

Paolo Monella

Encoding pre-modern writing systems



Ceci n'est pas un linguiste.

paolo.monella@gmx.net

The scholarly digital edition and the humanities. Theoretical approaches and alternative tools
DiXiT Workshop, Rome, 4 December 2014

Saussure

Saussure

Ferdinand De
Saussure

Relational nature
of signs
within a semiotic
system



Ceci n'est pas un linguiste.

Saussure

MS A

a b c d e f g h **i** l m n o p q r s t **u** z . :

MS B

a b c d e f g h **i j** l m n o p q r s t **u v** z . , :

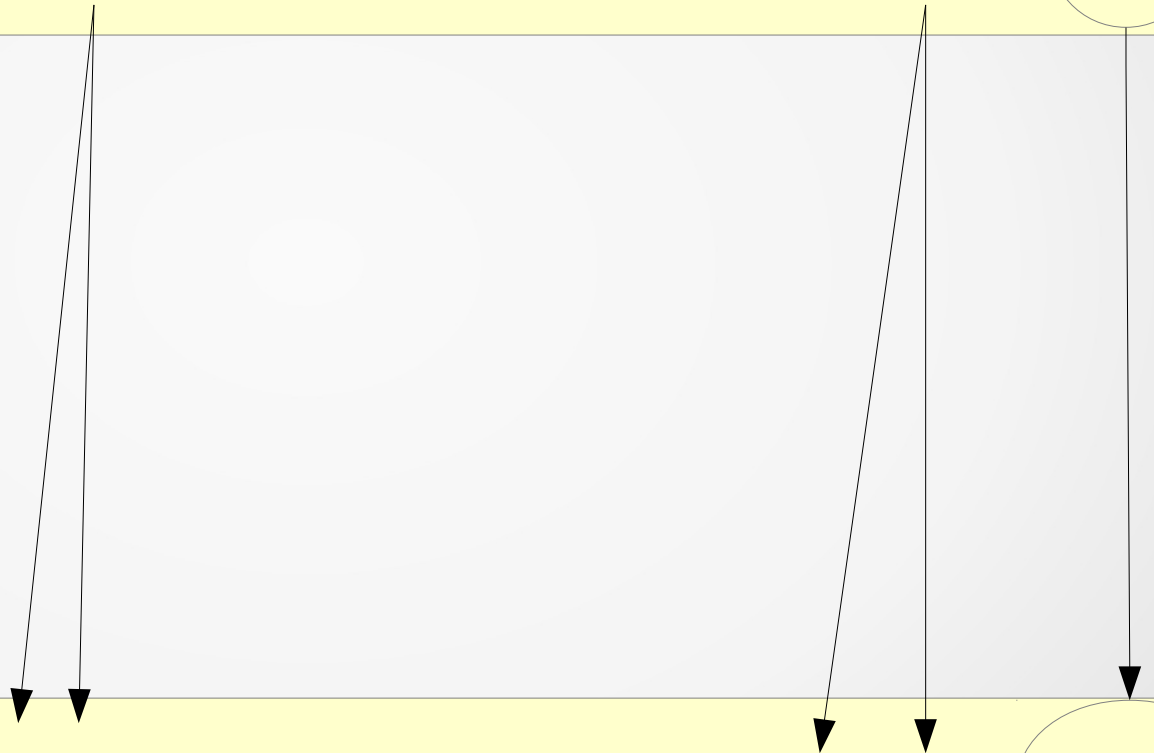
Saussure

MS A

a b c d e f g h **i** l m n o p q r s t **u** z . :

MS B

a b c d e f g h **i** **j** l m n o p q r s t **u** **v** z . , :



Saussure

MS A

a b c d e f g h i l m n o p q r s t u z . :

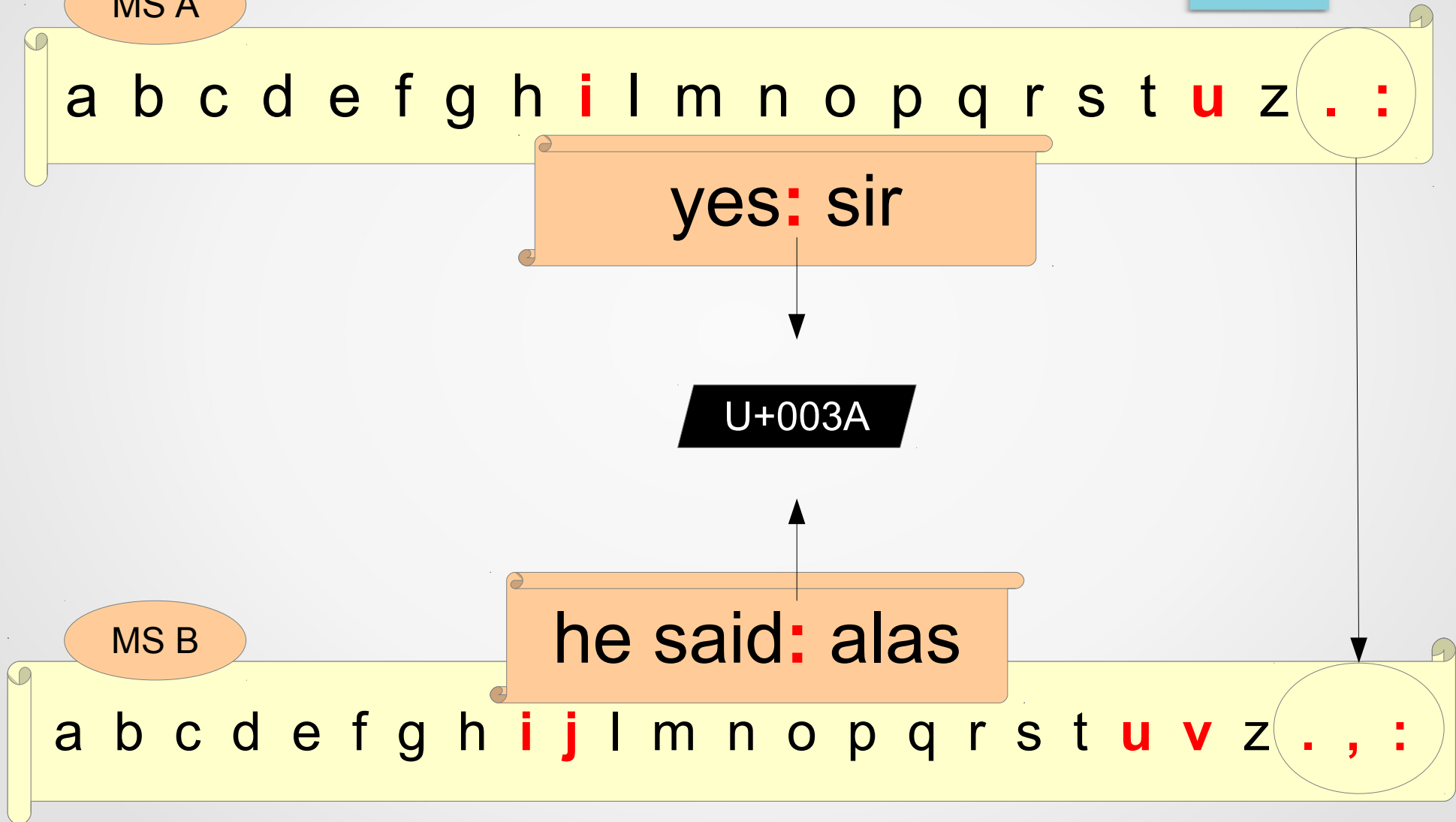
yes: sir

U+003A

MS B

a b c d e f g h i j l m n o p q r s t u v z . , :

he said: alas



Saussure

MS A

a b c d e f g h i l m n o p q r s t u z . :

Every noun

U+0075

≠

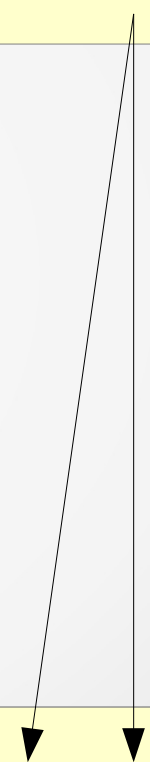
U+0076

U+0075

MS B

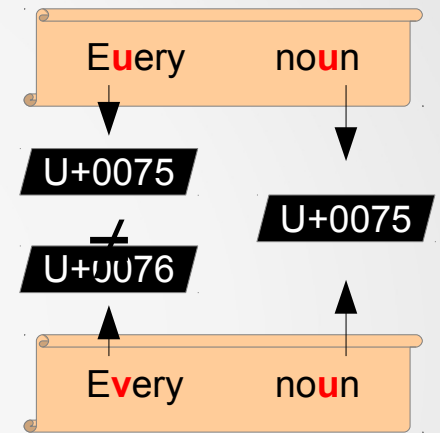
a b c d e f g h i j l m n o p q r s t u v z . , :

Every noun



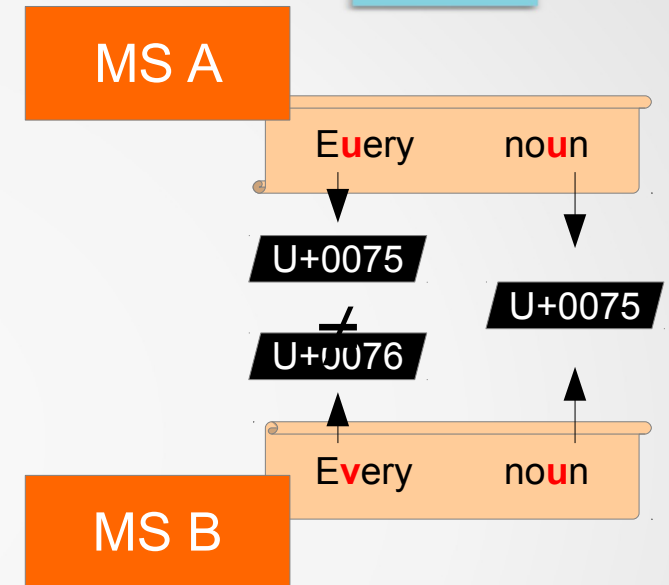
Saussure

- Why comparing texts?



Saussure

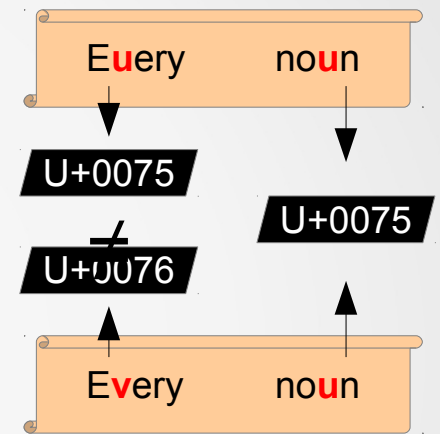
- Why comparing texts?
 - Textual Criticism



Saussure

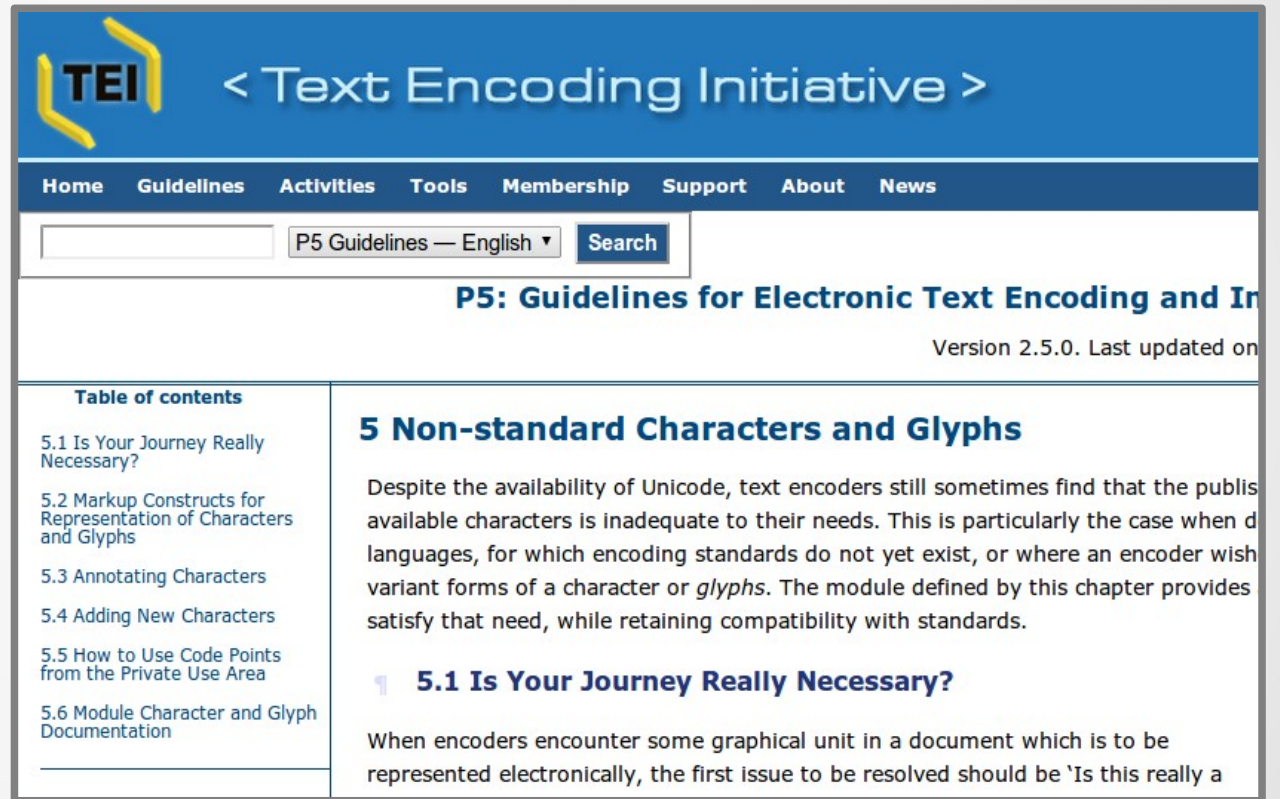
- Why comparing texts?
 - Textual Criticism
 - Processing (e. g. cross-corpus search)

query: "every"



Saussure

- TEI: Unicode (a “u” is a “u”)
 - P5.vi
 - P5.5



The screenshot shows the TEI website interface. At the top is a blue header with the TEI logo and the text "< Text Encoding Initiative >". Below the header is a navigation bar with links: Home, Guidelines, Activities, Tools, Membership, Support, About, and News. A search bar is located below the navigation bar, with a dropdown menu set to "P5 Guidelines — English" and a "Search" button. The main content area is titled "P5: Guidelines for Electronic Text Encoding and Interpretation" with the version "Version 2.5.0. Last updated on" (the date is partially obscured). On the left side, there is a "Table of contents" section listing various chapters, including "5.1 Is Your Journey Really Necessary?". The main content area displays the title "5 Non-standard Characters and Glyphs" and a paragraph explaining that despite the availability of Unicode, text encoders still sometimes find that the published available characters are inadequate to their needs. It mentions that this is particularly the case when dealing with languages for which encoding standards do not yet exist, or where an encoder wishes to represent variant forms of a character or *glyphs*. The module defined by this chapter provides a way to satisfy that need, while retaining compatibility with standards. Below this paragraph, the section "5.1 Is Your Journey Really Necessary?" is introduced, followed by a paragraph stating that when encoders encounter some graphical unit in a document which is to be represented electronically, the first issue to be resolved should be 'Is this really a'.

TEI < Text Encoding Initiative >

Home Guidelines Activities Tools Membership Support About News

P5 Guidelines — English

P5: Guidelines for Electronic Text Encoding and Interpretation
Version 2.5.0. Last updated on

Table of contents

- 5.1 Is Your Journey Really Necessary?
- 5.2 Markup Constructs for Representation of Characters and Glyphs
- 5.3 Annotating Characters
- 5.4 Adding New Characters
- 5.5 How to Use Code Points from the Private Use Area
- 5.6 Module Character and Glyph Documentation

5 Non-standard Characters and Glyphs

Despite the availability of Unicode, text encoders still sometimes find that the published available characters is inadequate to their needs. This is particularly the case when dealing with languages, for which encoding standards do not yet exist, or where an encoder wishes to represent variant forms of a character or *glyphs*. The module defined by this chapter provides a way to satisfy that need, while retaining compatibility with standards.

5.1 Is Your Journey Really Necessary?

When encoders encounter some graphical unit in a document which is to be represented electronically, the first issue to be resolved should be 'Is this really a'

Saussure

- TEI: Unicode
(a “u” is a “u”)
- Corpus-wide
normalization:
Canterbury
Tales
[documentation](#)

2. The practice of this transcription

A specially designed computer screen font was used for the transcription.

a b c d e f g h i j k l m n o p q r s t u v w x y
A B C D E F G H I J K L M N O P Q R S T U V W X Y

These were supplemented by the following Middle English characters:

þ ȝ þ ȝ

The following signs of abbreviation were available:

p p P P² 9 9 v f q ȝ

These characters are available as superscripts:

a e i r t u

Characters usually occurring at word ends are:

ȝ H h d

Marks of punctuation are:⁹

¶ , ; : . ()

Saussure

- T. Orlandi: **SDE**
and the theory
of systems

→ Graphic
system
of MS D

a b c d e f g h i j l m n o p q r s t u v x z

A B C D E F G H I J L M N O P Q R S T U V X Z

[spazio] . , :

XML/TEI P5 Gaiji

XML/TEI P5 Gaiji



2. The practice of this transcription

A specially designed computer screen font was used for the transcription.

a b c d e f g h i j k l m n o p q r s t u v w x y
A B C D E F G H I J K L M N O P Q R S T U V W X Y

These were supplemented by the following Middle English characters:

þ ȝ ȝ

The following signs of abbreviation were available:

p p P P 3- 9 + f q ȝ

These characters are available as superscripts:

a e i r t u

Characters usually occurring at word ends are:

ȝ H h d

Marks of punctuation are:⁹

¶ , ; : . ()

a
b
funny-b
c
d
...

XML/TEI P5 Gaiji

<charDecl>

~~a
b~~

<glyph xml:id="funny-b">



2. The practice of this transcription

A specially designed computer screen font was used for the transcription.

abcdefghijklmnopqrstuvwxyz
ABCDEFGHIJKLMNOPQRSTUVWXYZ

These were supplemented by the following Middle English characters:

١٣٣٣

The following signs of abbreviation were available:

$p, p, p, p, 2-9, 9, 5, 6, 9, 7$

These characters are available as superscripts:

acirtu

Characters usually occurring at word ends are:

7 H h d

Marks of punctuation are:⁹

¶, ; : . ()

a
b
funny-b
c
d
...

XML/TEI P5 Gaiji → <char>

<charDecl>

~~a~~
~~b~~

<glyph xml:id="funny-b">

Guide-
lines

<charDecl>

<char xml:id= "uv">

<charName>

SMALL LATIN U OR V

</charName>

<desc>

Expression:

U-shaped when lowercase,
V-shaped when uppercase.

Content:

either letter Latin u or
letter Latin v

</desc>

XML/TEI P5 Gaiji → <char>

<charDecl>

~~a~~

~~b~~

<glyph xml:id="funny-b">

Guide-
lines

<charDecl>

<char xml:id= "uv">

<charProp>

<localName>

Expression

</localName>

<value>

U+0075

</value>

<localName>

Content

</localName>

<value>

u|v

</value>

XML/TEI P5 Gaiji → <char>

<charDecl>

~~a~~
~~b~~

<glyph xml:id="funny-b">

Guide-
lines

<charDecl>

<char xml:id= "v">

<charName>Premodern Latin
uncial lowercase v</charName>

<charProp>

<localName>**Expression**

</localName>

<value>U+0076</value>

<localName>**Content**

</localName>

<value>v</value>

<charProp>

<mapping type="standard">v

</mapping>

<graphic url="v.jpg"/

</char>

<charDecl>

XML/TEI P5 Gaiji → Comparing

MS A

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

```
<g ref="#uv" />
```

MS B

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

```
<char xml:id= "v">
  <charProp>
    <localName>Expression
  </localName>
    <value>U+0076</value>
    <localName>Content
  </localName>
    <value>v</value>
```

```
<g ref="#v" />
```

XML/TEI P5 Gaiji → Comparing

MS A

<char xml:id= "uv">

<mapping type="mapto_u">u
</mapping>

<mapping type="mapto_v">v
</mapping>

<g ref="#uv" type="mapto_v" />
<g ref="#uv" type="mapto_u" />

MS B

<char xml:id= "u">

<mapping type="standard">u
</mapping>

<char xml:id= "v">

<mapping type="standard">v
</mapping>

XML/TEI P5 Gaiji → Comparing

MS A

<char xml:id= "uv">

<mapping type="mapto_u">u
</mapping>

<mapping type="mapto_v">v
</mapping>

<g ref="#uv" type="mapto_v" />

MS B

<char xml:id= "u">

<mapping type="standard">u
</mapping>

<char xml:id= "v">

<mapping type="standard">v
</mapping>

<g ref="#v" />

XML/TEI P5 Gaiji → All signs?

<charDecl>

~~a~~
~~b~~

<char xml:id="uv">

Guide-
lines

<charDecl>

<char xml:id="a">

<char xml:id="b">

<char xml:id="uv">

a

b

c

...

uv

...

XML/TEI P5 Gaiji → All signs?

<charDecl>

~~a~~
~~b~~

<char xml:id="uv">

Guide-
lines

<charDecl>

<char xml:id="a">

<char xml:id="b">

<char xm:id="uv">

<body>

a

b

<g ref="#uv">

XML/TEI P5 Gaiji → All signs?

<charDecl>

~~a~~

~~b~~

<char xml:id="uv">

Guide-
lines

<charDecl>

<char xml:id="a">

<char xml:id="b">

<char xm:id="uv">

<body>

a

b

<g ref="#uv">

XML/TEI P5 Gaiji → All signs?

<charDecl>

~~a~~
~~b~~

<char xml:id="uv">

Guide-
lines

<charDecl>

<char xml:id="a">

<char xml:id="b">

<char xm:id="uv">

<body>

a

b

<g ref="#uv">

<body>

<g ref="#a">

<g ref="#b">

<g ref="#uv">

XML/TEI P5 Gaiji → All signs?

<charDecl>

~~a~~
~~b~~

<char xml:id="uv">

Guide-
lines

<charDecl>

<char xml:id="a">

<char xml:id="b">

<char xm:id="uv">

<body>

a

b

<g ref="#uv">

<body>

<g ref="#a">

<g ref="#b">

<g ref="#uv">

XML/TEI P5 Gaiji → All signs?

<charDecl>

~~a~~
~~b~~

<char xml:id="uv">

Guide-
lines

<charDecl>

<char xml:id="a">

<char xml:id="b">

<char xm:id="uv">

<body>

a

b

<g ref="#uv">

<body>

<g ref="#a">

<g ref="#b">

<g ref="#uv">

Vespa Project

www.unipa.it/paolo.monella/lincei/edition.html

XML/TEI P5 Gaiji → All signs?

<charDecl>

~~a~~
~~b~~

<char xml:id="uv">

Guide-
lines

<charDecl>

<char xml:id="a">

<char xml:id="b">

<char xm:id="uv">

<body>

a

b

<g ref="#uv">

Techn.

<body>

<g ref="#a">

<g ref="#b">

<g ref="#uv">

<body>

a

b

<g ref="#uv">

<g ref="#uv" />ir

<g ref="#uv" />
<g ref="#i" />
<g ref="#r" />

<g ref="#uv" />ir

XML/TEI P5 Gaiji → What would it take?

XML/TEI P5 Gaiji → What would it take?

<charDecl>

~~a~~
~~b~~

<glyph id="fun-b">

Guide-
lines

<charDecl>

<char xml:id="a">

<char xml:id="b">

<glyph id="fun-b">

<body>

a

b

<g ref="#fun-b">

Techn.

<body>

<g ref="#a">

<g ref="#b">

<g ref="#fun-b">

<body>

a

b

<g ref="#fun-b">

XML/TEI P5 Gaiji → What would it take?

```
<charDecl>
```

~~a~~
~~b~~

```
<glyph id="fun-b">
```

Guide-
lines

```
<charDecl>
```

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

```
<char xml:id="b">
```

```
<glyph id="fun-b">
```

```
<body>
```

a

b

```
<g ref="#fun-b">
```

Techn.

```
<body>
```

```
<g ref="#a">
```

```
<g ref="#b">
```

```
<g ref="#fun-b">
```

```
<body>
```

a

b

```
<g ref="#fun-b">
```

- More lines of code

XML/TEI P5 Gaiji → What would it take?

<charDecl>

~~a~~
~~b~~

<glyph id="fun-b">

Guide-
lines

<charDecl>

<char xml:id="a">

↑<char xml:id="b">

<glyph id="fun-b">

<body>

a

b

<g ref="#fun-b">

Techn.

<body>

<g ref="#a">

<g ref="#b">

<g ref="#fun-b">

<body>

a

b

<g ref="#fun-b">

- More lines of code
- Guidelines: <char>

XML/TEI P5 Gaiji → What would it take?

<charDecl>

~~a~~
~~b~~

<glyph id="fun-b">

Guide-
lines

<charDecl>

<char xml:id="a">

<char xml:id="b">

<glyph id="fun-b">

<body>

a

b

<g ref="#fun-b">

Techn.

<body>

<g ref="#a">

<g ref="#b">

<g ref="#fun-b">

<body>

a

b

<g ref="#fun-b">

- More lines of code
- Guidelines: <char>

- Technical: no <g>

XML/TEI P5 Gaiji → What would it take?

<charDecl>

~~a~~
~~b~~

<glyph id="fun-b">

Guide-
lines

<charDecl>

<char xml:id="a">

<char xml:id="b">

<glyph id="fun-b">

<body>

a

b

<g ref="#fun-b">

Techn.

<body>

<g ref="#a">

<g ref="#b">

<g ref="#fun-b">

<body>

a

b

<g ref="#fun-b">

- More lines of code
- Guidelines: <char>

- Technical: no <g>
- Interoperability: <mapping>

XML/TEI P5 Gaiji → What would it take?

<charDecl>

~~a~~
~~b~~

<glyph id="fun-b">

Guide-
lines

<charDecl>

<char xml:id="a">

<char xml:id="b">

<glyph id="fun-b">

<body>

a

b

<g ref="#fun-b">

Techn.

<body>

<g ref="#a">

<g ref="#b">

<g ref="#fun-b">

<body>

a

b

<g ref="#fun-b">

- More lines of code
- Guidelines: <char>

- Technical: no <g>
- Interoperability: <mapping>

Vespa Project

Vespa Project → Table of signs

Graphemes ID	Content (alphabemes ID)	Expression
t	t	Latin minuscule uncial t
u	uv	Latin minuscule uncial u/v (u-shaped, not v-shaped)
ae	a, e	Latin minuscule uncial e with tail bottom left: img/ax.jpg
b_	b, i, s	Latin minuscule uncial b with macron top right
.		Middle dot

Vespa Project → Source file

Graphemes ID	Words ID
Judici(u_)	nou,[iudicium],n,s,iudicium
coci	nou,[cocus],g,s,coci
et	con,[et],et
pistoris	nou,[pistor],g,s,pistoris
-lb	

Vespa Project → Generated files

graphematic.xml

- `<g id="1.1" ref="#J" />`
- `<g id="1.2" ref="#u" />`
- `<g id="1.3" ref="#d" />`
- `<g id="1.4" ref="#i" />`
- `<g id="1.5" ref="#c" />`
- `<g id="1.6" ref="#i" />`
- `<g id="1.7" ref="#u_" />`

alphabetic.xml

- `<c id="1.1.1" ref="#j" />`
- `<c id="1.2.1" ref="#uv" />`
- `<c id="1.3.1" ref="#d" />`
- `<c id="1.4.1" ref="#i" />`
- `<c id="1.5.1" ref="#c" />`
- `<c id="1.6.1" ref="#i" />`
- `<c id="1.7.1" ref="#uv" />`
- `<c id="1.7.2" ref="#m" />`

linguistic.xml

- `<w id="1">`
 `nou,[iudicium],n,s,iudicium`
 `</w>`

Vespa Project → Alignment

graphematic.xml

- `<g id="1.1" ref="#J" />`
- `<g id="1.2" ref="#u" />`
- `<g id="1.3" ref="#d" />`
- `<g id="1.4" ref="#i" />`
- `<g id="1.5" ref="#c" />`
- `<g id="1.6" ref="#i" />`
- `<g id="1.7" ref="#u_" />`

alphabetic.xml

- `<c id="1.1.1" ref="#j" />`
- `<c id="1.2.1" ref="#uv" />`
- `<c id="1.3.1" ref="#d" />`
- `<c id="1.4.1" ref="#i" />`
- `<c id="1.5.1" ref="#c" />`
- `<c id="1.6.1" ref="#i" />`
- `<c id="1.7.1" ref="#uv" />`
- `<c id="1.7.2" ref="#m" />`

align_alph_graph.xml

- `<link targets="graphematic.xml#1.7 #1.7" />`
- `<ptr id="1.7" targets=
 "alphabetic.xml#1.7.1
 alphabetic.xml#1.7.2" />`

Looking for solutions

Looking for solutions

- Medieval Unicode Font Initiative →
- Unicode Private Use Area (PUA) →
- Gaiji <glyph> →
- SGML/TEI P3 Writing System Declaration (WSD) →

Looking for solutions

- SGML/TEI P3 Writing System Declaration (WSD) →

```
<writingSystemDeclaration lang='eng' name=' ... ' date='1993-05-29'>  
  <language iso639='...'><!-- name of language here --></language>  
  <script><!-- description of script here ... --></script>  
  <direction chars=LR lines=TB>  
  <characters>  
    <!-- description of character inventory here ... -->  
  </characters>  
</writingSystemDeclaration>
```

Looking for solutions

- SGML/TEI P3 Writing System Declaration (WSD) →

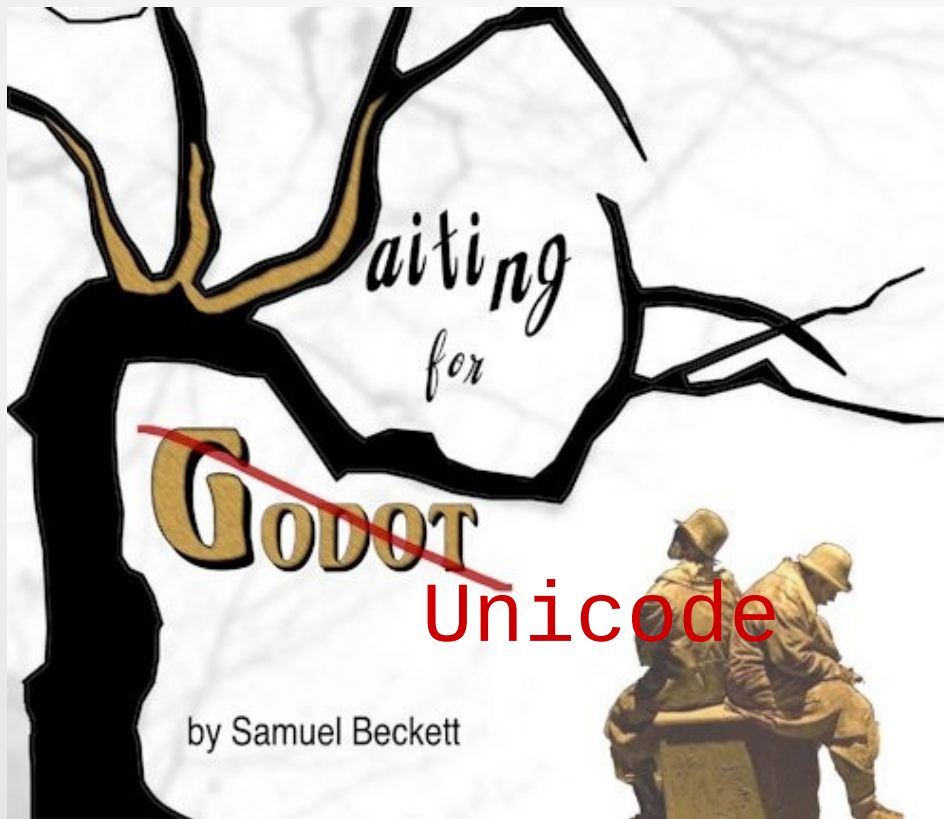
25.4.1 Base Components of the WSD

[...] in the <characters> element:

- reference to an international standard
- reference to a public set of SGML entities
- reference to another WSD
- **formal declaration of each graphic unit in the writing system**
- a combination of the above

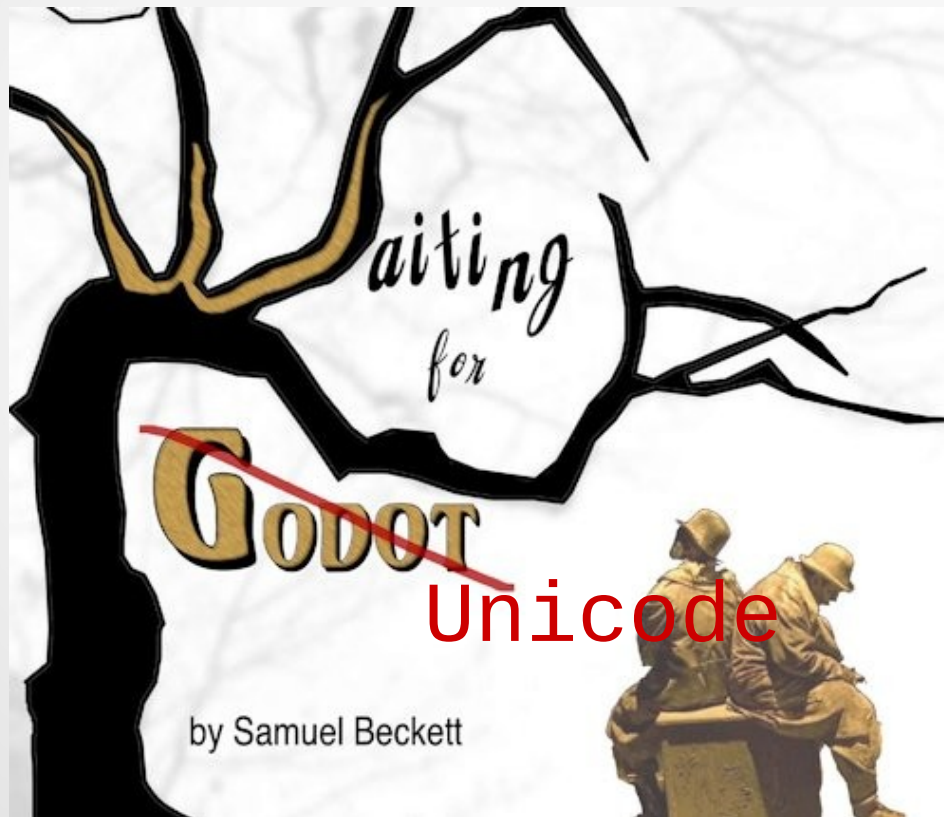
Looking for solutions

- SGML/TEI P3 Writing System Declaration (WSD) →



Looking for solutions

- SGML/TEI P3 Writing System Declaration (WSD) →



... then Unicode arrived!

Birnbaum,
Clemenson,
Kempgen &
Ribarov 2008 →

Looking for solutions

<charDecl>

~~a~~
~~b~~

<char xml:id="uv">

Guide-
lines

<charDecl>

<char xml:id="a">

<char xml:id="b">

<char xm:id="uv">

<body>

a

b

<g ref="#uv">

Techn.

<body>

<g ref="#a">

<g ref="#b">

<g ref="#uv">

<body>

a

b

<g ref="#uv">