Paolo Monella
paolo.monella@gmx.net

In the Tower of Babel

Modelling primary sources of multi-testimonial textual transmissions

Digital Classicist & Institute of Classical Studies Seminar 2012
London, 20 July 2012

Outline

- The 'Babel issue'
  - The 'Babel issue' in a nutshell
  - Comparing texts
  - In the Tower
- Two issues
  - A. Comparison at the linguistic layer
  - B. Comparison at the graphemic layer

Outline

- The 'Babel issue'
  - The 'Babel issue' in a nutshell
  - Comparing texts
  - In the Tower
- Two issues
  - A. Comparison at the linguistic layer
  - B. Comparison at the graphemic layer

The 'Babel issue' in a nutshell

- Each primary source uses a different writing (encoding, semiotic) system
- We want to compare the texts they bear

Outline

- The 'Babel issue'
  - The 'Babel issue' in a nutshell
  - Comparing texts
  - In the Tower
- Two issues
  - A. Comparison at the linguistic layer
  - B. Comparison at the graphemic layer

Comparing texts

- Each primary source uses a different writing (encoding, semiotic) system
- We want to compare the texts they bear
  - Textual Criticism
  - Processing (e. g. cross-corpus search)

Comparing texts

- Each primary source uses a different writing (encoding, semiotic) system
- We want to compare the texts they bear
  - Textual Criticism
  - Processing (e. g. cross-corpus search)

MS A	MS B	Print ed. C
p̲ui’$_G$	peruius$_G$	pervius$_G$

Comparing texts

- Each primary source uses a different writing (encoding, semiotic) system
- We want to compare the texts they bear
  - Textual Criticism
  - Processing (e. g. cross-corpus search)

MS A
p̲ui’$_G$

MS B
peruius$_G$

Query
pervius$_G$

Comparing texts

- Simplification: same writing system (no 'Babel' issue)
  - simple variant (at **graphemic** layer)

MS A

w_G	↔	w_G
y_G	↔	y_G
f_G	↔	f_G
	↮	$\mathbf{f_G}$
e_G	↔	e_G

MS B

MS A

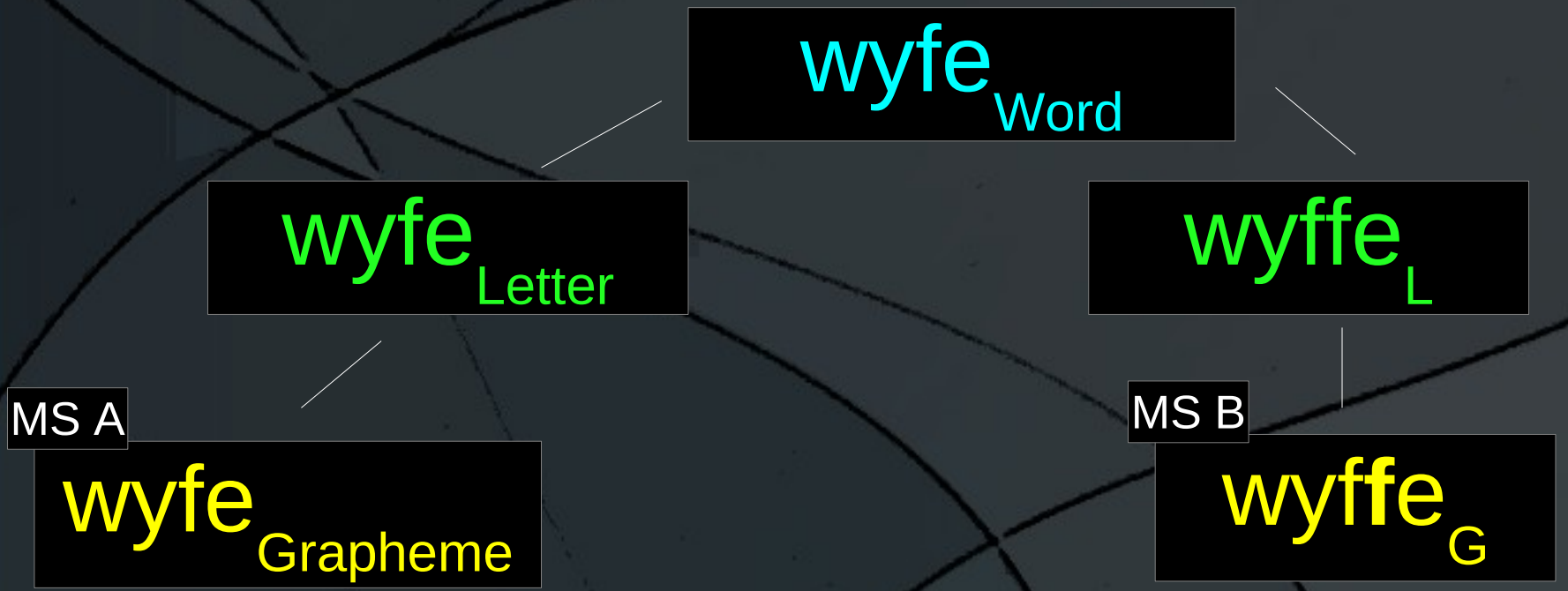
$wyfe_{Grapheme}$

MS B

$wyffe_G$

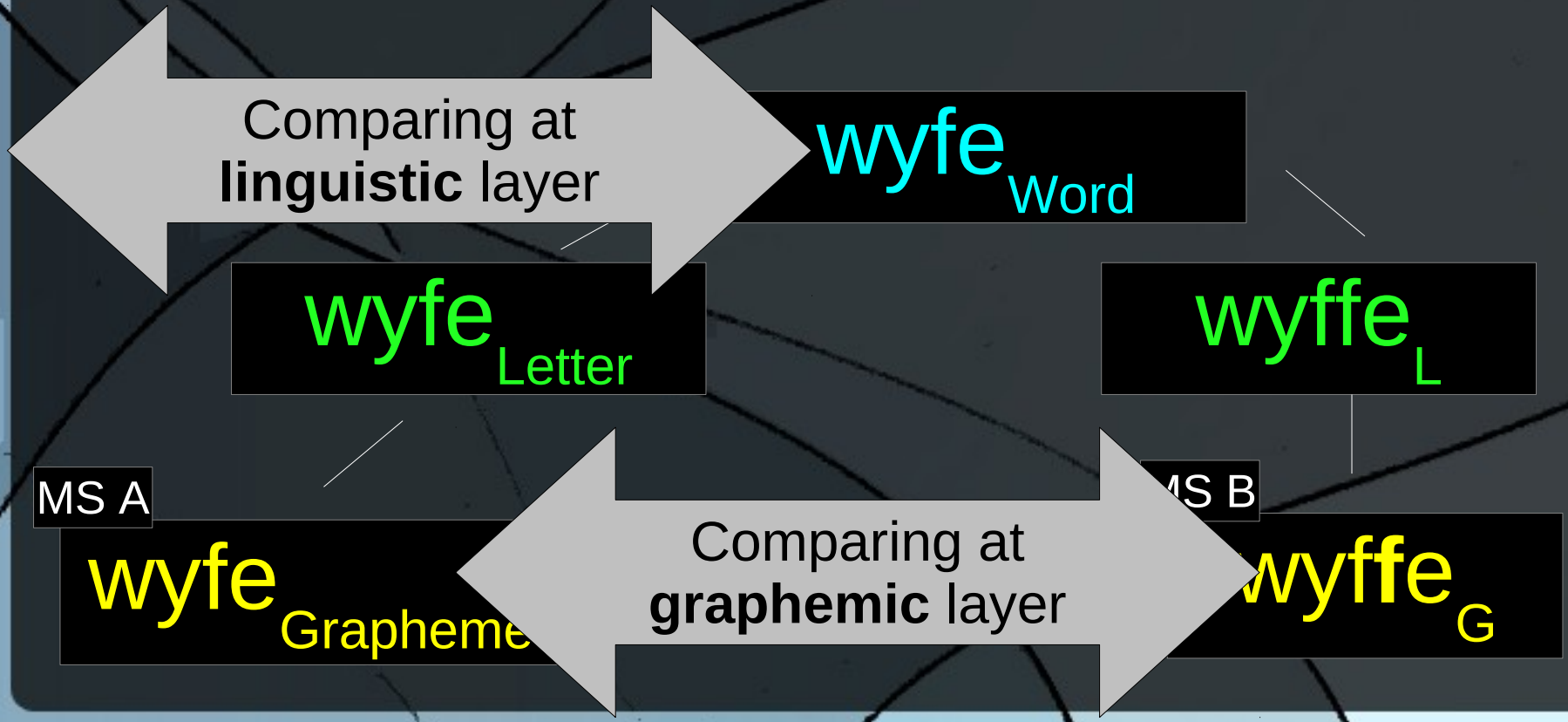
Comparing texts

- Simplification: same writing system (no 'Babel' issue)
  - no variant (at **linguistic** layer)



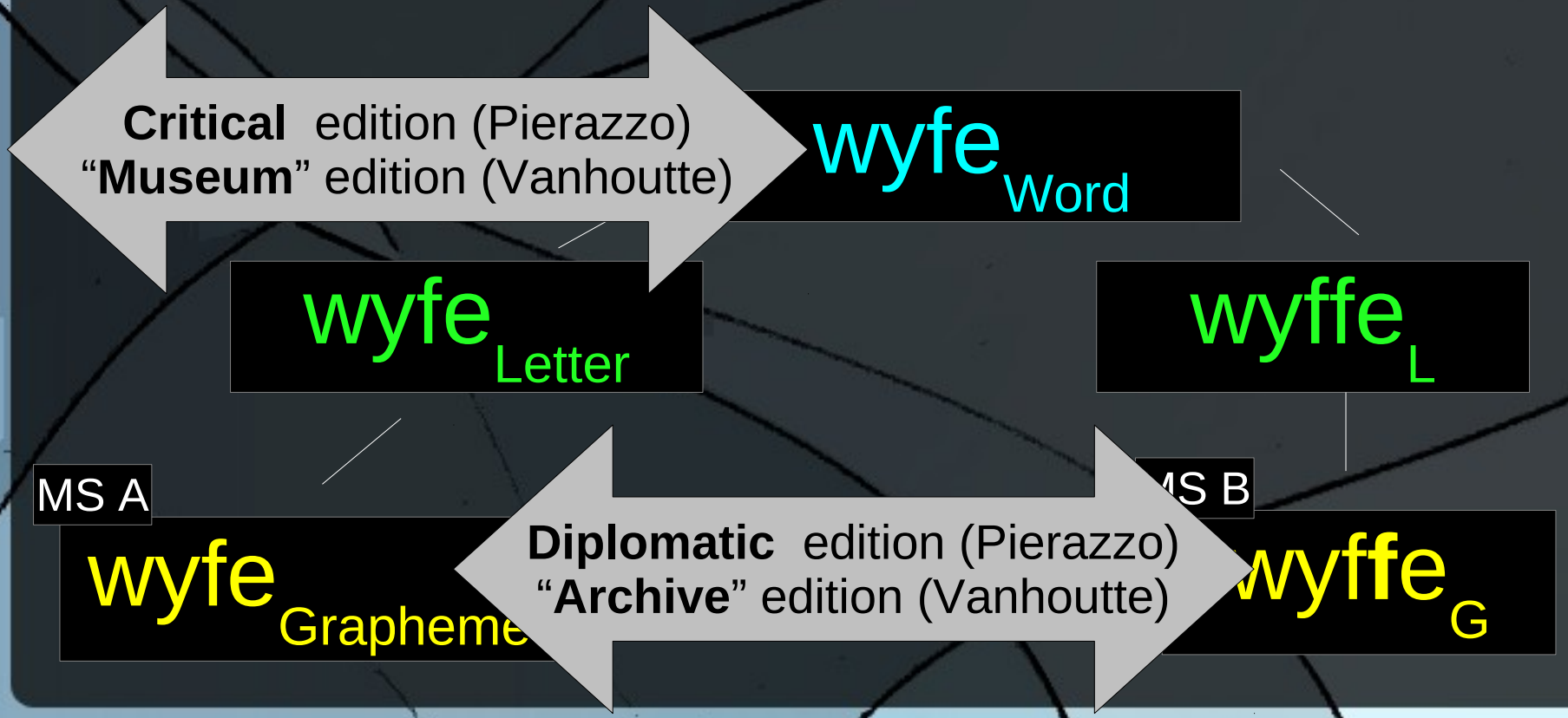
Comparing texts

- The same text (reading)?
  - Comparing texts at different layers



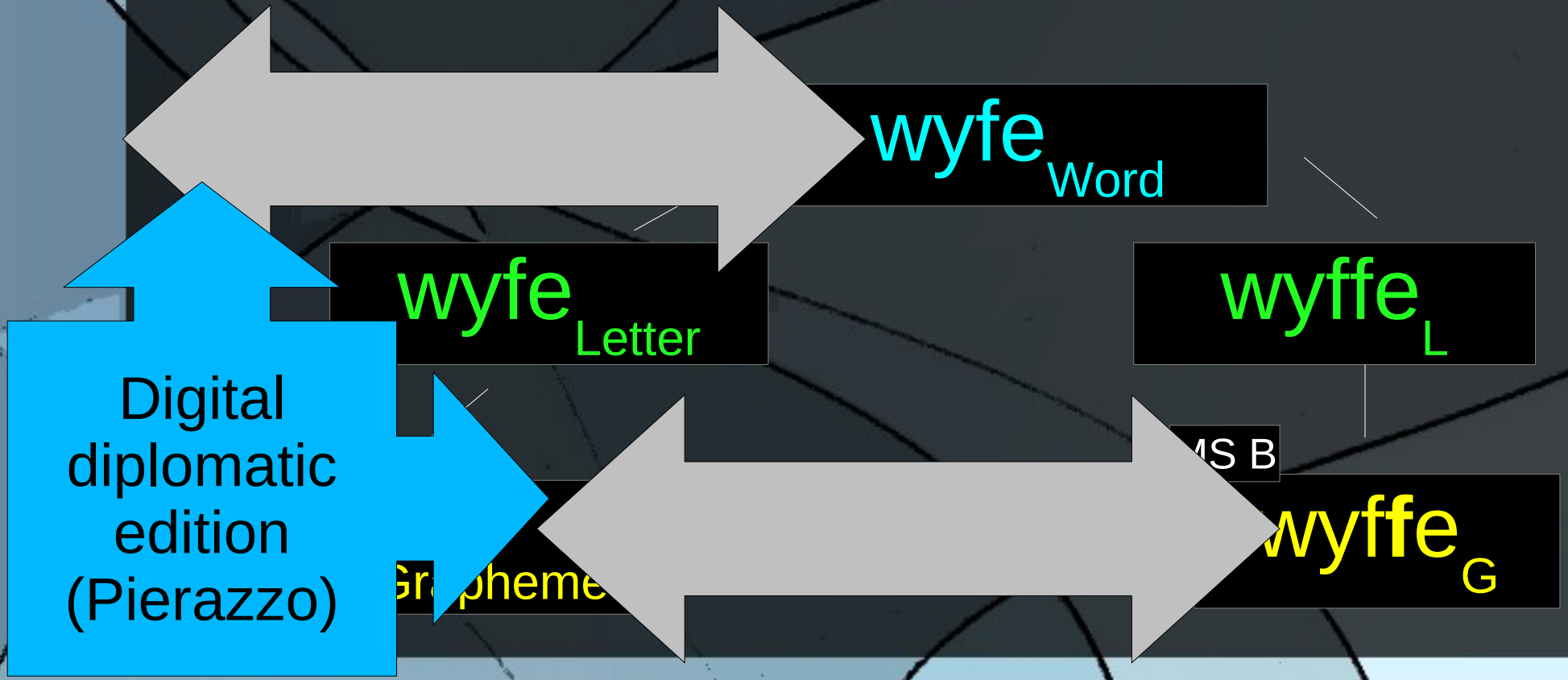
Comparing texts

- Types of edition in the **print** age



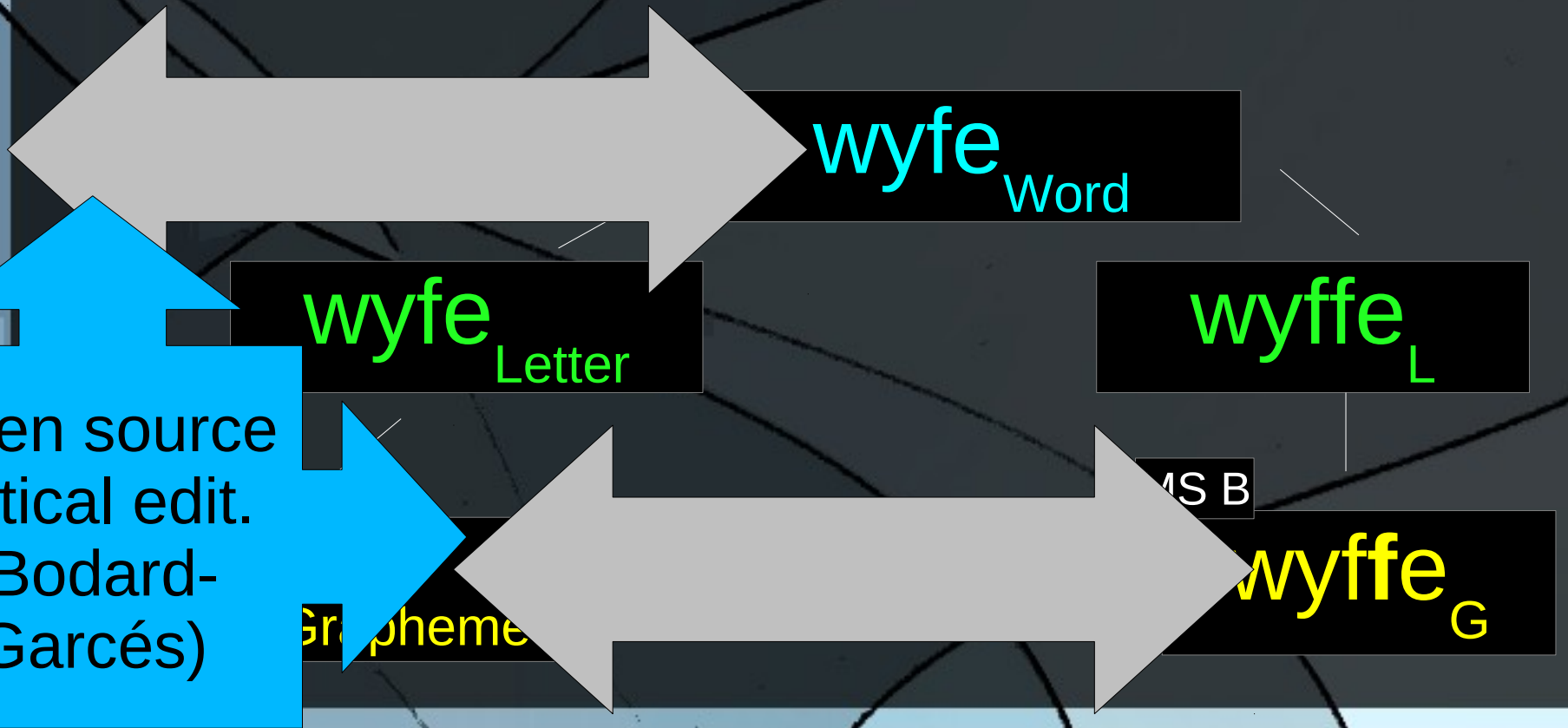
Comparing texts

- **Digital** edition



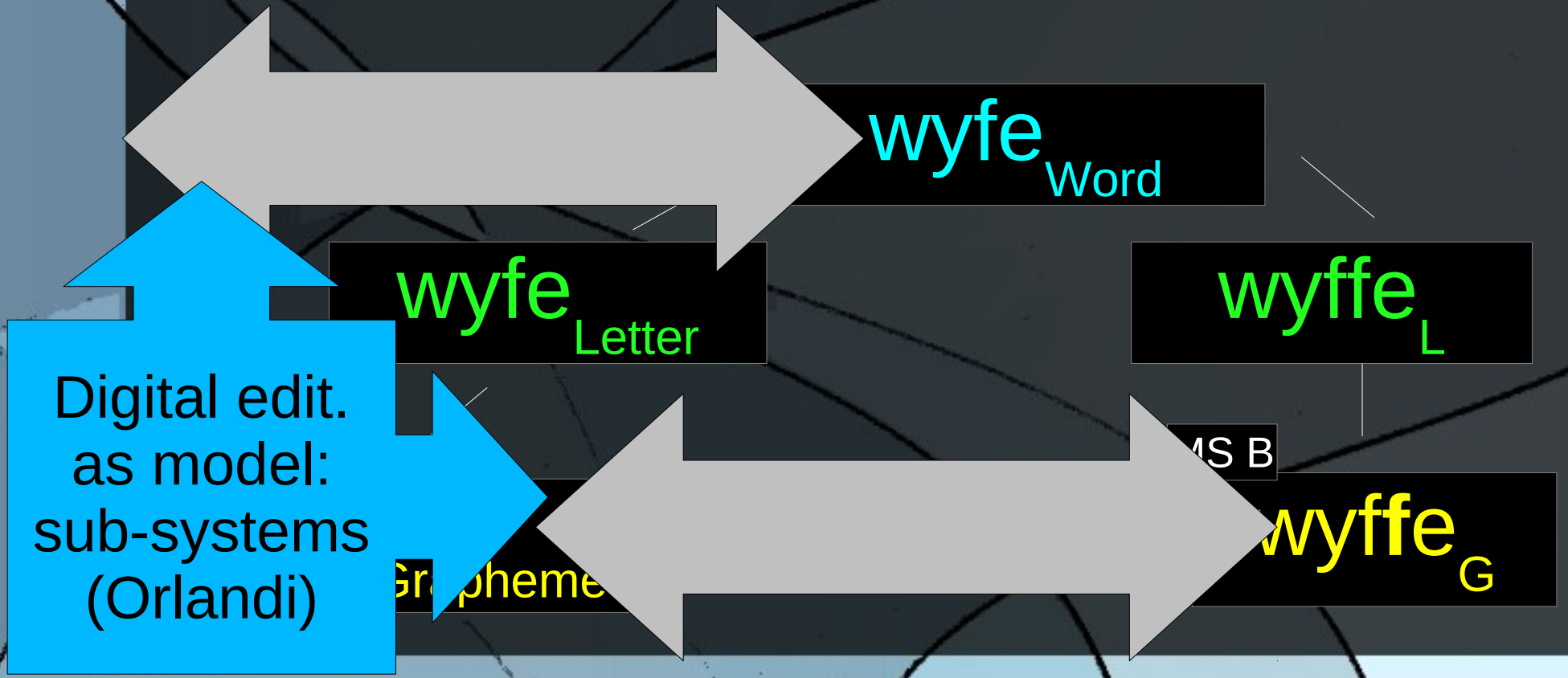
Comparing texts

- **Digital** edition



Comparing texts

- **Digital** edition

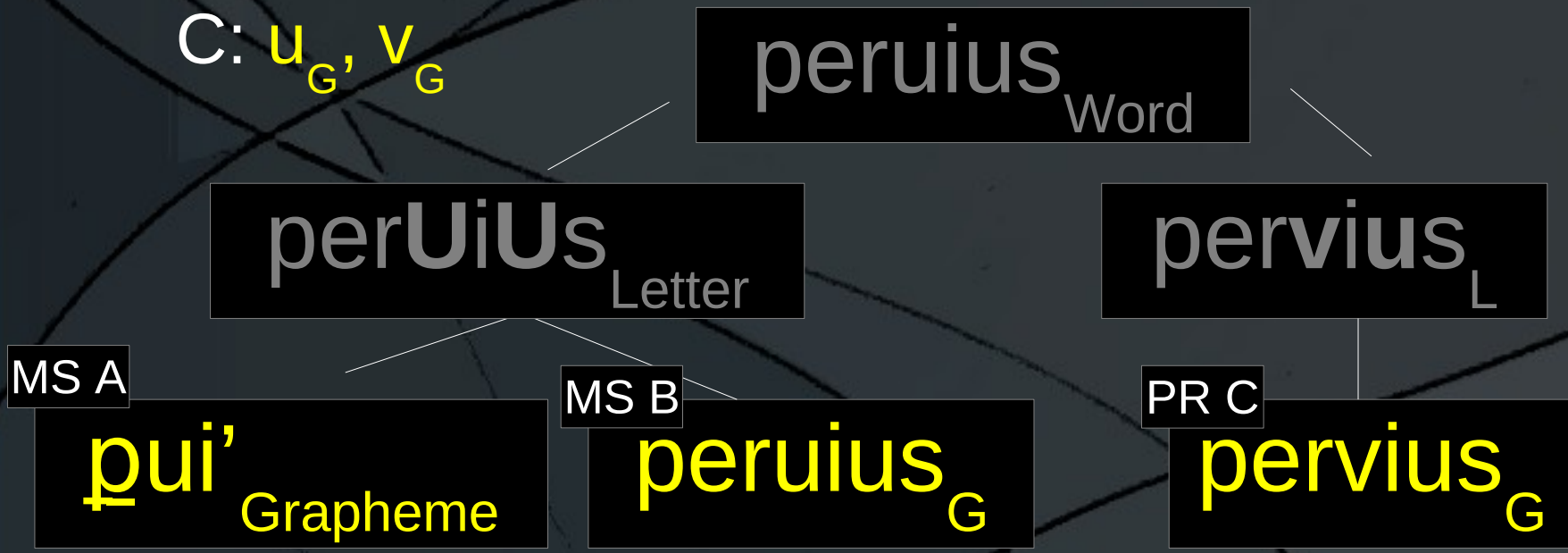


Outline

- The 'Babel issue'
  - The 'Babel issue' in a nutshell
  - Comparing texts
  - In the Tower
- Two issues
  - A. Comparison at the linguistic layer
  - B. Comparison at the graphemic layer

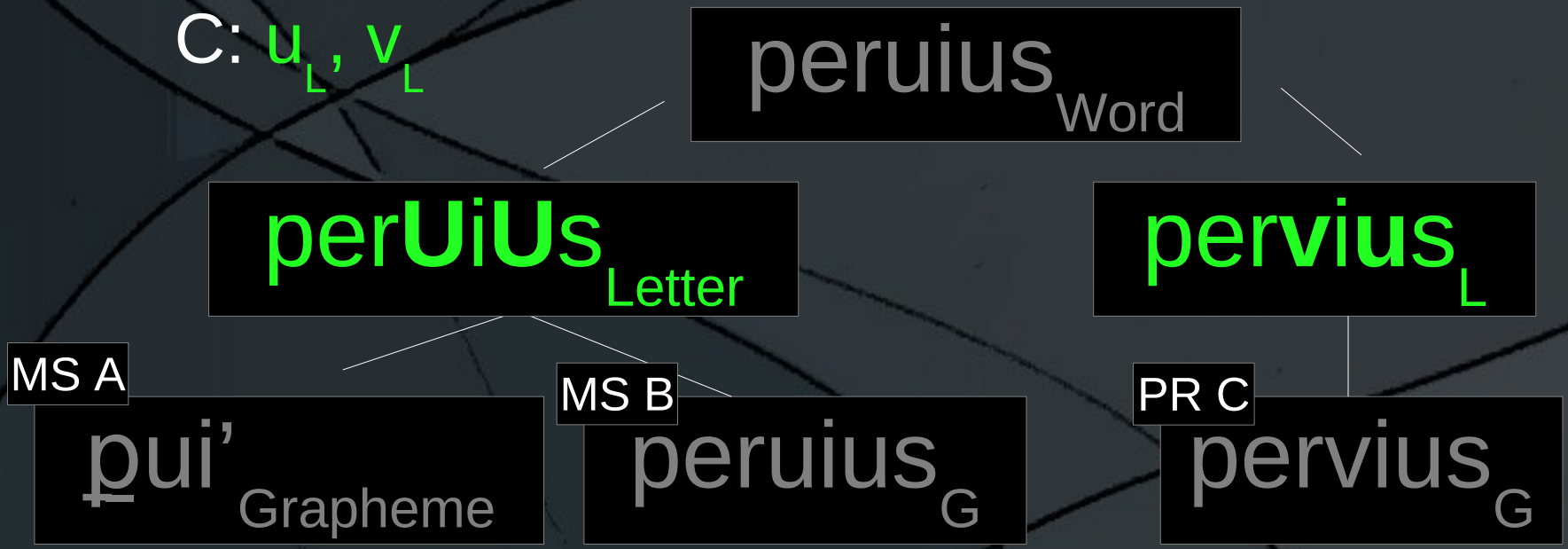
In the Tower

- Different Writing Systems (Graphemes):
  - A: $\underline{p}_G$, u_G (no u_G/v_G distinction)
  - B: u_G (no u_G/v_G distinction)
  - C: u_G, v_G



In the Tower

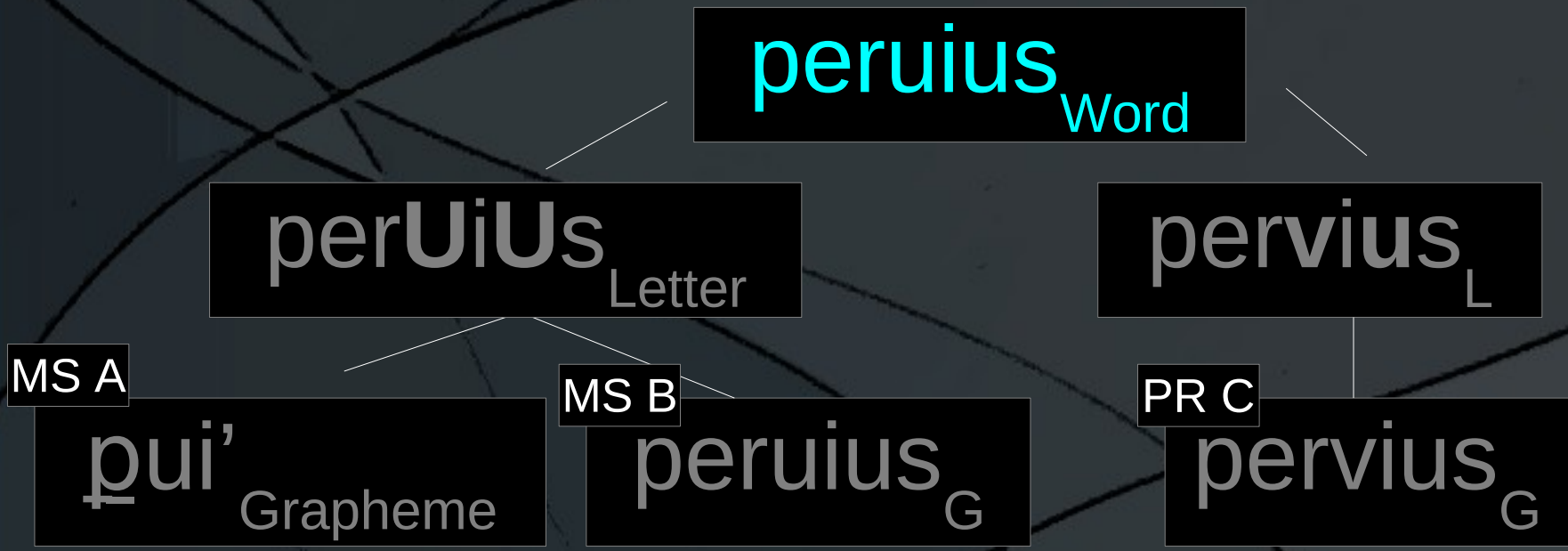
- Different Alphabets (Letters):
  - A: u_L (no u_L/v_L distinction)
  - B: u_L (no u_L/v_L distinction)
  - C: u_L, v_L



In the Tower

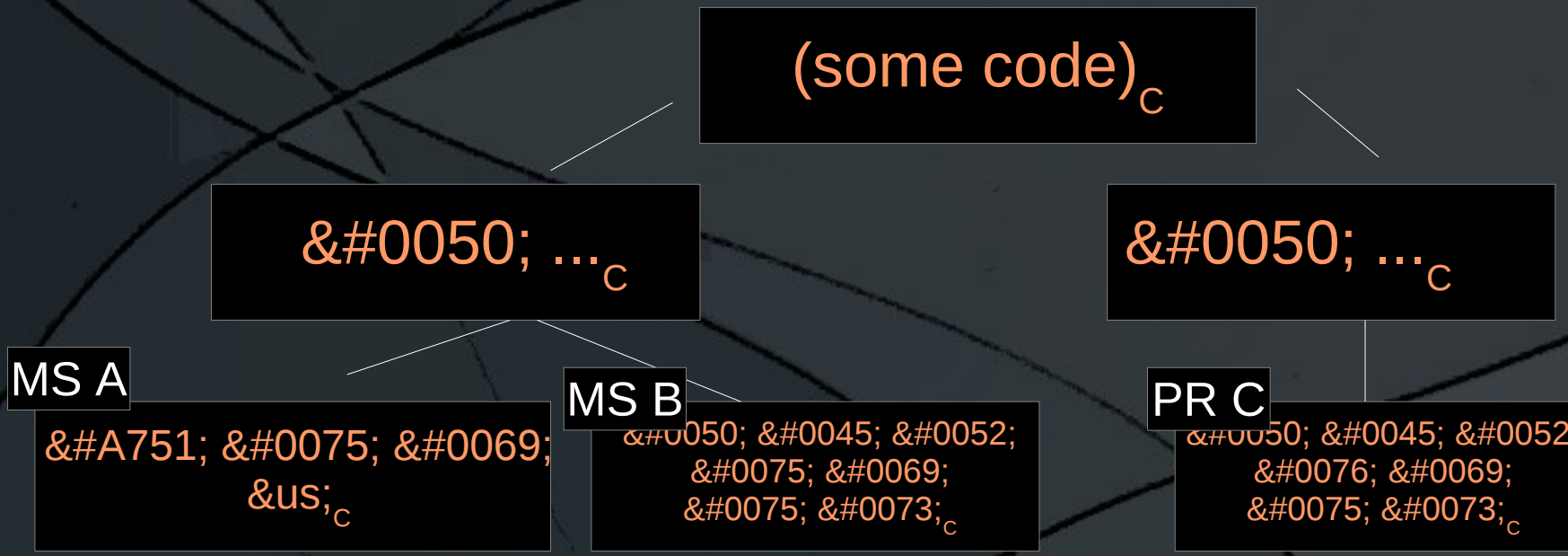
- Same text at Linguistic layer (Inflected word)

  “Perius”, nom. sing. masc. of lemma “perius, -a, um”



In the Tower

- Digital edition: Formalisation



In the Tower

- Life in the Tower of Babel (recapitulation):
  - Each builder, a language
    - Each primary source, a semiotic system
  - ...yet we want builders to interact
    - Textual Criticism
    - Processing (e. g. cross-corpus search)
  - ...at different layers
    - Graphs, allographs, graphemes, linguistic...
  - … all the sudden, all builders are robots!
    - Formalisation (no human intuition)

Outline

- The 'Babel issue'
  - The 'Babel issue' in a nutshell
  - Comparing texts
  - In the Tower
- Two issues
  - A. Comparison at the linguistic layer
  - B. Comparison at the graphemic layer

Two issues

A. Comparison at linguistic layer
  - Substantial readings ($wyfe_W \neq lyf_W$)
  - Critical edition
  - The 'text' (Inflected words)

B. Comparison at graphemic layer
  - Accidentals ($wyfe_G \neq wyffe_G$)
  - Diplomatic edition
  - The 'spelling' (orthography, Graphemes)

Two issues

A. Comparison at linguistic layer

- How do you identify elements at linguistic layer (Inflected words) digitally?
  - The Canterbury Tales Project: "Regularized spelling"
  - Tito Orlandi: Linguistic entities

Two issues

B. Comparison at graphemic layer

- Can you compare graphemes through different graphemic systems?
  - The Canterbury Tales Project: Unicode; Corpus-wide modelling of the graphemic system
  - Tito Orlandi: MS-wide complete modelling of the graphemic system (and of other systems at other textual layers)

Two issues

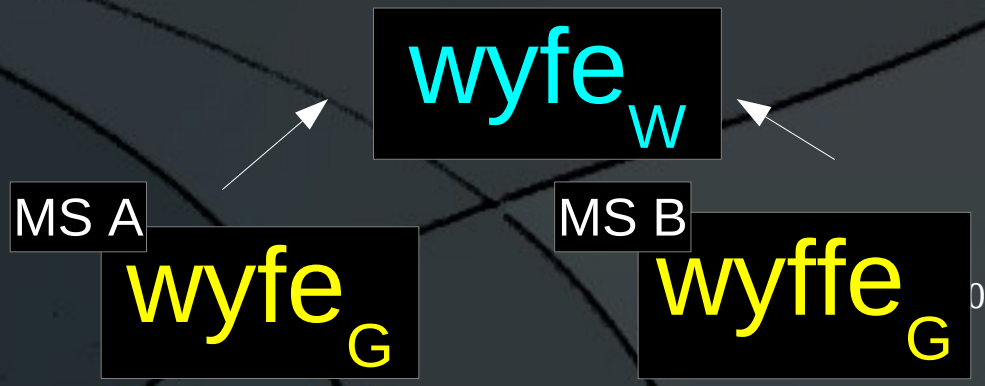
- The Canterbury Tales Project
- Tito Orlandi, *Informatica Testuale*, Laterza: Roma 2010
- My own project of an experimental edition from the *Anthologia Latina*

Outline

- The 'Babel issue'
  - The 'Babel issue' in a nutshell
  - Comparing texts
  - In the Tower
- Two issues
  - A. Comparison at the linguistic layer
  - B. Comparison at the graphemic layer

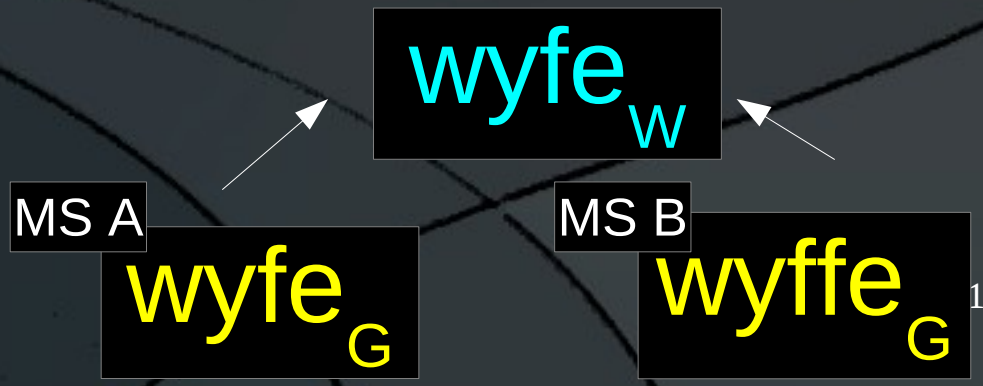
A. Comparison at linguistic layer

- Why comparing Words (at linguistic layer)? The Canterbury Tales Project:
  - Textual criticism
    - Relations between MSS
  - Processing
    - Indexing (“Spelling database”)



A. Comparison at linguistic layer

- From Grapheme to Word, not Lemma
  - Inflected word ($wyfe_w \neq wyves_w$)
    - Not as element of language (*langue*)
    - But as as element of that text (*parole*)
  - “Substantial reading”
  - Critical edition

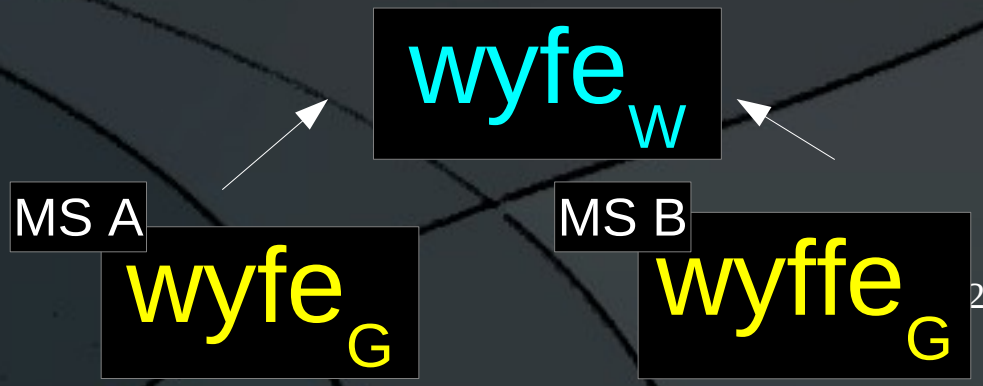


A. Comparison at linguistic layer

- From Grapheme to Word: semi-automatic
  - Not computable so far
  - Linguistic competence required
    - Grapheme → Letter: ambiguous graphemes

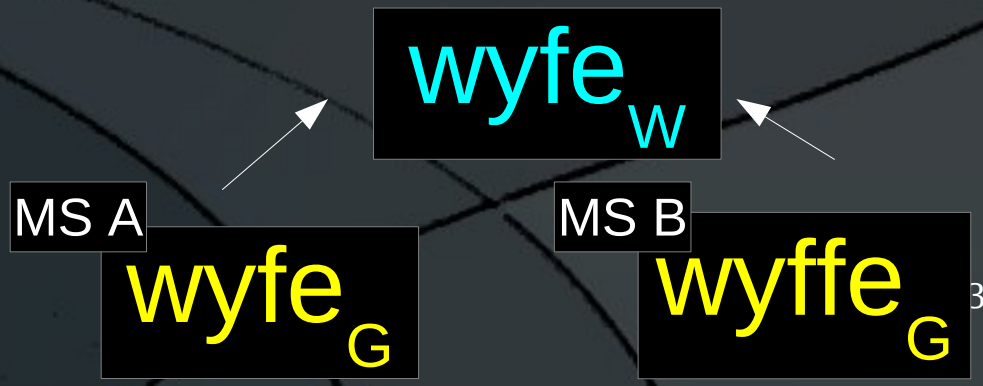
      lon_G → Jon_L / lon_L → Jon_W / lon_W

      $\underline{p}do_G$ → $perdo_L$ / $prodo_L$ → $perdo_W$ / $prodo_W$



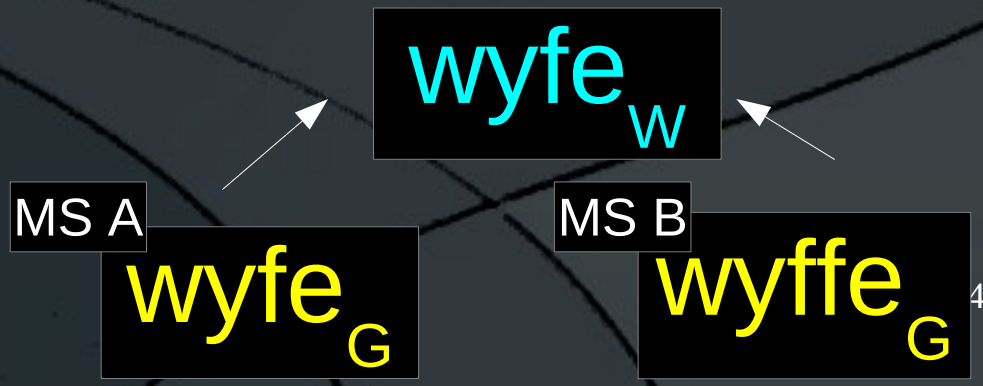
A. Comparison at linguistic layer

- From Grapheme to Word: semi-automatic
  - Not computable so far
  - Linguistic competence required
    - Letter → Word: omographs
      est_G → est_L → est (she is)$_W$ / est (she eats)$_W$



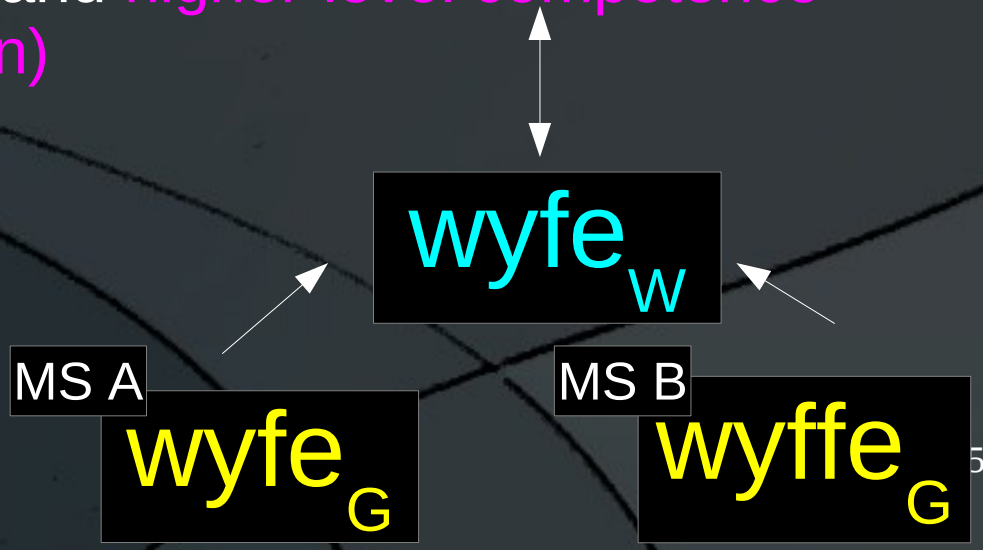
A. Comparison at linguistic layer

- From Grapheme to Word: semi-automatic
  - Not computable so far
  - Linguistic competence required
    - Letter → Word: different spellings
      $wyfe_w$ / $wyffe_w$ → $wyfe_w$ / $wyffe_w$ → $wyfe_w$



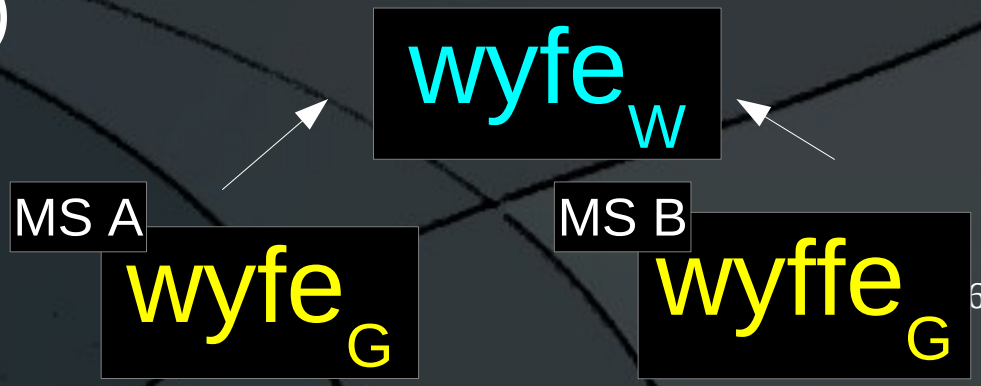
A. Comparison at linguistic layer

- From Grapheme to Word: semi-automatic
  - Not computable so far
  - Linguistic competence required
  - 'Context': interpretation of the whole text
    - Linguistic and higher-level competence (Jon / Ion)



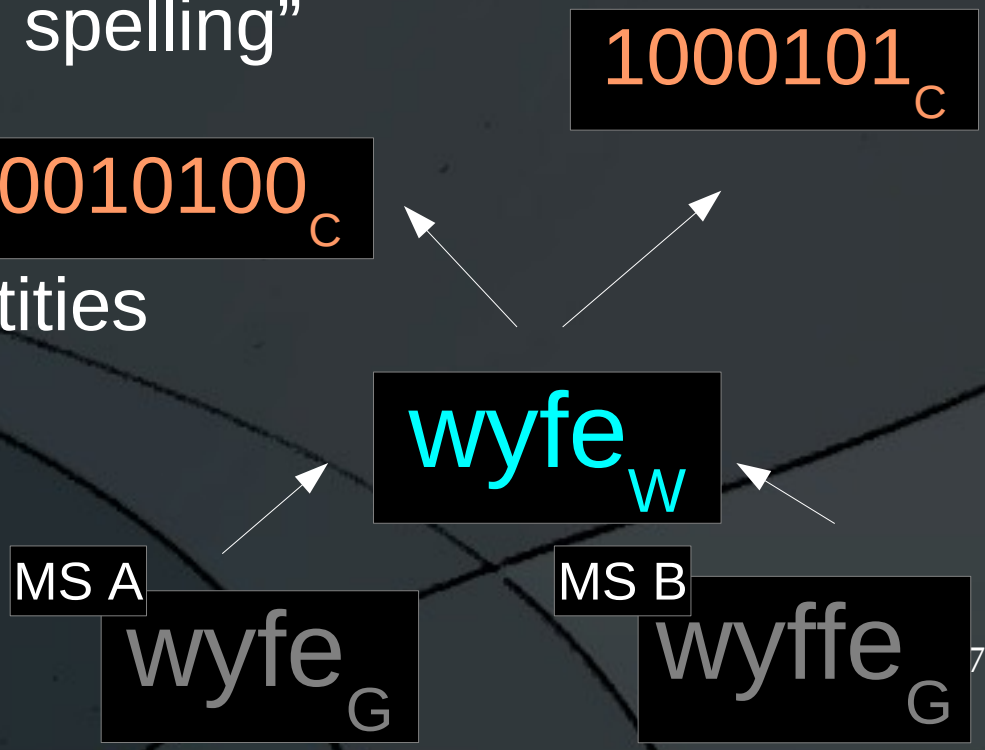
A. Comparison at linguistic layer

- From Grapheme to Word: semi-automatic
  - Semi-automatic procedures (human-driven, computer-assisted)
  - "The computer collation program we are using (Collate) permits regularization as part of the collation process” [...] “regularized spelling" (Robinson-Solopova →)



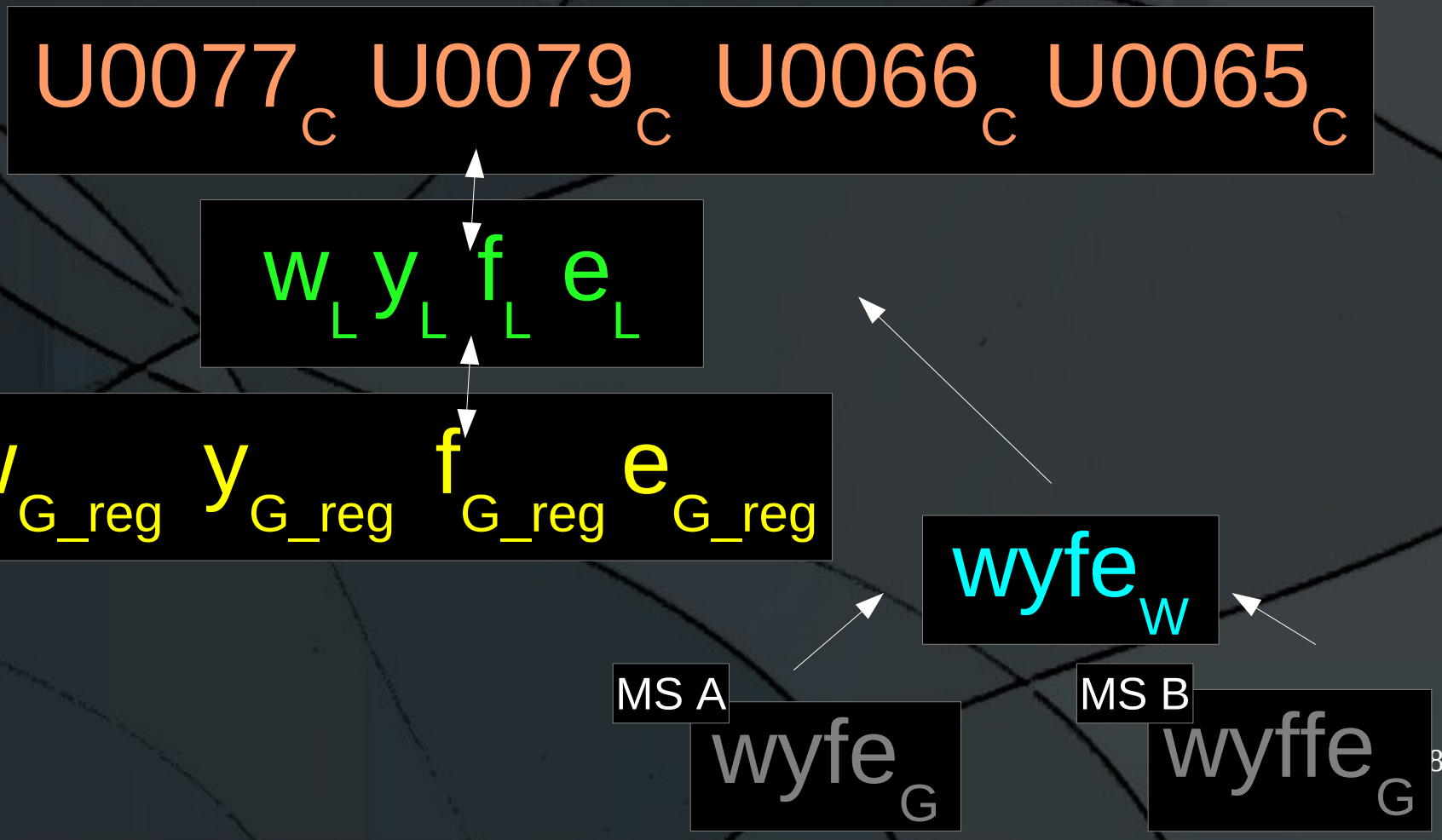
A. Comparison at linguistic layer

- How do you identify elements at linguistic layer (Inflected words) digitally?
  - The Canterbury Tales Project: "Regularized spelling"
  - Tito Orlandi: Linguistic entities



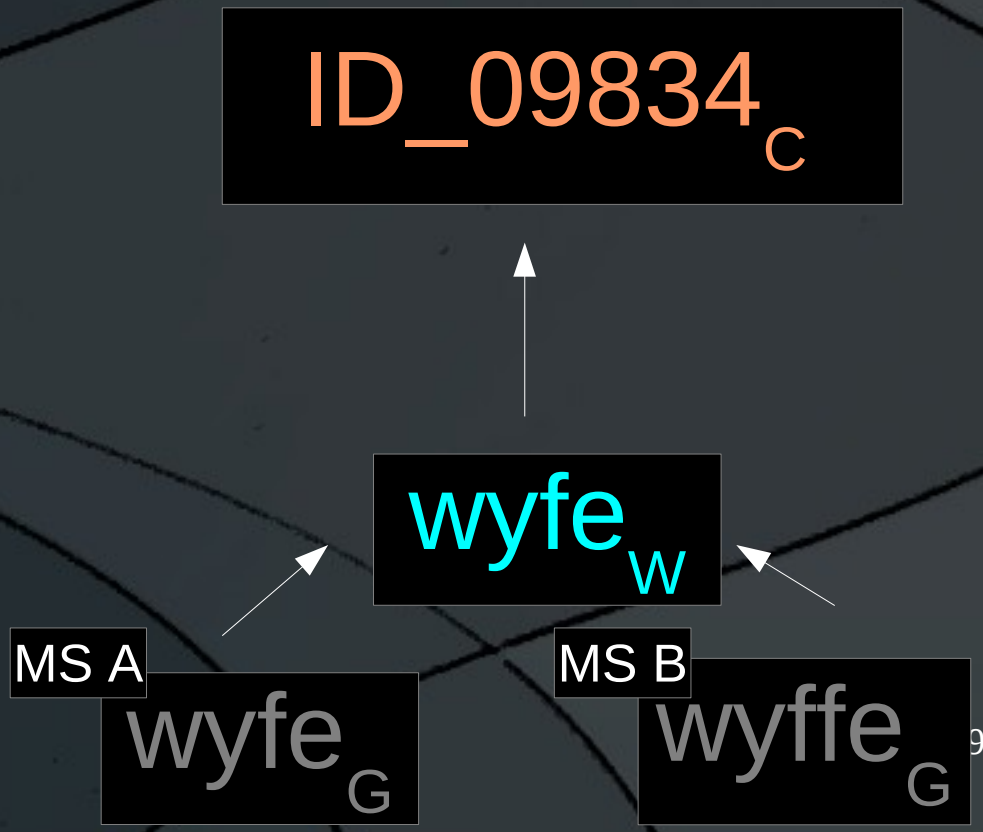
A. Comparison at linguistic layer

- Canterbury: “Regularized spelling”



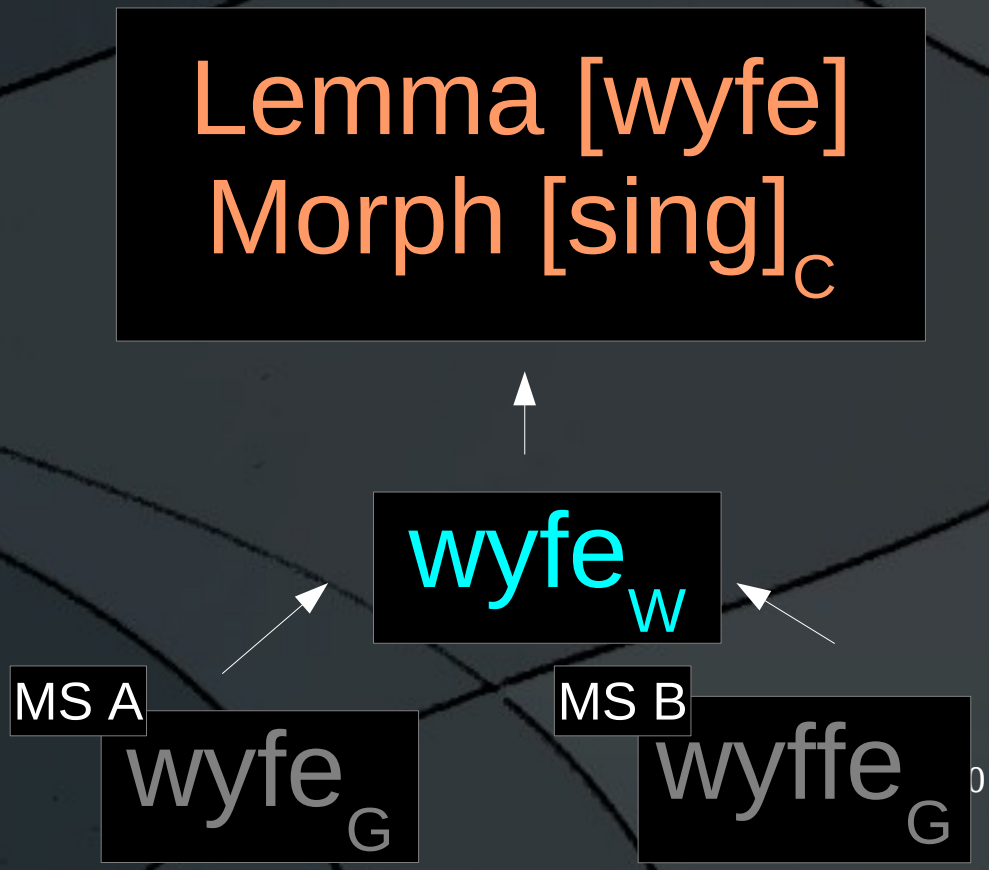
A. Comparison at linguistic layer

- Orlandi: Discrete linguistic entities in a database



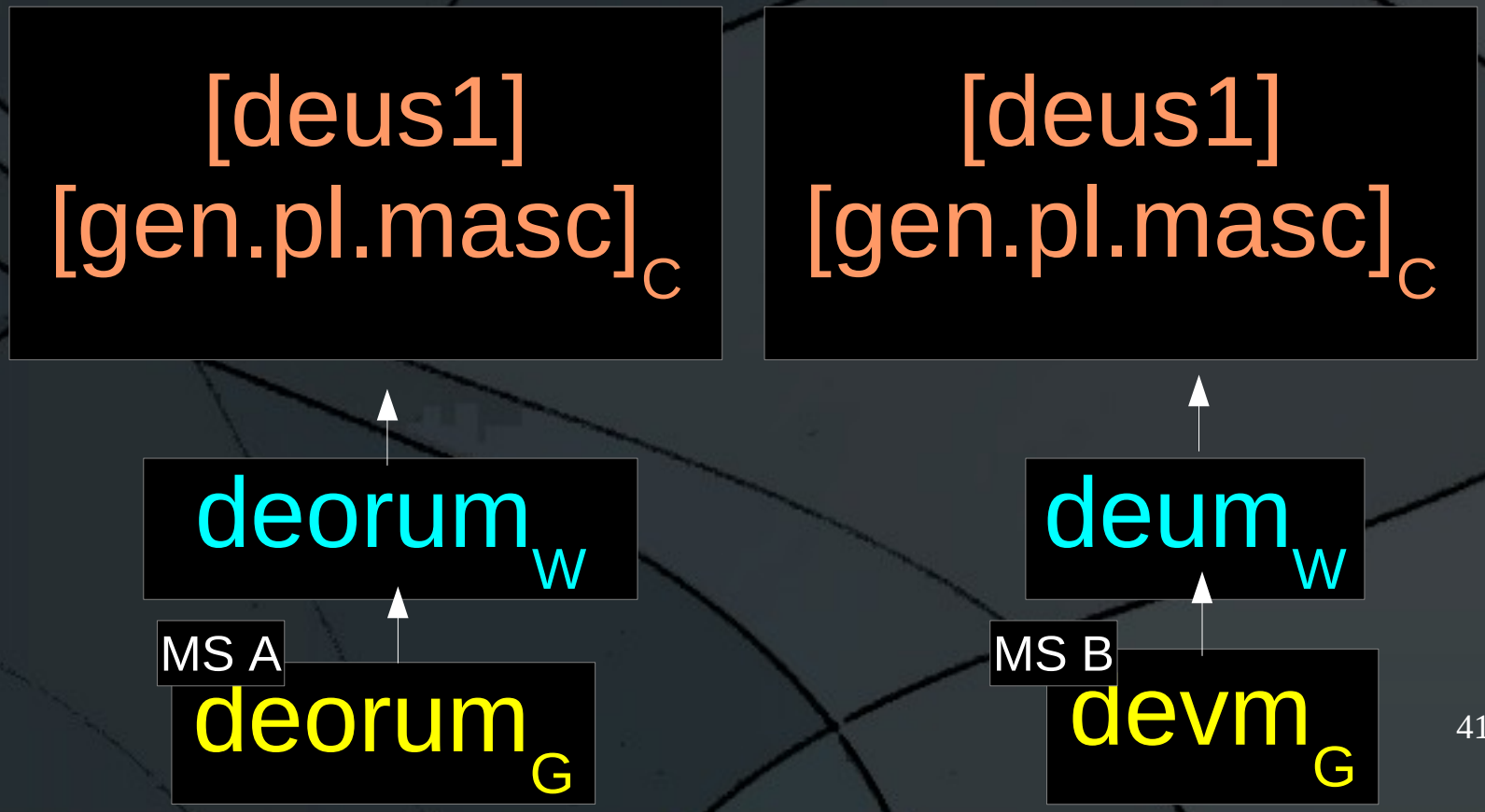
A. Comparison at linguistic layer

- Orlandi: Discrete linguistic entities in a database → How do you identify them?



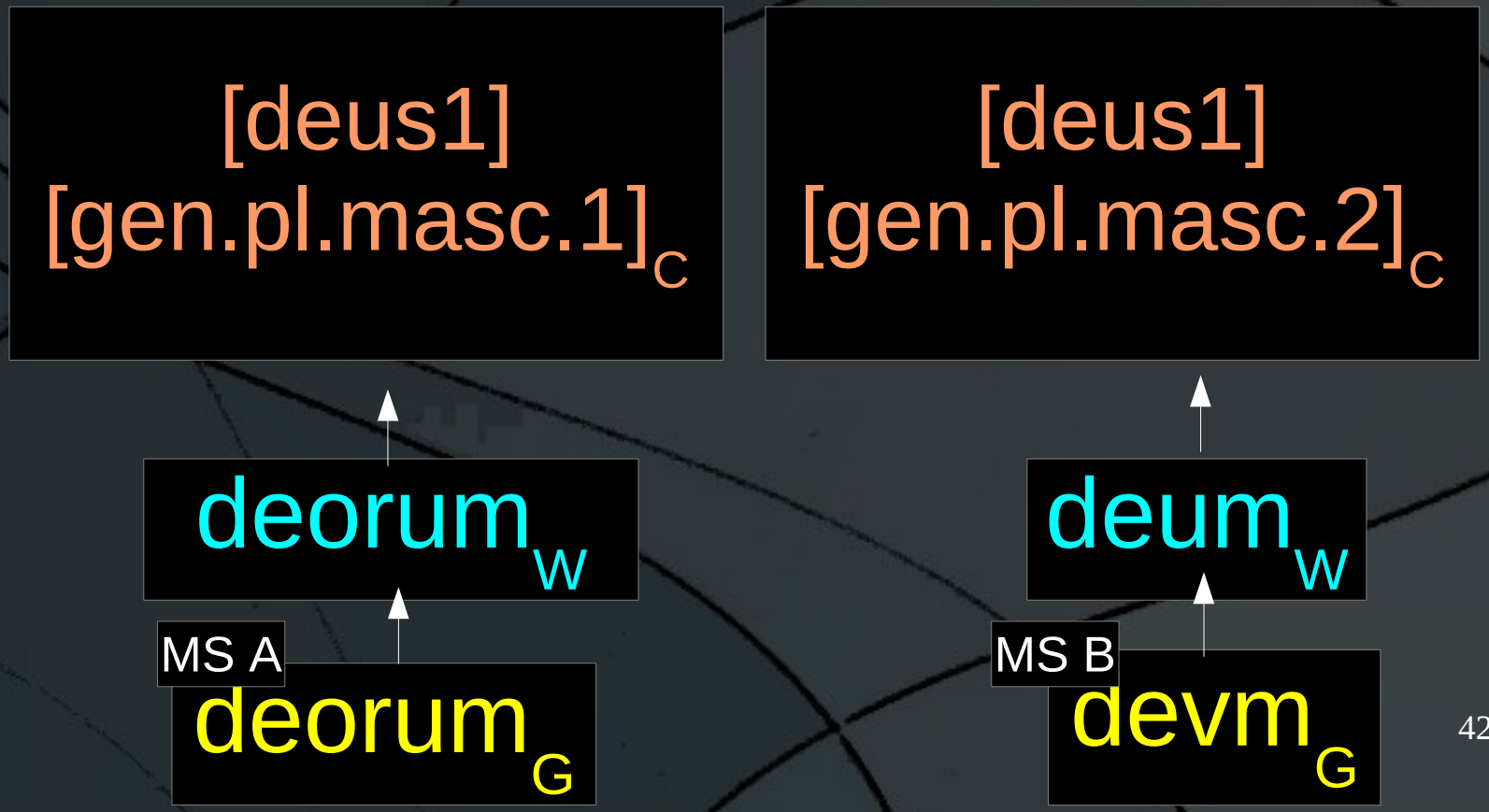
A. Comparison at linguistic layer

- Orlandi: Discrete linguistic entities in a database → How do you identify them?



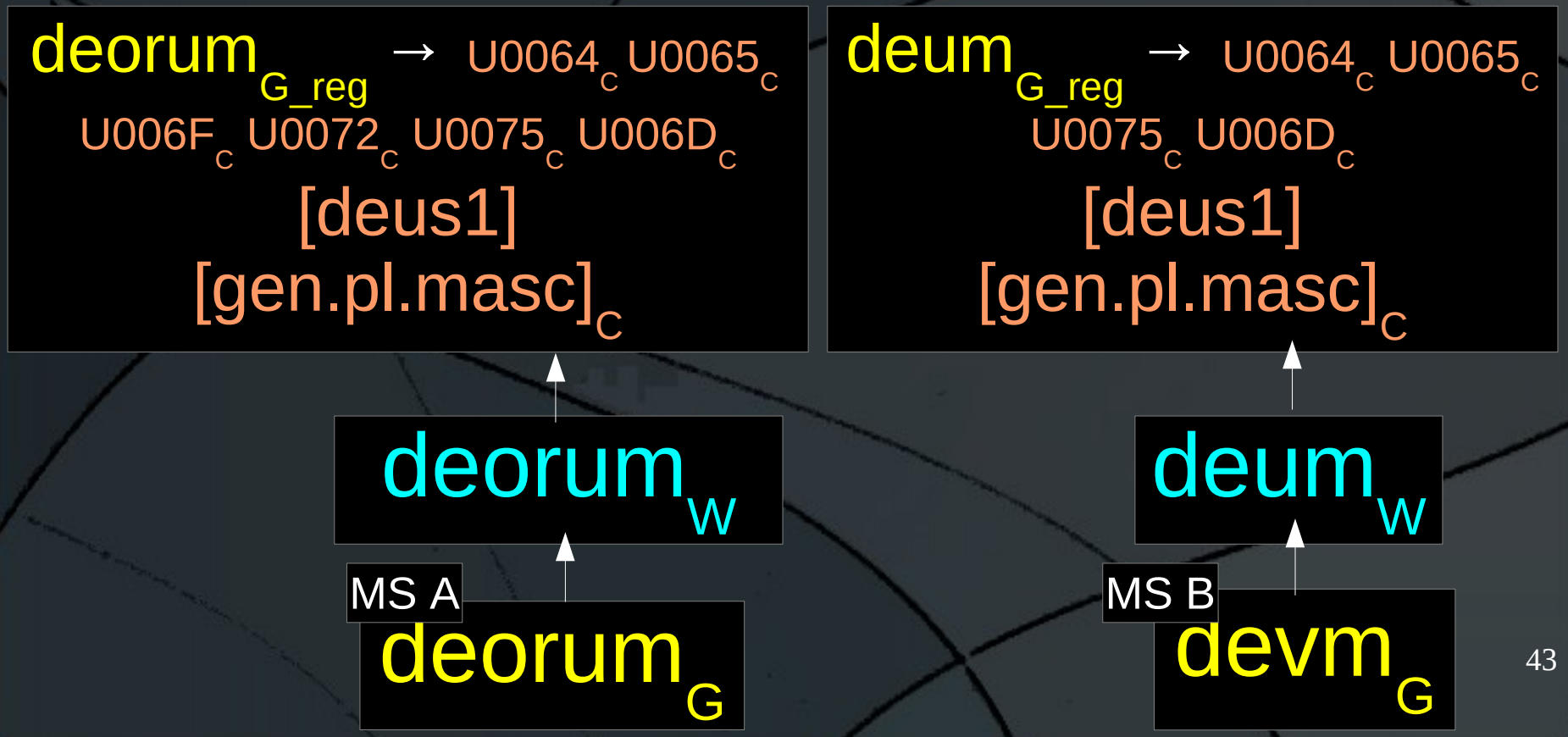
A. Comparison at linguistic layer

- Orlandi: Discrete linguistic entities in a database → How do you identify them?



A. Comparison at linguistic layer

- Orlandi: Discrete linguistic entities in a database → How do you identify them?



A. Comparison at linguistic layer

- Orlandi/Monella: is it worth it?

Canterbury

$deum_{G\_reg}$

Orlandi/Monella

$deum_{G\_reg}$

[deus1] [gen.plur.masc]$_C$

$deum_W$

$devm_G$

A. Comparison at linguistic layer

Canterbury, MS A

$deum_{G\_reg}$

=

Canterbury, MS B

$deum_{G\_reg}$

Orlandi/Monella, MS A

$deum_{G\_reg}$ [deus1] [gen.plur.masc]$_C$

≠

Orlandi/Monella, MS B

$deum_{G\_reg}$ [deus1] [acc.sing.masc]$_C$

MS A] *Pater devm*

MS B] Timeo devm

$deum_W$ ≠ $deum_W$

A. Comparison at linguistic layer

- I'ts only worth it
  - if words *are not* their regularised spelling, i. e.
  - if homograph words exist

deum$_w$	≠	deum$_w$
Pater deum		*Timeo deum*

A. Comparison at linguistic layer

- I'ts only worth it
  - if substantial readings *are not* their regularised spelling, i. e.
  - if homograph substantial readings exist (do they? → Concept of “reading”)

MS A *Pater, devm Neptvnvm odi*

MS B *Pater deum, Neptunum odi*

deum ≠ *deum?*

A. Comparison at linguistic layer

Canterbury, MS A

$deum_{G\_reg}$

=

Canterbury, MS B

$deum_{G\_reg}$

Orlandi/Monella, MS A

$deum_{G\_reg}$
[deus1] [gen.plur.masc]$_C$

≠

Orlandi/Monella, MS B

$deum_{G\_reg}$
[deus1] [acc.sing.masc]$_C$

MS A

Pater,
devm Neptvnvm odi

MS B

Pater deum,
Neptunum odi

deum ≠ *deum?*

A. Comparison at linguistic layer

Canterbury, MS A

$deum_{G\_reg}$

=

Canterbury, MS B

$deum_{G\_reg}$

Orlandi/Monella, MS A

$deum_{G\_reg}$
[deus1] [gen.plur.masc]$_C$

≠

Orlandi/Monella, MS B

$deum_{G\_reg}$
[deus1] [acc.sing.masc]$_C$

MS A

Pater,
devm Neptvnvm odi

MS B

Pater deum,
Neptunum odi

deum = *deum?*

A. Comparison at linguistic layer

- Normally done via regularised spelling
- Inflected words are not their regularised spelling (omographs)
- Are there omograph substantial readings (*deum*)?
- If so:
  - No regularised spelling
  - But formal identifiers for inflected words (spelling, lemma, identifier)

Outline

- The 'Babel issue'
  - The 'Babel issue' in a nutshell
  - Comparing texts
  - In the Tower
- Two issues
  - A. Comparison at the linguistic layer
  - B. Comparison at the graphemic layer

B. Comparison at graphemic layer

- Why comparing spellings (at graphemic layer)?
  - Historical Linguistics
    - Evidence for lexical, morphological and phonetic evolution

B. Comparison at graphemic layer

- Why comparing spellings (at graphemic layer)?
  - Palaeography
    - Indirect evidence for letters (Benskin: letter þ *thorn* →)
      - Overlapping with Historical Linguistics (Emiliano: “Scripto-linguistic change” →)
    - Though only graphemes
      - Not allographs (graphetes, s/ſ) or graphs (bitmap)

B. Comparison at graphemic layer

- Why comparing spellings (at graphemic layer)?
  - Textual Criticism
    - 'Orthographic' apparatus (The Hengwrt Chaucer Digital Facsimile, Collation function →)
    - “Although for most manuscripts collation of the regularized text will produce sufficient information **to place those manuscripts in genetic relation to one another** […]” (Robinson-Solopova →)

B. Comparison at graphemic layer

- *The loue of love*
- *Loue:* “A hill or mountain” ≠ *Love:* “Love”
- Middle English alphabet: u ≠ v

B. Comparison at graphemic layer

a_L

u_L

v_L

The loue of love

MS A

The loue of loue

MS B

B. Comparison at graphemic layer

a_L
u_L
v_L

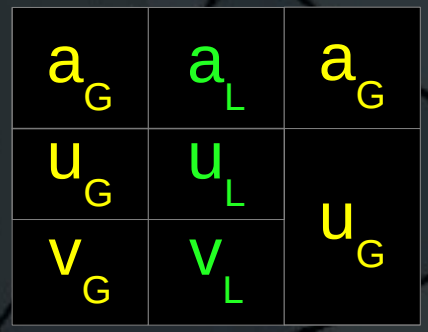
The loue of love$_G$

MS A

The loue of loue$_G$

MS B

B. Comparison at graphemic layer



The $lou_G e$ of $lov_G e$

The $lou_G e$ of $lou_G e$

MS A

MS B

B. Comparison at graphemic layer

a_G	a_L	a_G
u_G	u_L	u_G
v_G	v_L	

- Back in the Tower

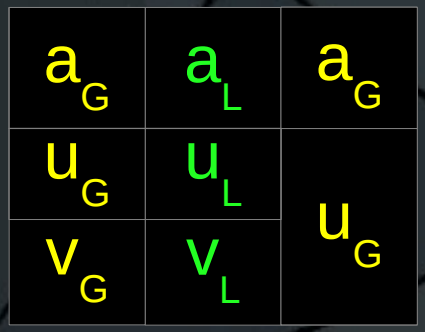
The lou_Ge of lov_Ge

MS A

The lou_Ge of lou_Ge

MS B

B. Comparison at graphemic layer



- Back in the Tower
  - "Grapheme. (1) A minimally distinctive unit of writing <u>in the context of a particular writing system</u>" (Unicode Glossary →).

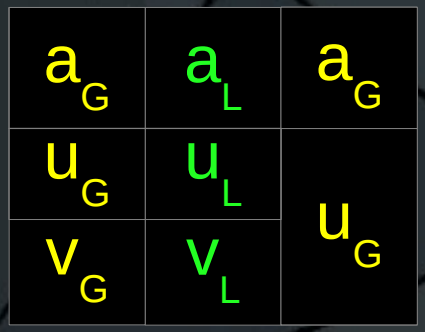
The lou$_G$e of lov$_G$e

MS A

The lou$_G$e of lou$_G$e

MS B

B. Comparison at graphemic layer



- Back in the Tower
  - De Saussure: relational nature of signs *within* a semiotic system

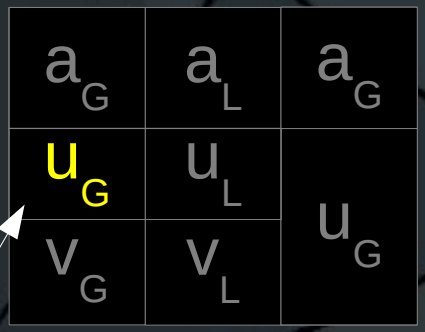
The lou$_G$e of lov$_G$e

MS A

The lou$_G$e of lou$_G$e

MS B

B. Comparison at graphemic layer



- Back in the Tower
  - De Saussure: relational nature of signs *within* a semiotic system
  - This is a different grapheme than this, as the former is in contrast with this, while the latter is not

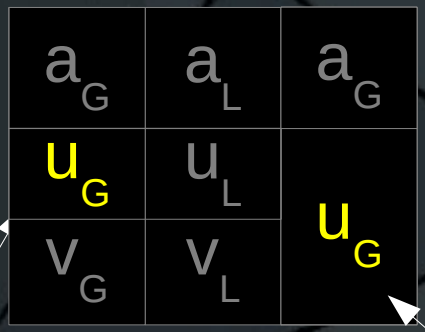
The lou$_G$e of lov$_G$e

MS A

The lou$_G$e of lou$_G$e

MS B

B. Comparison at graphemic layer



- Back in the Tower
  - De Saussure: relational nature of signs *within* a semiotic system
  - This is a different grapheme than this, as the former is in contrast with this, while the latter is not

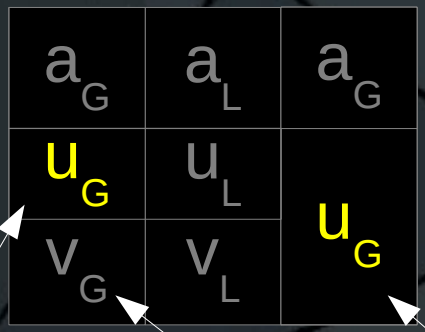
The lou$_G$e of lov$_G$e

MS A

The lou$_G$e of lou$_G$e

MS B

B. Comparison at graphemic layer



- Back in the Tower
  - De Saussure: relational nature of signs *within* a semiotic system
  - This is a different grapheme than this, as the former is in contrast with this, while the latter is not

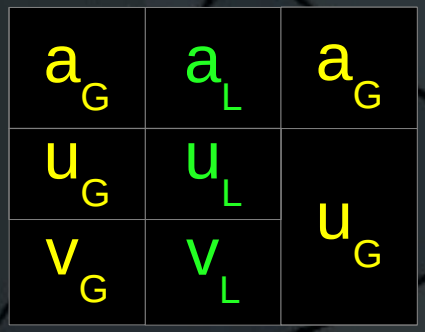
The lou$_G$e of lov$_G$e

MS A

The lou$_G$e of lou$_G$e

MS B

B. Comparison at graphemic layer



- How do you encode graphemes digitally?

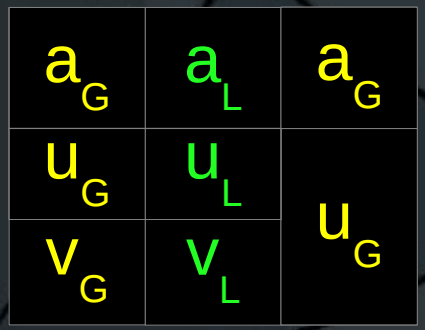
The $lou_G e$ of $lov_G e$

MS A

The $lou_G e$ of $lou_G e$

MS B

B. Comparison at graphemic layer



- TEI
  - Unicode → →
- The Canterbury Tales Project
  - Corpus-wide definition of the graphemic system →

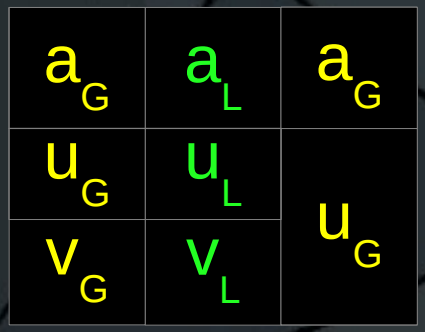
The lou$_G$e of lov$_G$e

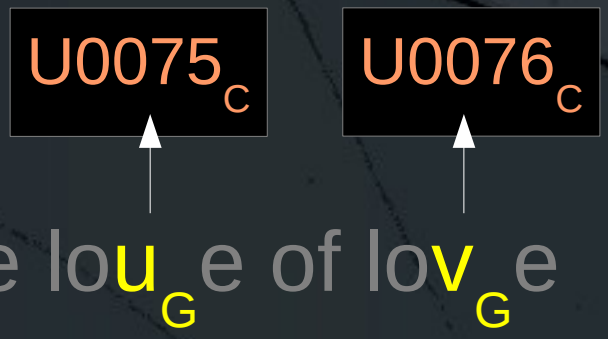
MS A

The lou$_G$e of lou$_G$e

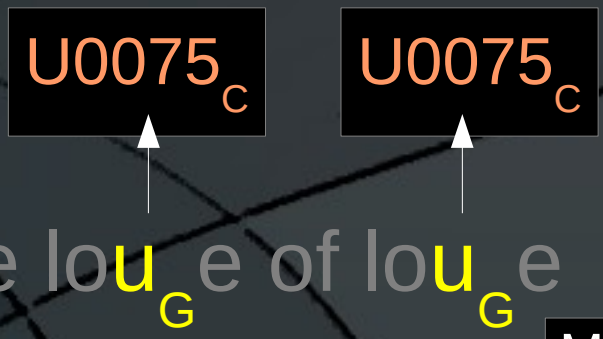
MS B

B. Comparison at graphemic layer



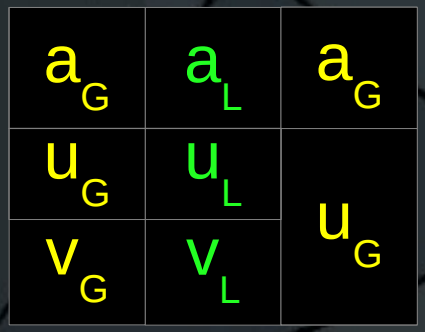


MS A



MS B

B. Comparison at graphemic layer



No match:
different graphemes

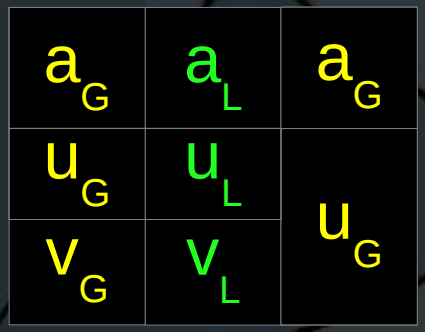
The lou$_G$e of lo**v**$_G$e

The lou$_G$e of lo**u**$_G$e

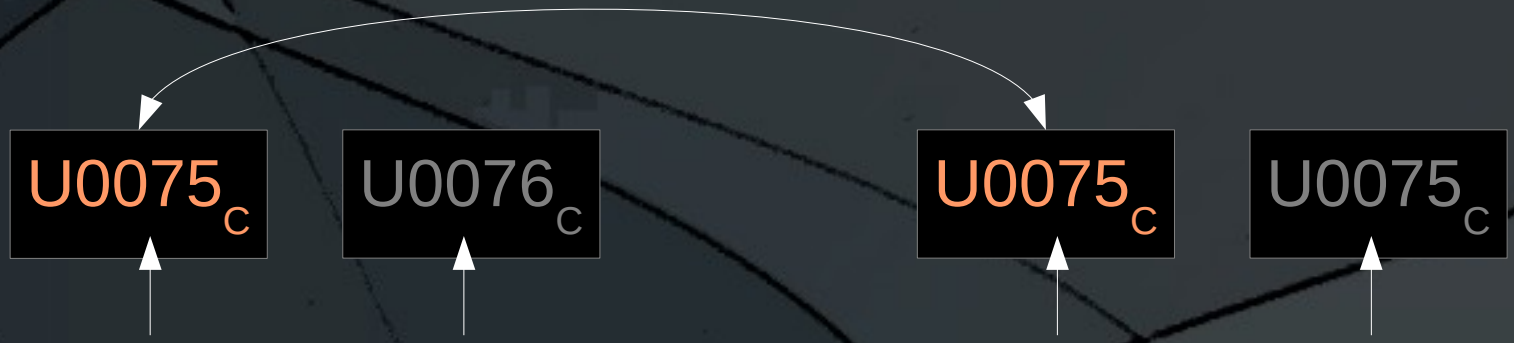
MS A

MS B

B. Comparison at graphemic layer



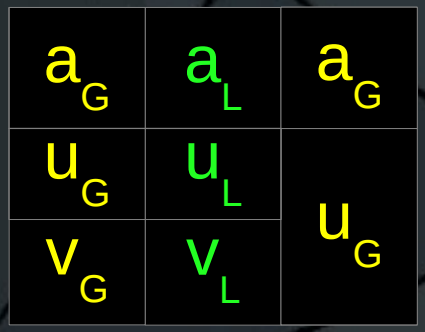
Match:
same grapheme



MS A

MS B

B. Comparison at graphemic layer



- Orlandi
  - *MS*-wide *complete* definition of graphemic system
    - Not corpus-wide (Canterbury)
    - Not world-wide (Unicode)

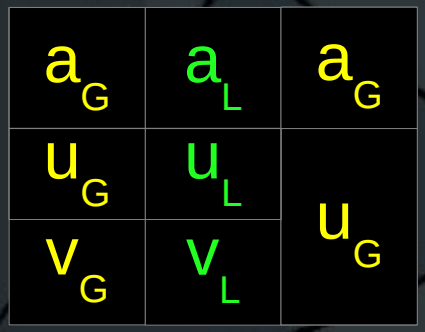
The louGe of lovGe

MS A

The louGe of louGe

MS B

B. Comparison at graphemic layer




```
<teiHeader>
    <encodingDesc>
        <charDecl>
            <char xml:id="uv">
            <glyph xml:id="long_s">
```

Graphemes

Allographs

The lou$_G$e of lov$_G$e

MS A

The lou$_G$e of lou$_G$e

MS B

B. Comparison at graphemic layer

a_G	a_L	a_G
u_G	u_L	u_G
v_G	v_L	

```
<body>
    <p>
        <g ref="uv" />
        <!-- or, better: -->
        <!ENTITY uv '<g ref="#uv" / >
        &uv;
```

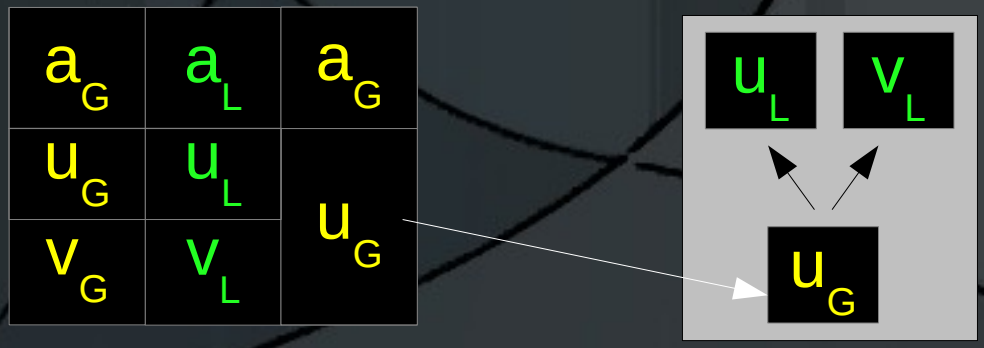
The lou_Ge of lov_Ge

MS A

The lou_Ge of lou_Ge

MS B

B. Comparison at graphemic layer



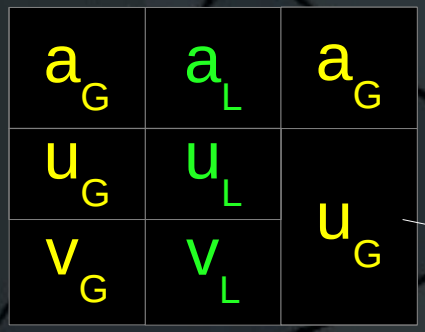
The lou_Ge of lov_Ge

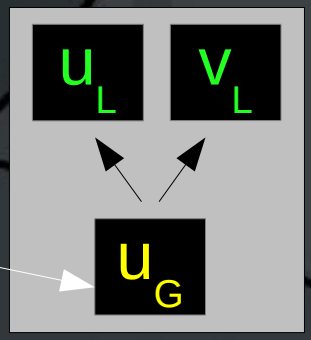
MS A

The lou_Ge of lou_Ge

MS B

B. Comparison at graphemic layer





MS A

```
<charDesc>
 <char xml:id= "a">
 <char xml:id= "u">
 <char xml:id= "v">
```

MS B

```
<charDesc>
 <char xml:id= "a">
 <char xml:id= "uv">
```

$U0075_C$ $U0076_C$

The lou_Ge of lov_Ge

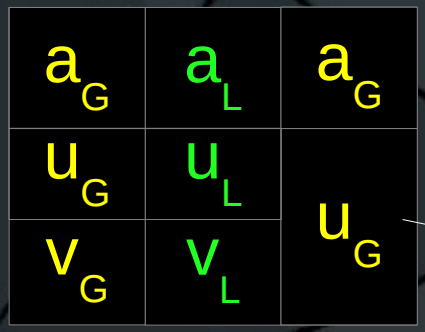
MS A

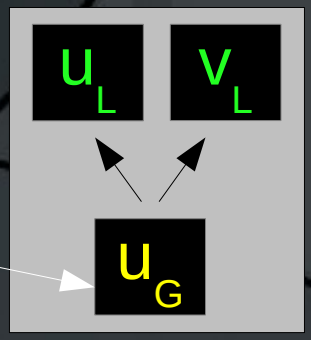
$\&uv_C$ $\&uv_C$

The lou_Ge of lou_Ge

MS B

B. Comparison at graphemic layer





MS B

```
<charDesc>
 <char xml:id= "uv">
  <charName>SMALL LATIN U OR V</charName>
  <desc>U-shaped when lowercase, V-shaped when
   uppercase. Content: either small letter
   Latin u or smal letter Latin v</desc>
```

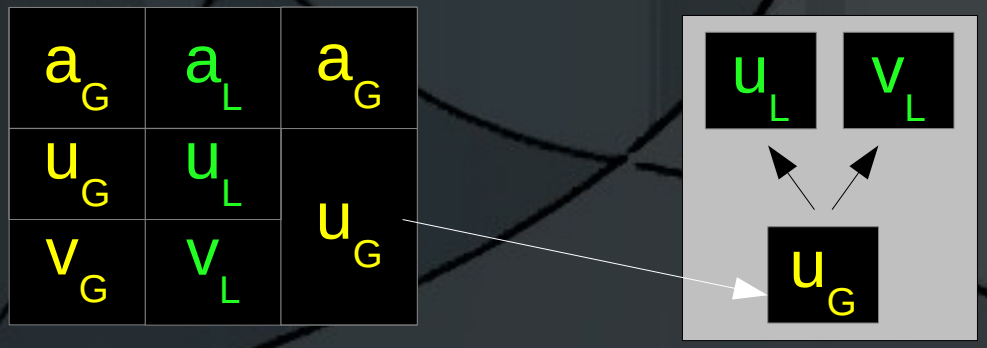
The $lou_G e$ of $lov_G e$

MS A

The $lou_G e$ of $lou_G e$

MS B

B. Comparison at graphemic layer



MS B

```
<charDesc>
 <char xml:id= "uv">
  <charProp>
    <localName>Expression</localName>
    <value>U+0075</value>
    <localName>Content</localName>
    <value>u|v</value>
```

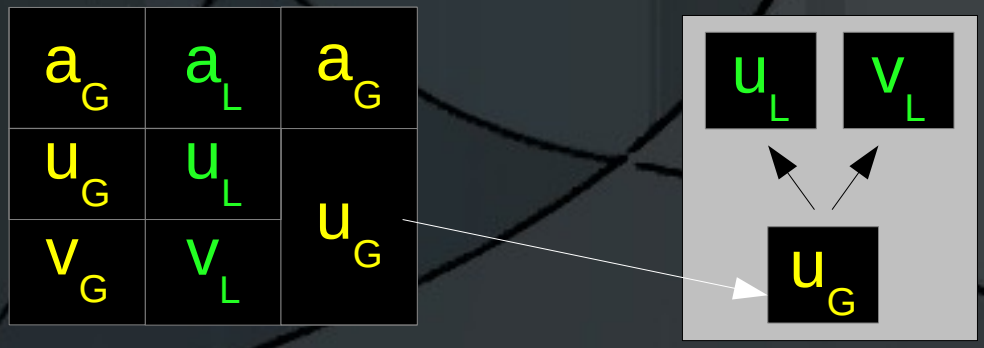
The lou$_G$e of lov$_G$e

MS A

The lou$_G$e of lou$_G$e

MS B

B. Comparison at graphemic layer



The $lou_G e$ of $lov_G e$

MS A

The $lou_G e$ of $lou_G e$

MS B

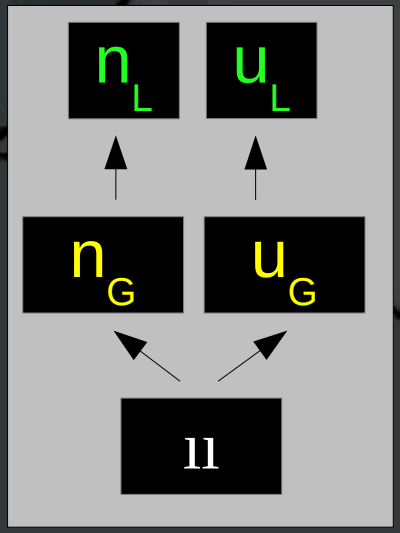
B. Comparison at graphemic layer

Two minims, lowercase

Expression (shape): indistinguishable.
Content (letter): n or u.

The reader does not identify the grapheme from its shape, but guessing its content.

Graphematic information is not conveyed by graphic information, but by linguistic information (context), so the scribe was confident in our Linguistic competence to tell the graphemes apart.



The $lou_G e$ of $lov_G e$

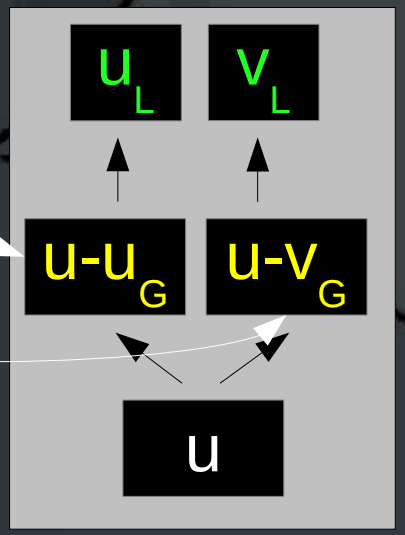
MS A

The $lou_G e$ of $lou_G e$

MS B

B. Comparison at graphemic layer

a_G	a_L	a_G
u_G	u_L	$u\text{-}u_G$
v_G	v_L	$u\text{-}v_G$



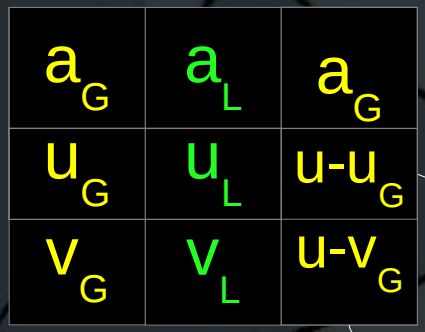
The lou_Ge of lov_Ge

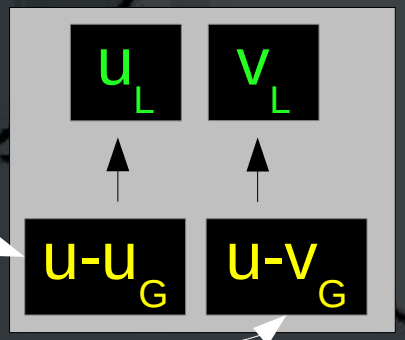
MS A

The lou_Ge of lou_Ge

MS B

B. Comparison at graphemic layer



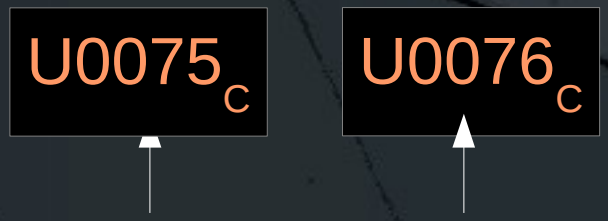


MS A

```
<charDesc>
 <char xml:id= "a">
 <char xml:id= "u">
 <char xml:id= "v">
```

MS B

```
<charDesc>
 <char xml:id= "a">
 <char xml:id= "u-u">
 <char xml:id= "u-v">
```



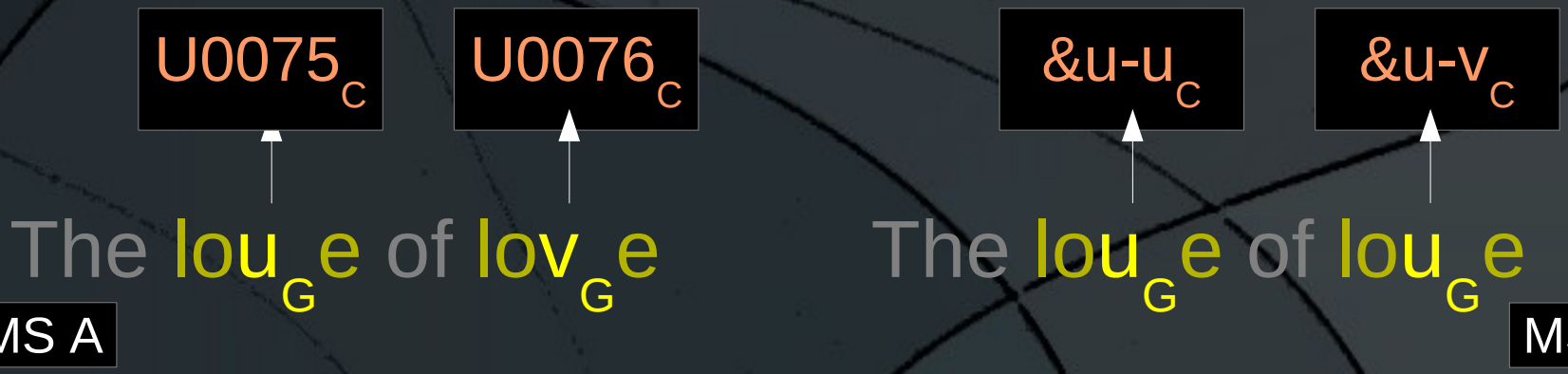
MS A

MS B

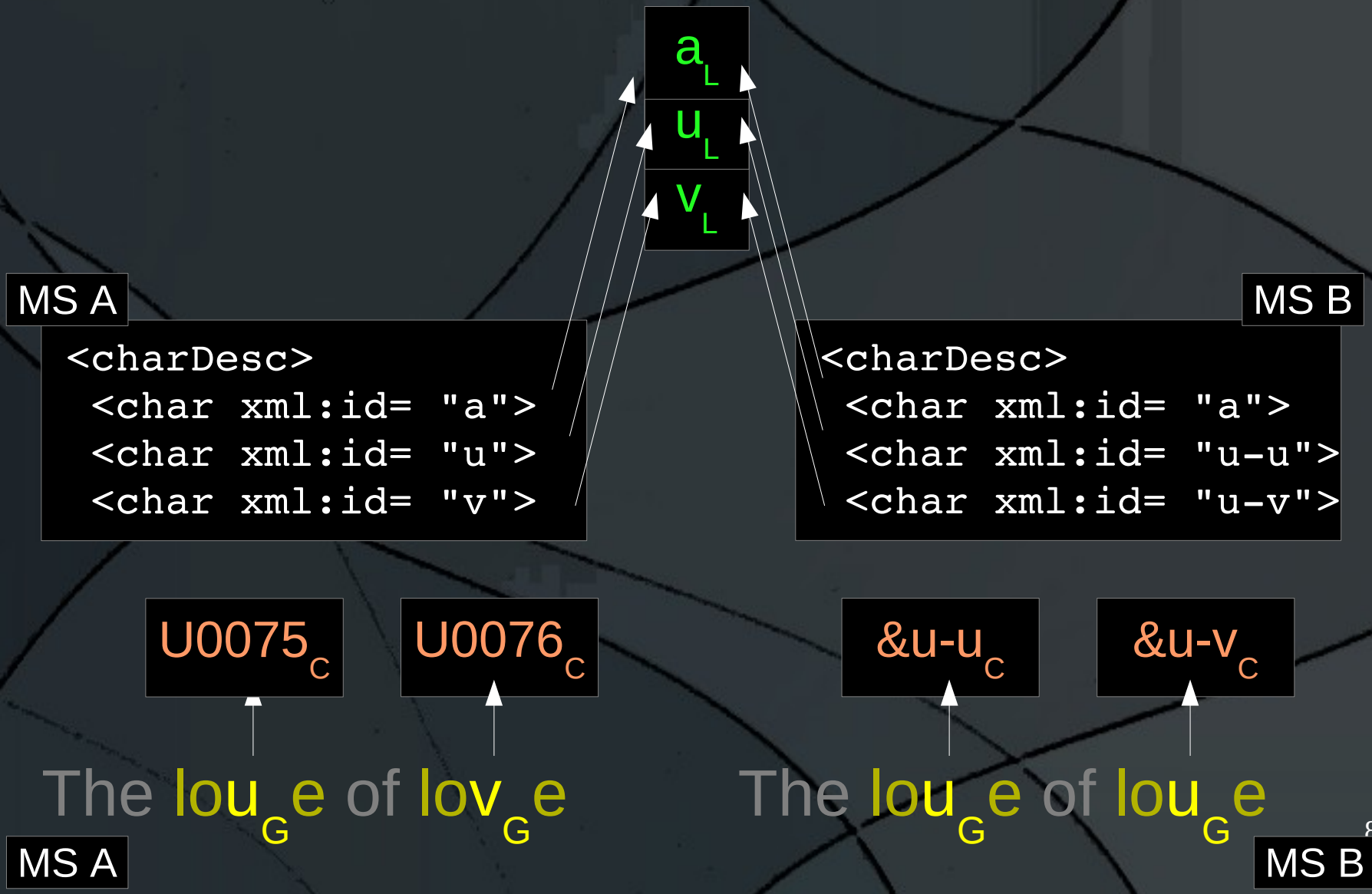
B. Comparison at graphemic layer

a_G	a_L	a_G
u_G	u_L	u-u_G
v_G	v_L	u-v_G

- Fun
- But how do you compare graphemes *now*?



B. Comparison at graphemic layer



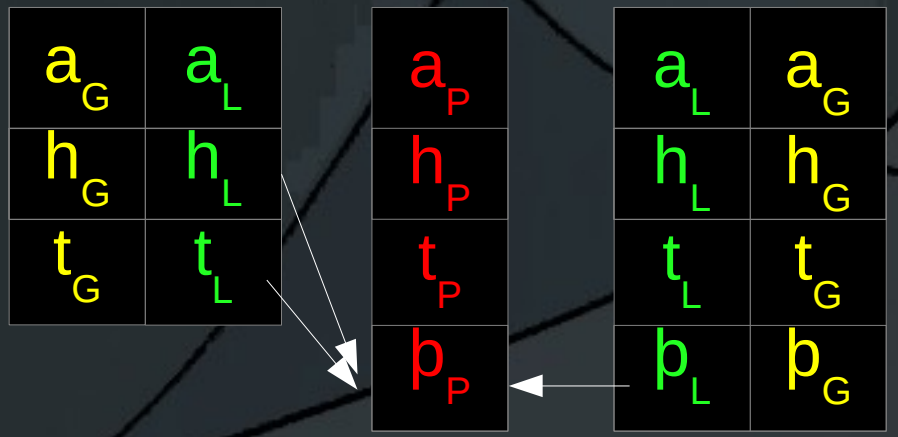
B. Comparison at graphemic layer

a_G	a_L
h_G	h_L
t_G	t_L

a_L	a_G
h_L	h_G
t_L	t_G
$þ_L$	$þ_G$

- When the game gets tough...
  - Different alphabets
    (Middle English, thorn letter)

B. Comparison at graphemic layer



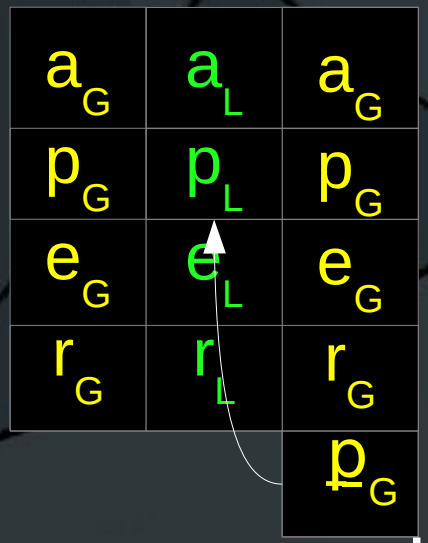
- When the game gets tough...
  - Different alphabets, shared phonetics (Middle English, thorn letter)

B. Comparison at graphemic layer

a_G	a_L	a_G
p_G	p_L	p_G
e_G	e_L	e_G
r_G	r_L	r_G
		$\underline{p}_G$

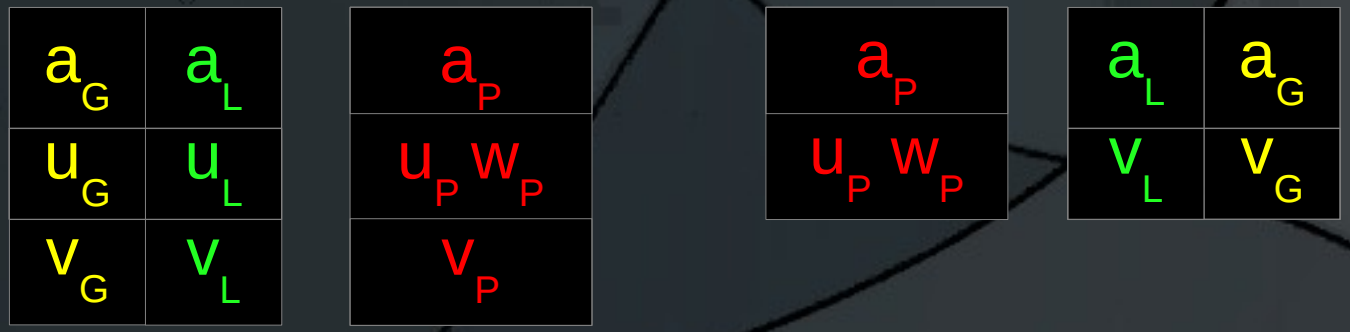
- When the game gets tougher...
  - No 1:1 grapheme / letter ratio (Huitfeldt →)
    - Brevigraphs as graphemes (up to 40%)
    - “Grapheme. (1) A minimally distinctive unit of writing in the context of a particular writing system” (Unicode Glossary →).

B. Comparison at graphemic layer



- When the game gets tougher...
  - No 1:1 grapheme / letter ratio (Huitfeldt →)
    - Brevigraphs as graphemes (up to 40%)
    - “Grapheme. (1) A minimally distinctive unit of writing in the context of a particular writing system” (Unicode Glossary →).

B. Comparison at graphemic layer



- When the game gets most tough...
  - (Classicists start to play)
  - Different alphabets, different phonetics (Ancient Latin)

B. Comparison at graphemic layer

- When the game gets most tough...
  - (Classicists start to play)
  - Different alphabets, different phonetics (Ancient Latin)
  - Still possible to formalise relationships

B. Comparison at graphemic layer

- Description of graphemic system
  - MS-wide
  - Complete
- Formalisation of relationships between graphemes in different graphemic systems
  - Involving higher-level entities (letters, phonemes)
- Intelligent searches, intelligent results

Outline

- The 'Babel issue'
  - The 'Babel issue' in a nutshell
  - Comparing texts
  - In the Tower
- Two issues
  - A. Comparison at the linguistic layer
  - B. Comparison at the graphemic layer