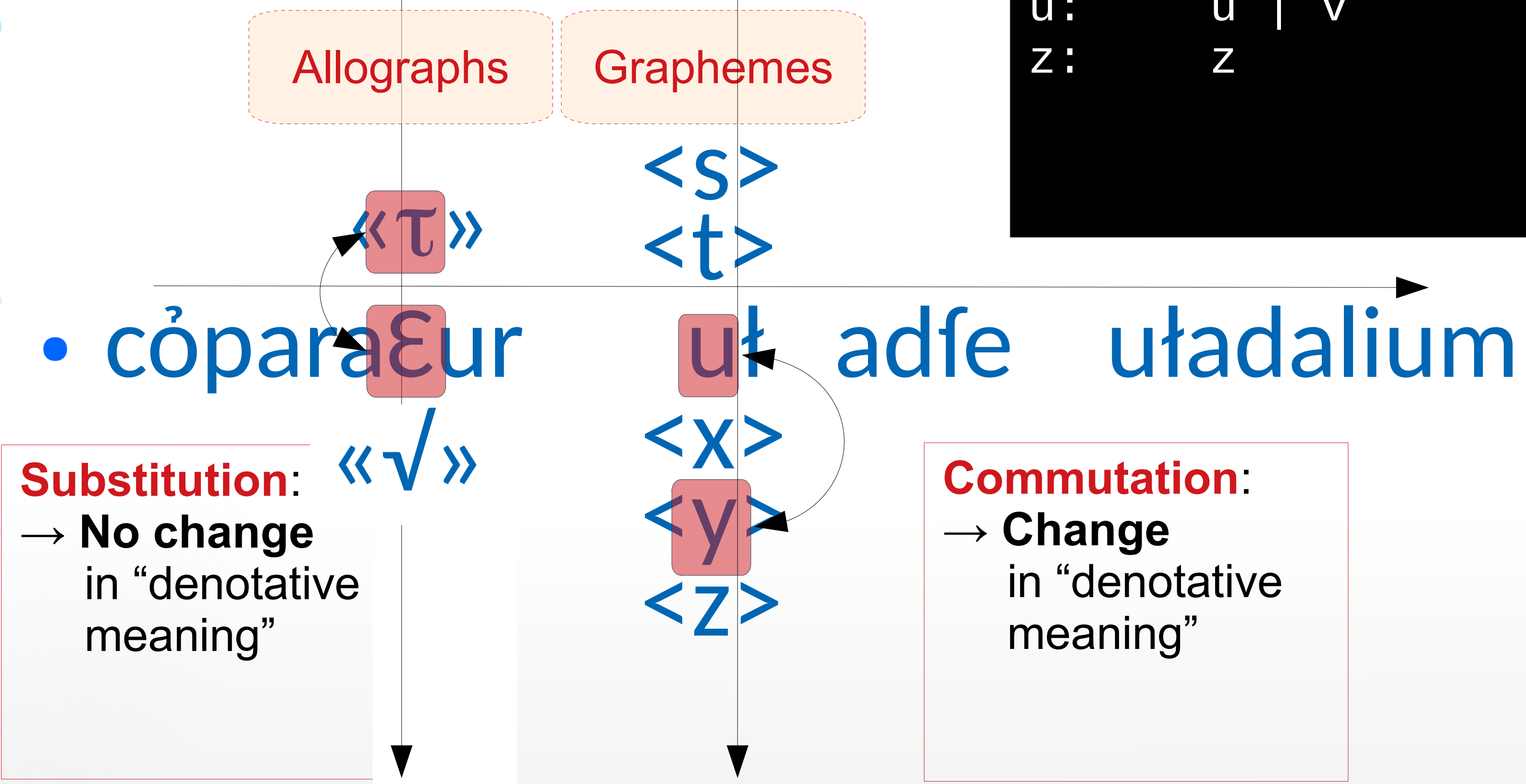
<z>

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
```



Allographs

Graphemes

<s>

«τ» <t>

- cỏparaƐur uł adſe uładalium

«√» <x>

<y>

<z>

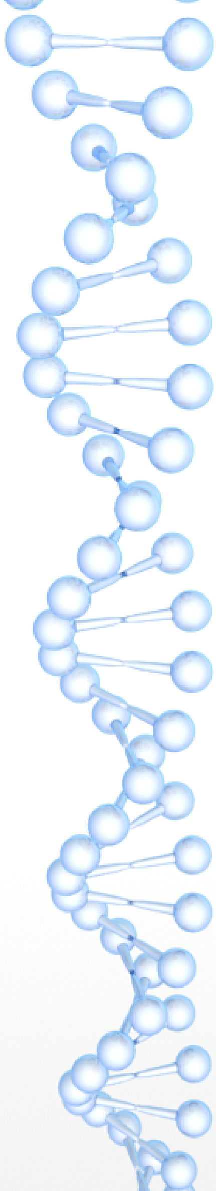
Substitution:
→ **No change**
in “denotative meaning”

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

Graphemes / allographs: 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...)

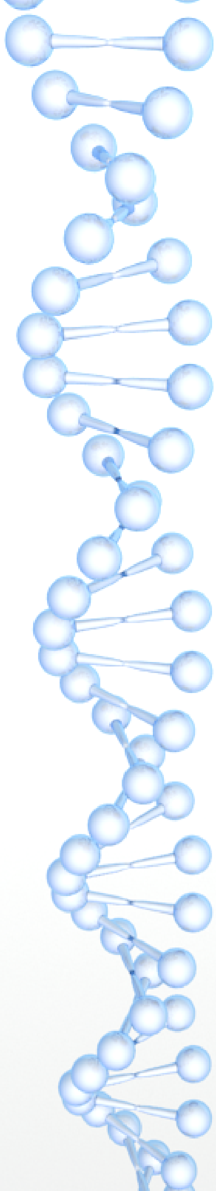
Good practices



Good practices

- See section Seminar readings on GitHub
  - Human-readable tables of signs
  - Machine-readable tables of signs

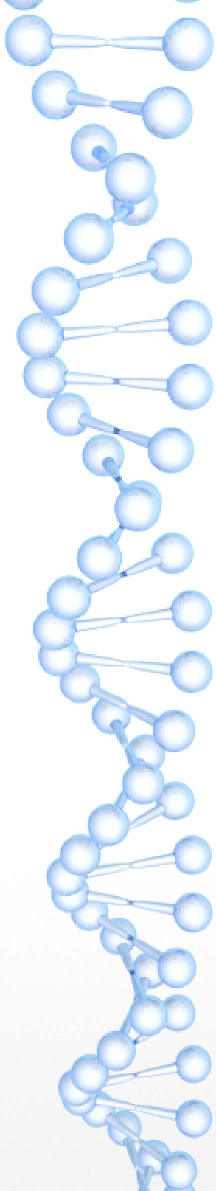
Exercise



Exercise

- See section Exercise on GitHub

Outline



Outline

- **Interoperability**
  of digital scholarly editions (DSEs)
  based on diplomatic transcriptions
  - The issue
  - Current solutions
  - Interoperability through modelling
- Orlandi’s **table of sign**
- **Graphemes/allographs**