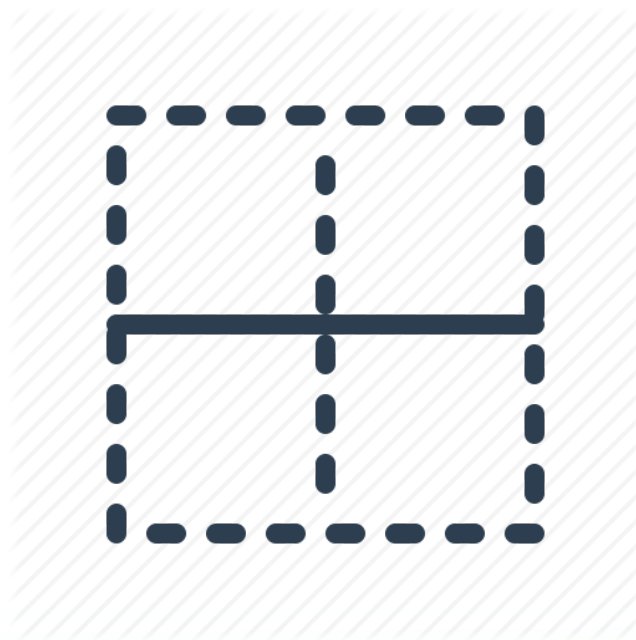


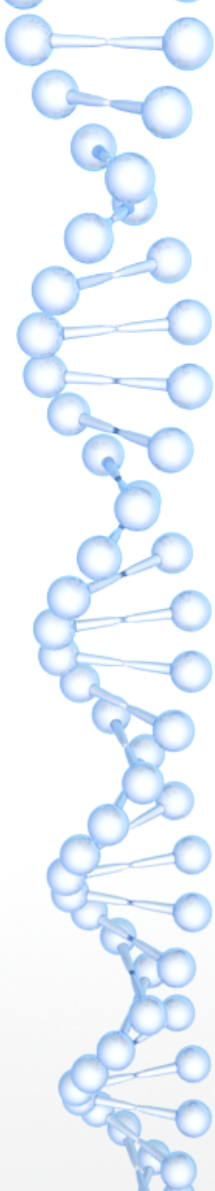
Paolo Monella

Encoding Pre-modern Writing Systems



*Sunoikisis Digital Classics,
Summer 2022 programme*
May 12, 2022





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



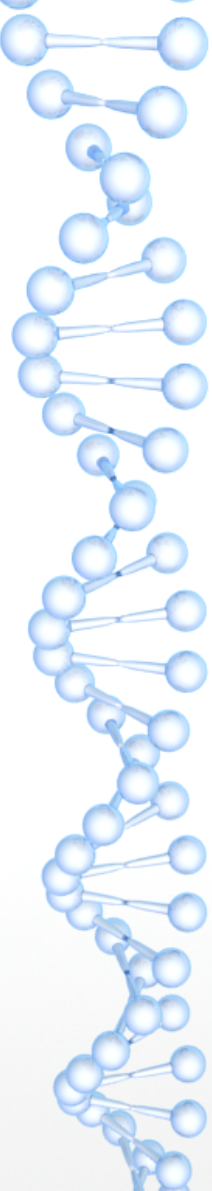
Outline

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

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.

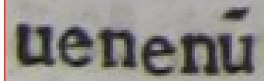
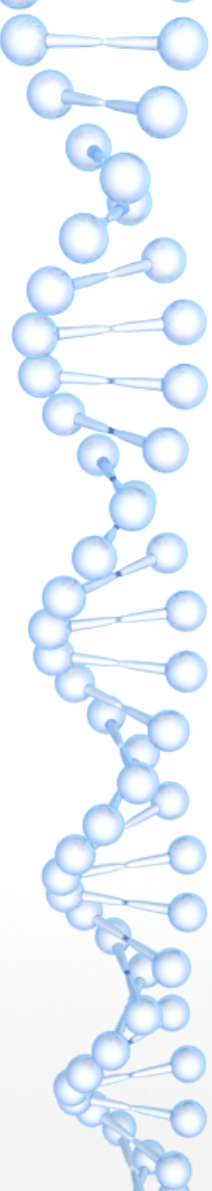
Interoperability: the issue

Interoperability: the issue



uenenũ

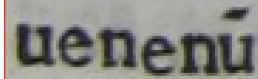
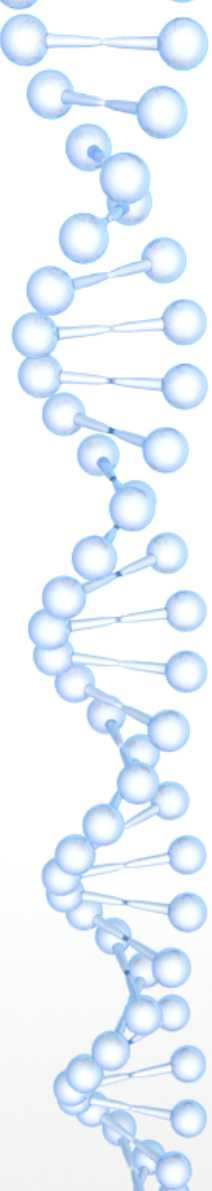
Interoperability: the issue



uenenū

- uenenū

Interoperability: the issue



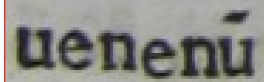
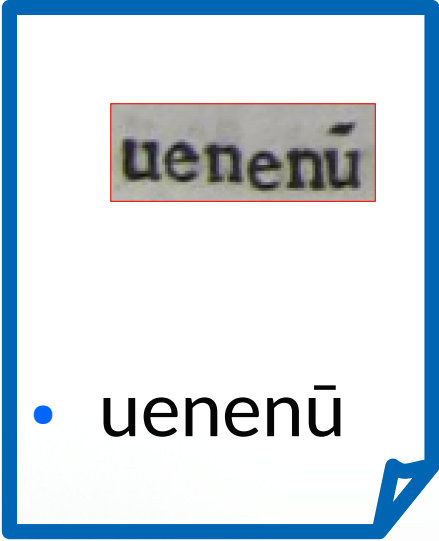
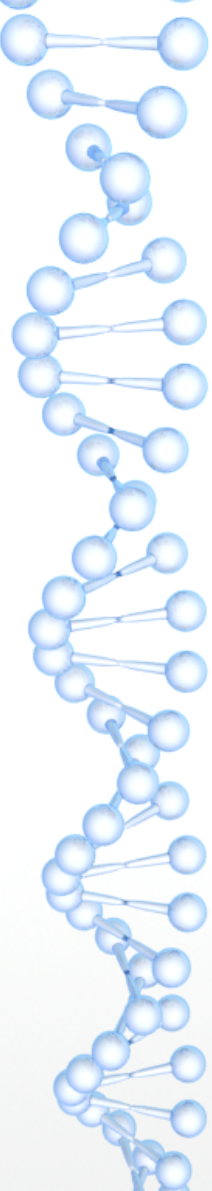
uenenū

- uenenū

Diplomatic

- Manual or HTR
 - Visualization
 - Processing

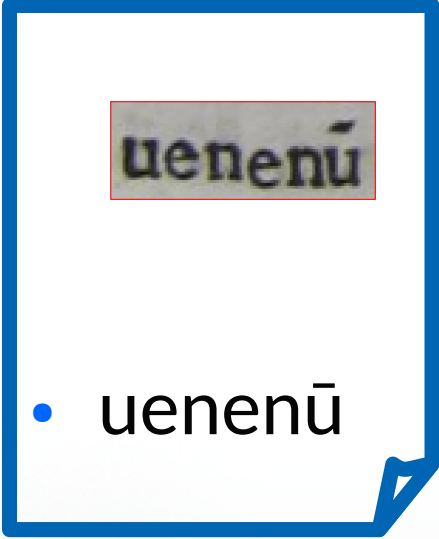
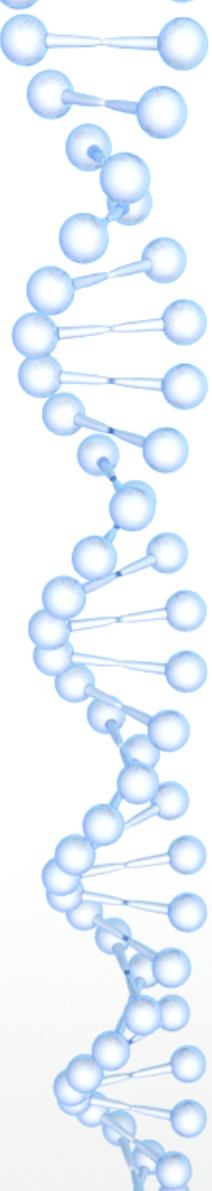
Interoperability: the issue



uēnenū

- uēnenū

Interoperability: the issue



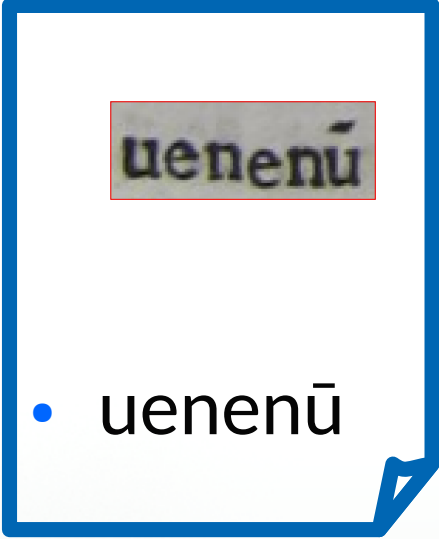
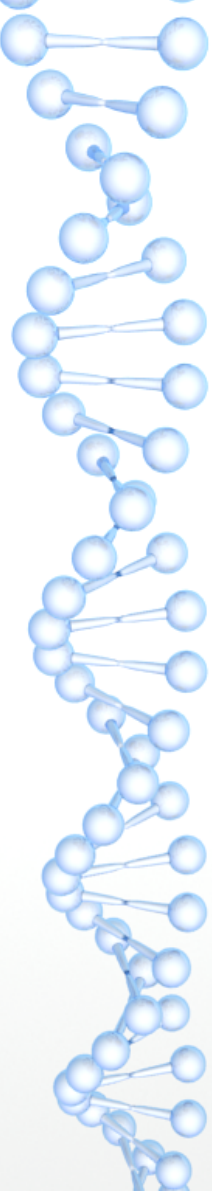
uenenū

- uenenū

- 
- uenenum

Interoperability: the issue

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

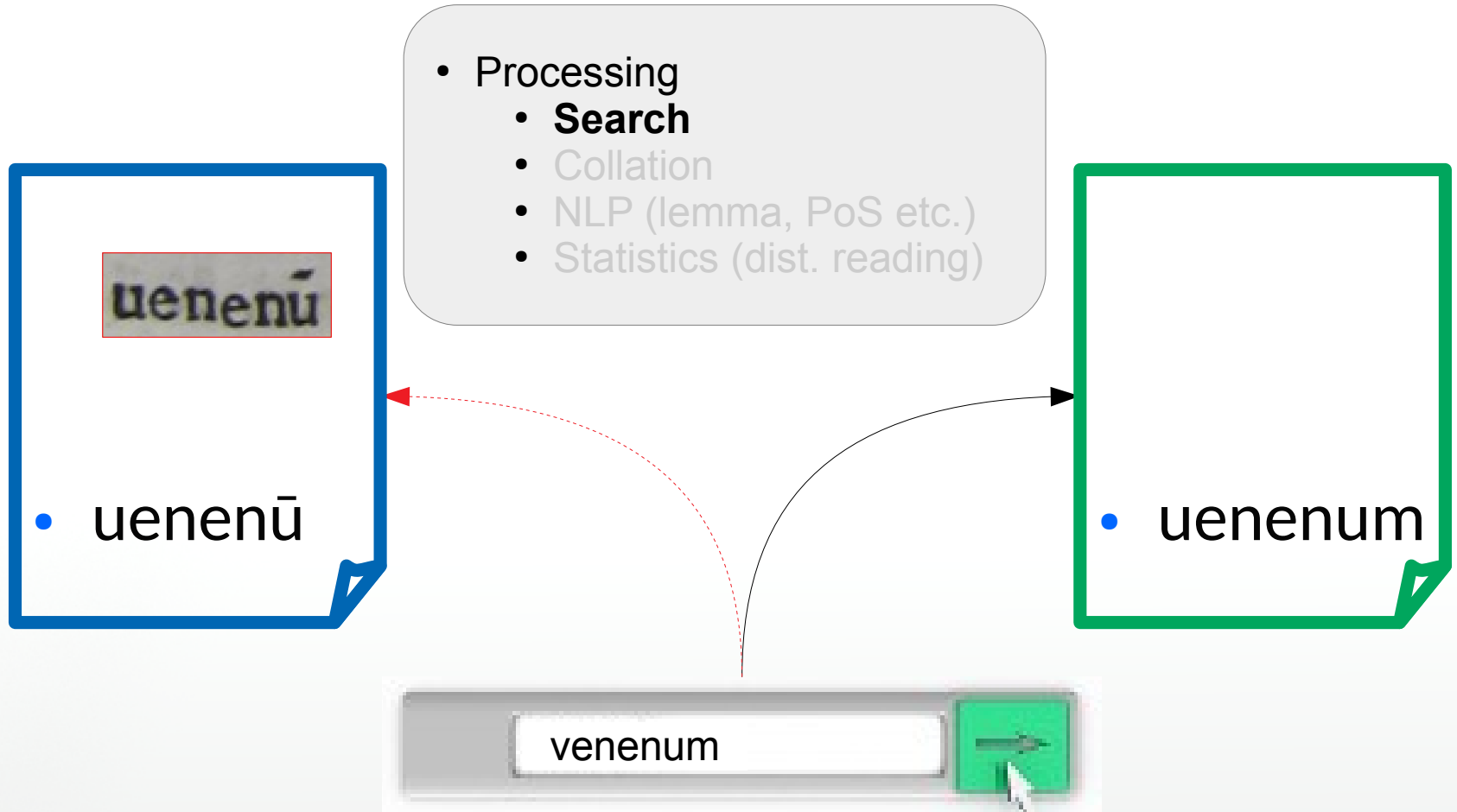


uenenū

- uenenū

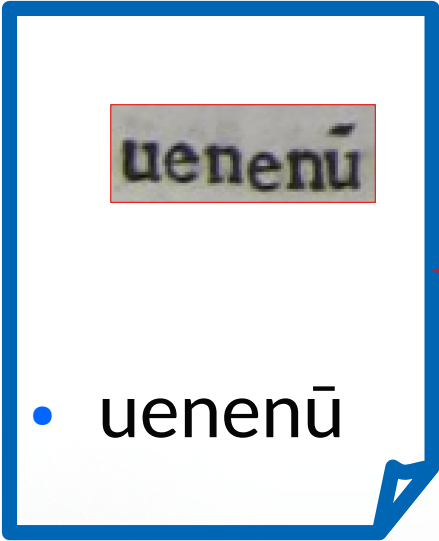
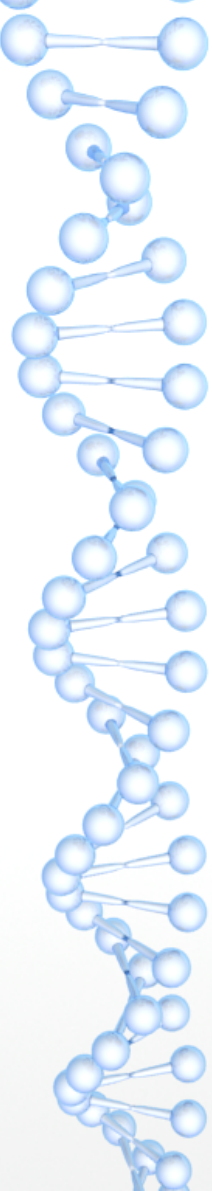
- 
- uenenum

Interoperability: the issue



Interoperability: the issue

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



uenenū

• uenenū

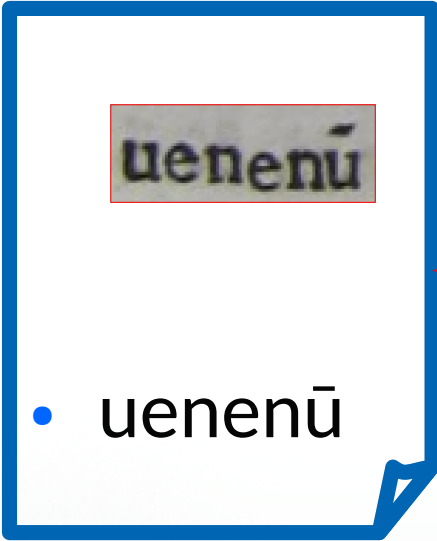
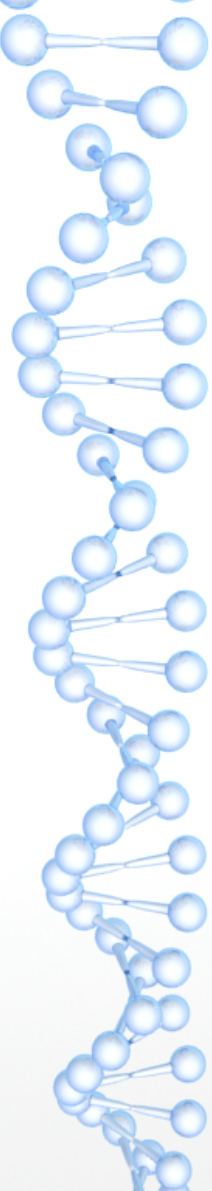
112 excedit] excedit corr. ex
exceditis R n70
114 obicitur] obiceretur V S n71
114 sunt] sint S n72



• uenenum

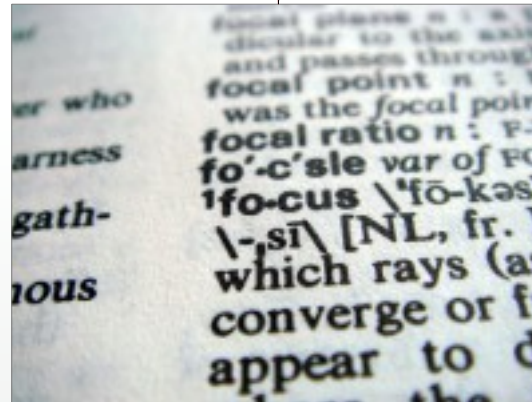
Interoperability: the issue

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



uenenū

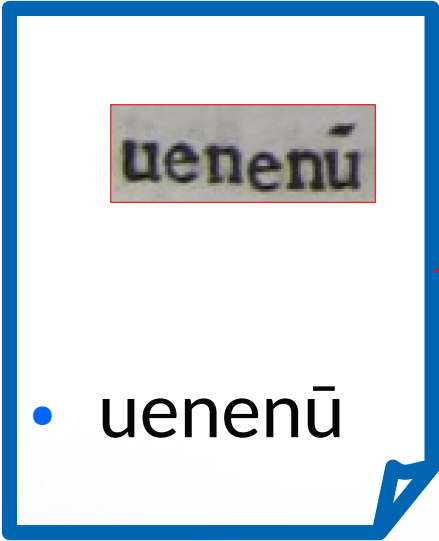
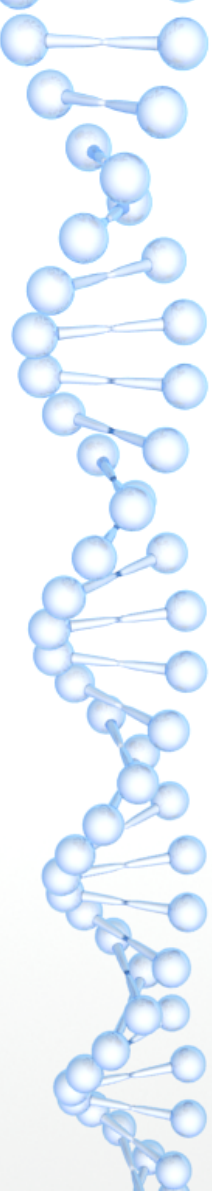
• uenenū



• uenenum

Interoperability: the issue

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



uenenū

• uenenū



• uenenum





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.

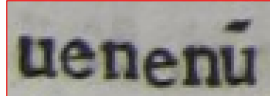
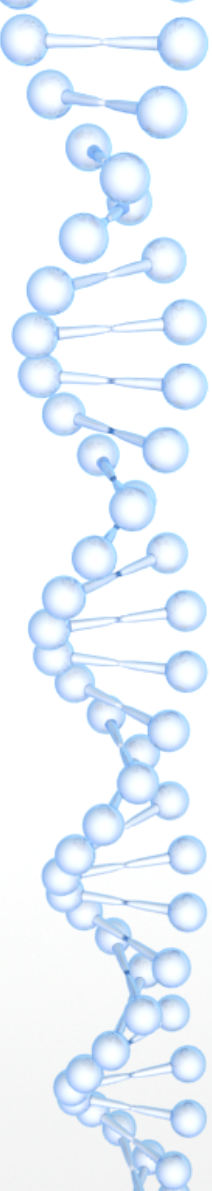
**Interoperability:
current solutions**



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 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



uenenū

- uenenū

Diplomatic

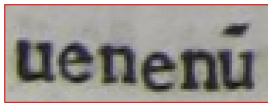
- Visualization
- Processing



Manual normalized transcription

- venenum

Normalized



uenenū

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

- uenenū

Diplomatic

- Visualization
- Processing

Manual normalized transcription

Not generated
by computer:
repeated manual
transcription

• **v**enenum

uenenū

Normalized

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

• **u**enenū


Diplomatic

- Visualization
- Processing

Application-specific pre-processing

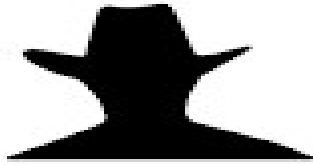
- Collation: pre-processing to normalize the text
 - Or instruct collation software to ignore *some* discrepancies (e.g. case, u/v etc.)
- Same with search/indexing/NLP/statistics software
- Disposable home-made solutions
 - ...for a shared issue



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**

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

Interoperability through modelling



WANTED

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

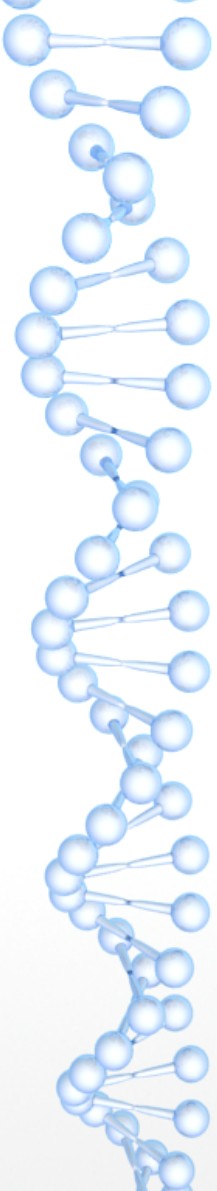
- Documenting project-specific modelling (transcription) and normalization practices
 - What am I transcribing (graphemes/allographs)? How?
 - What corresponds to what (f/s)?
- Documentation
 - In English prose (specifications for other programmers)
 - Formal (software code, tables)

Example ahead!

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 "Orlandi's table of signs" is centered within the rectangular body.

Orlandi's table of signs

Orlandi's table of signs



OCR/HTR
(witness A)



Manual
transcription
(witness B)

Orlandi's table of signs



Allographic
transcription

Vnder τhis Castτle



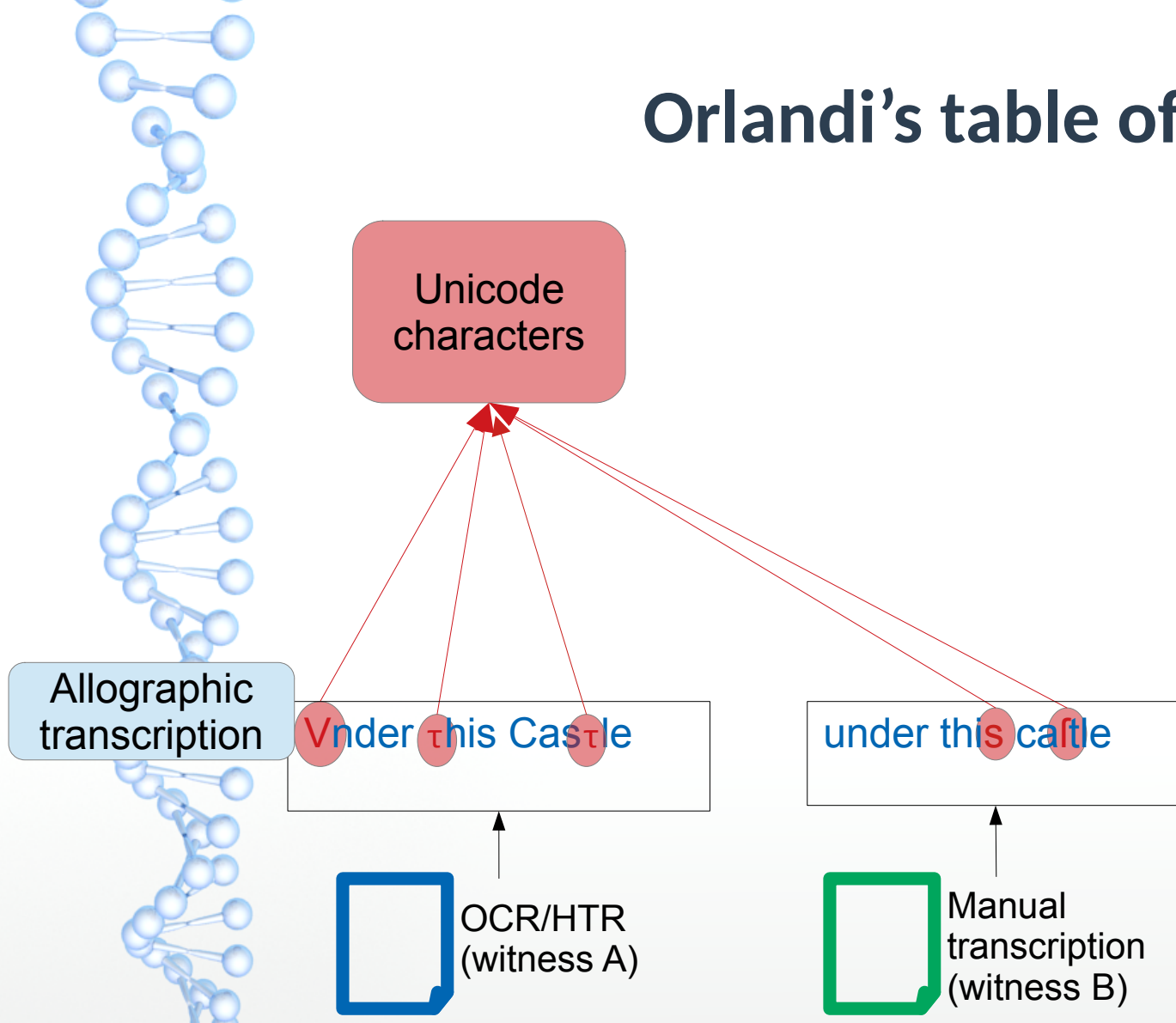
OCR/HTR
(witness A)

under this caτle



Manual
transcription
(witness B)

Orlandi's table of signs



Orlandi's table of signs



Allographic
transcription

Vnder τhis Castτle

OCR/HTR
(witness A)

under this caτle

Manual
transcription
(witness B)

Orlandi's table of signs



Allographic
transcription

Vnder this Castle

under this caſtle

OCR/HTR
(witness A)

Manual
transcription
(witness B)

Orlandi's table of signs



Allographic
transcription

Vnder τhis Castle

under this castle



OCR/HTR
(witness A)



Manual
transcription
(witness B)

Orlandi's table of signs



Allographic
transcription

Vnder τhis Castle



OCR/HTR
(witness A)

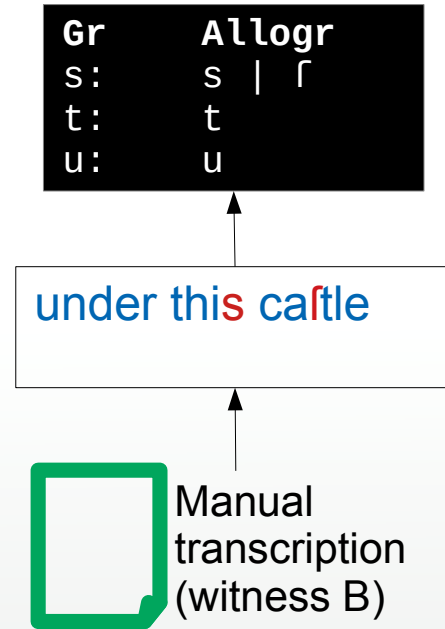
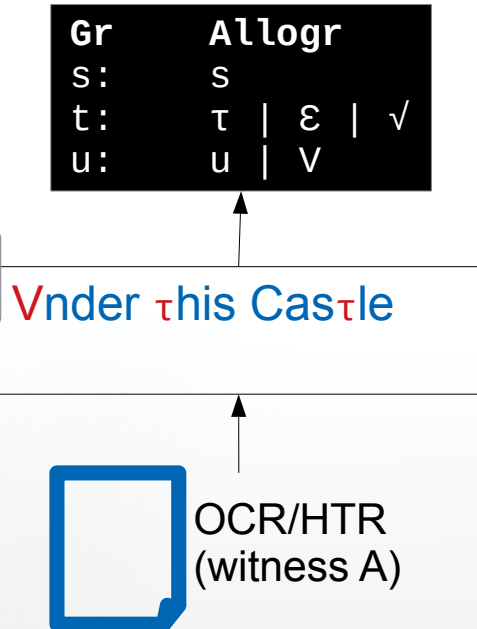
under this castle



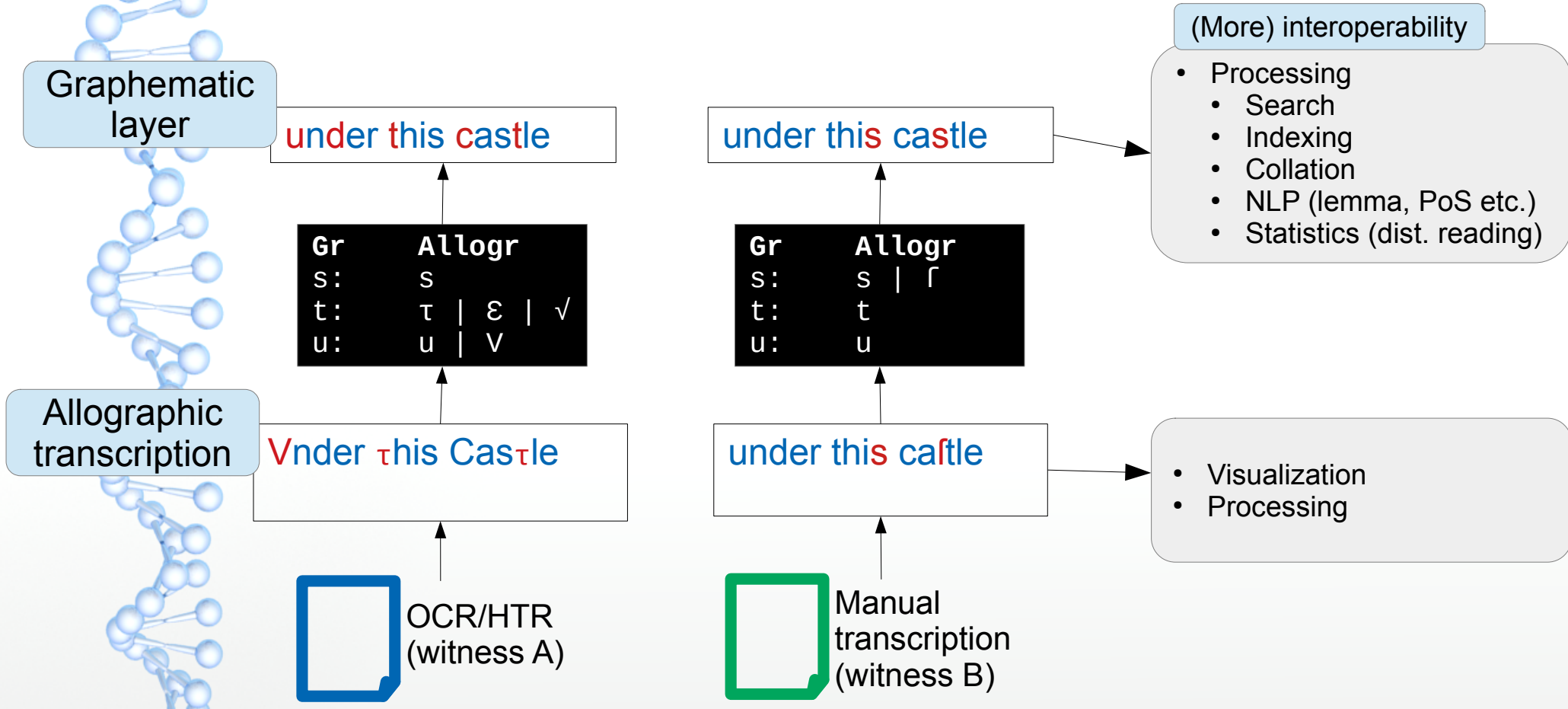
Manual
transcription
(witness B)

Orlandi's table of signs

Allographic transcription



Orlandi's table of signs



Orlandi's table of signs

Graphematic
layer

under this castle

Gr	Allogr
s:	s
t:	τ ε \vee
u:	u V

Allographic
transcription

Vnder τ his Castle



OCR/HTR
(witness A)

under this castle

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

under this caſtle



Manual
transcription
(witness B)

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

Interoperability through modelling



WANTED

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

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

Interoperability through modelling



WANTED

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

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

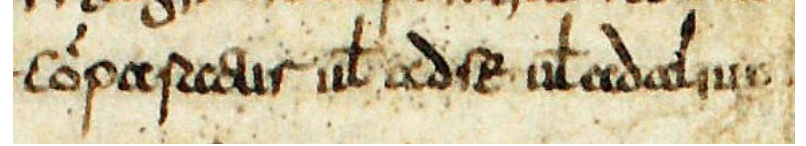


Graphemes/allographs

Graphemes/allographs: the commutation test

System

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



<s>

<t>

Text

- cōparaEUR

u

adfe

uʔadaliu

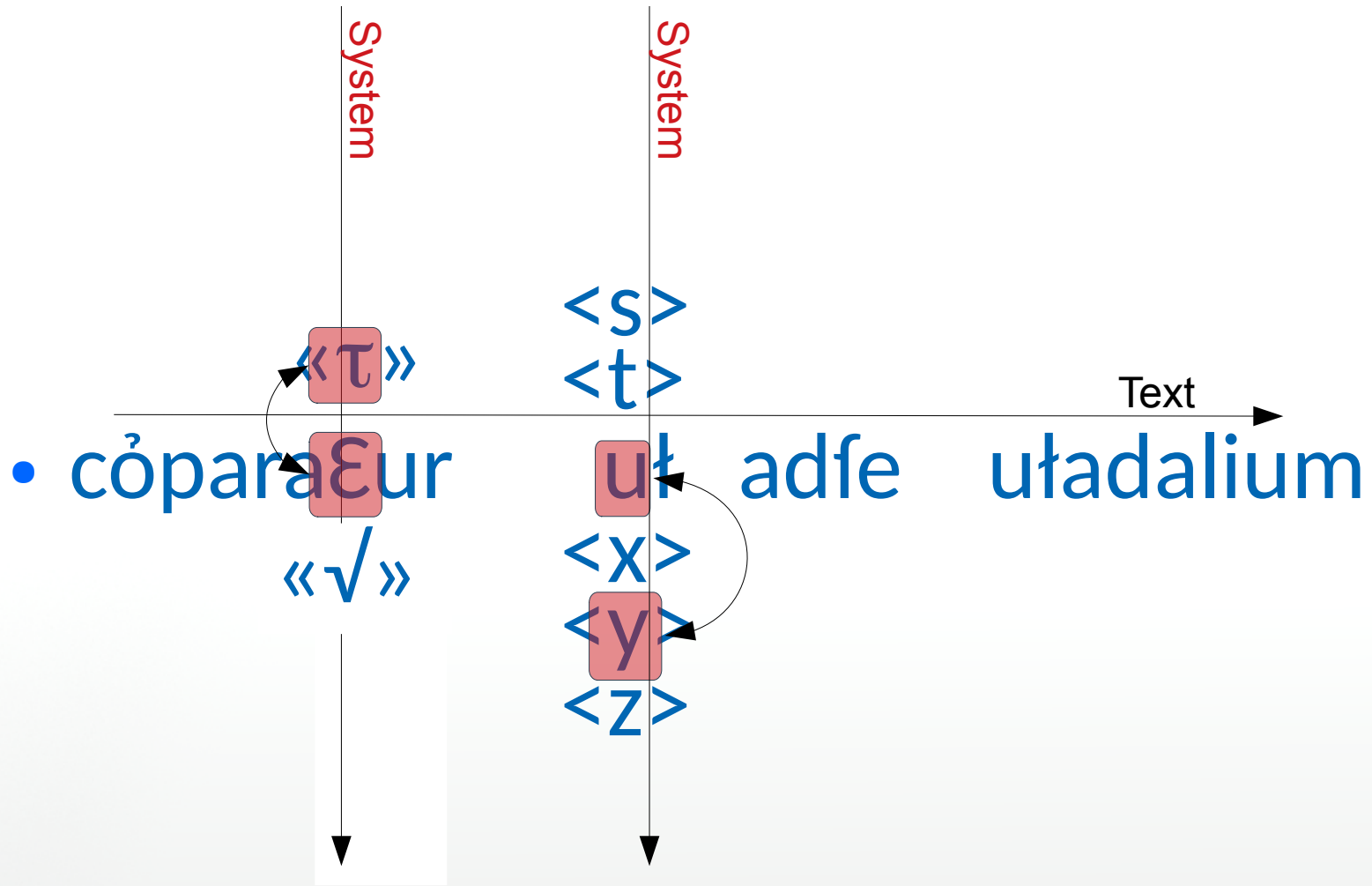
<x>

<y>

<z>



Graphemes/allographs: the commutation test



Graphemes/allographs: the commutation test

• còparaEur

Substitution:

→ No change
in “denotative
meaning”

«τ»

«√»

<s>
<t>

<x>

<y>

<z>

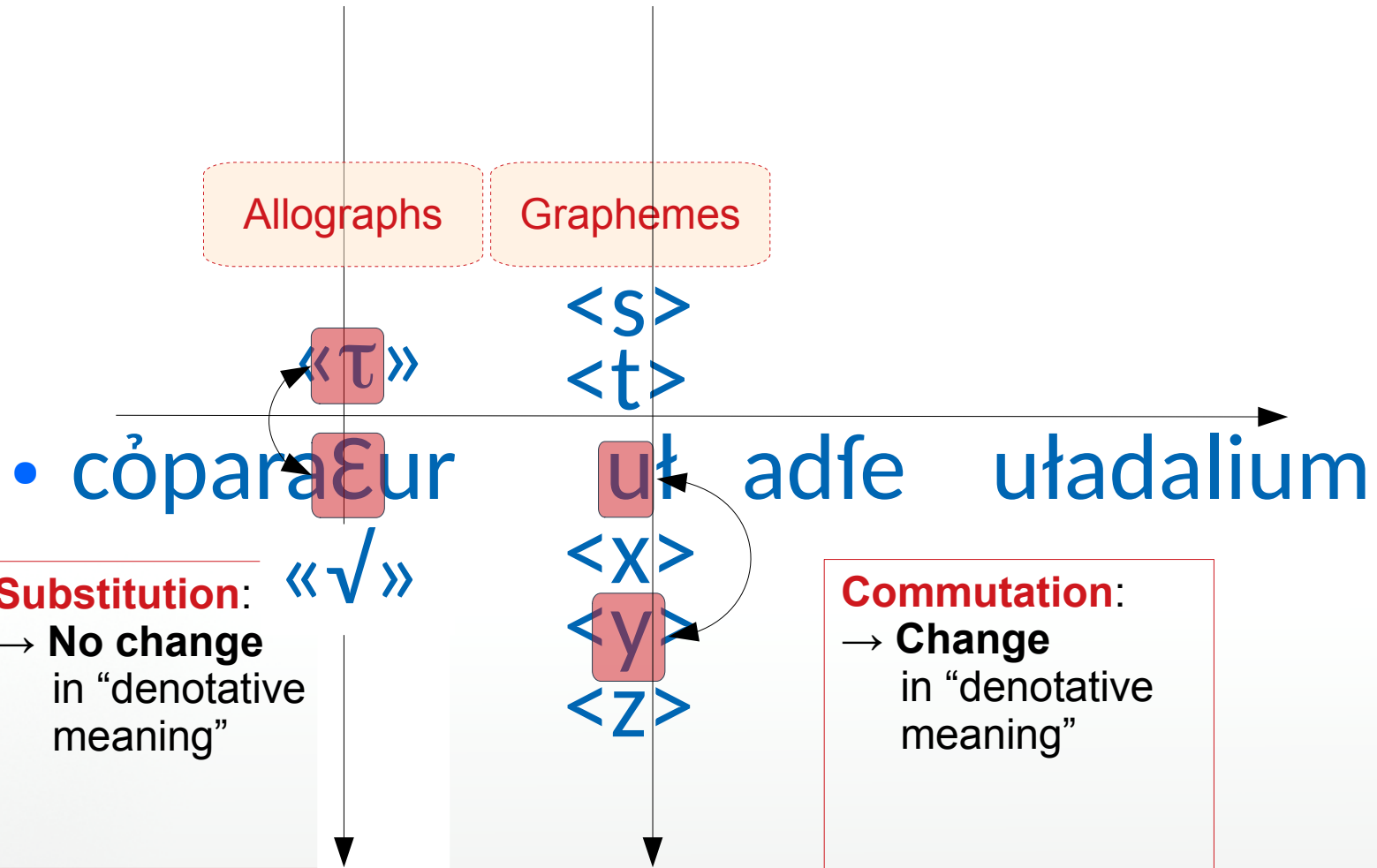
adfe

uɫadaliu

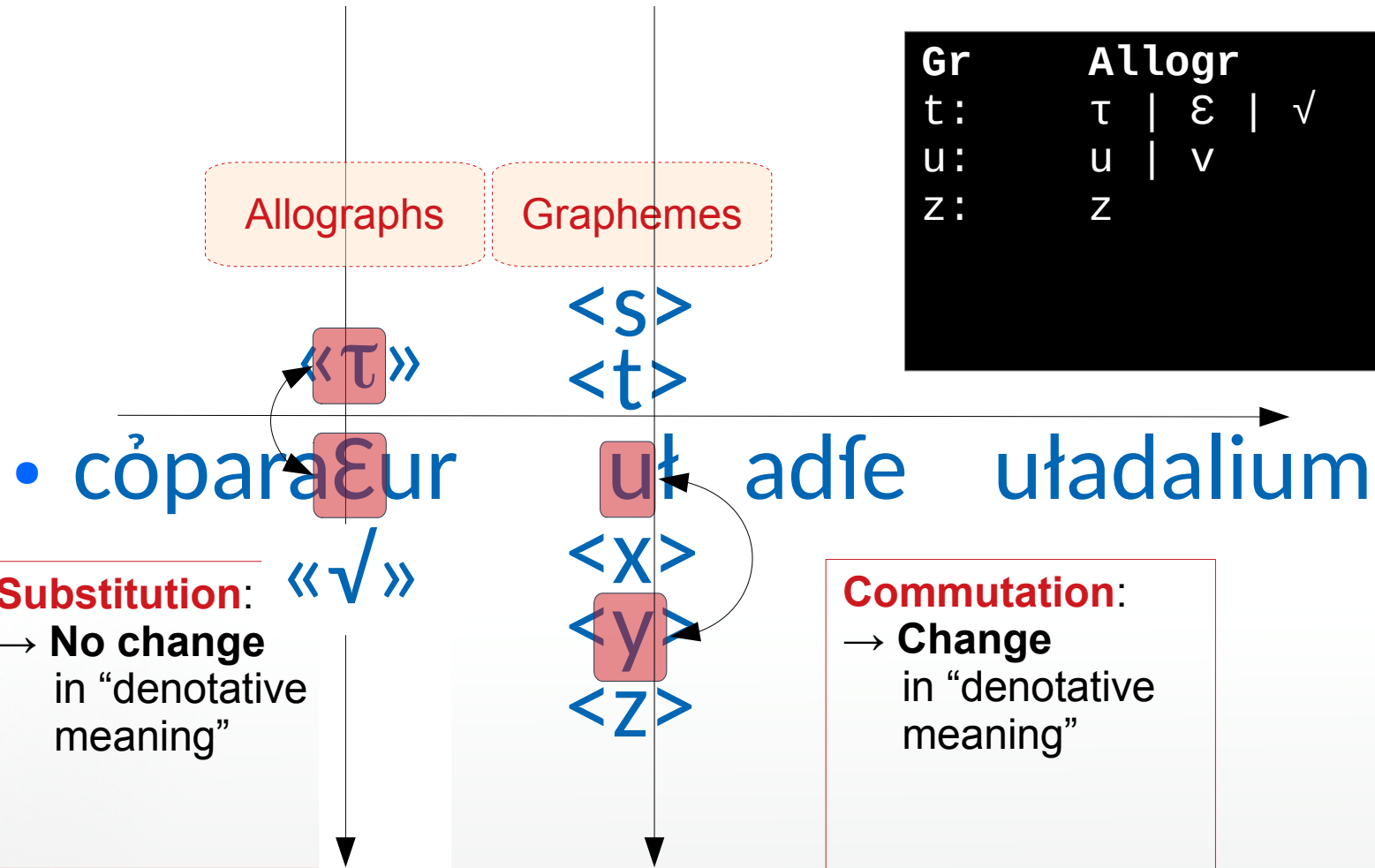
Commutation:

→ Change
in “denotative
meaning”

Graphemes/allographs: the commutation test

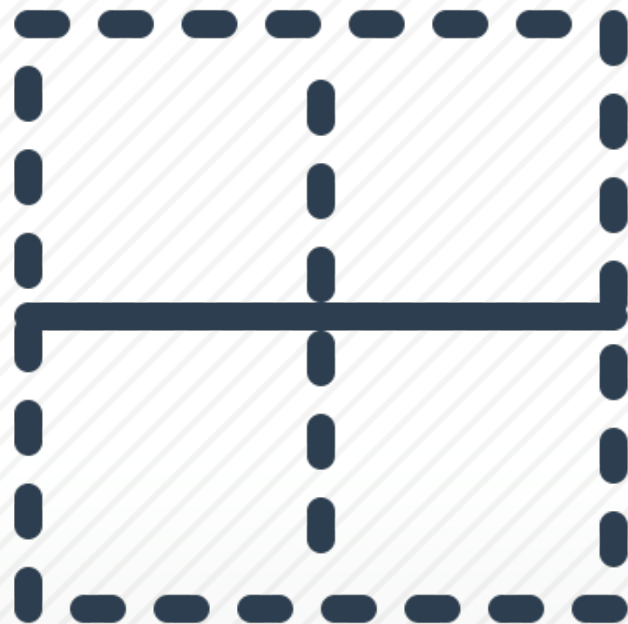


Graphemes/allographs: the commutation test



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

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. The word "Exercise" is centered within the rectangular body in a dark blue, bold, sans-serif font.

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**