

Encoding Allographs: (ab?) Using the <g> Element

Alexei Lavrentiev¹ & Dominique Stutzmann²

¹ ICAR Research Lab & ASLAN Labex, CNRS, Université de Lyon

² Institut de Recherche et d'Histoire des Textes, CNRS

“Connect, Animate, Innovate”. TEI Conference and Members’ Meeting
30 October 2015, Lyon, France

Outline

- Project Background
- Transcription Types
 - allographic / diplomatic / normalized / hybrid
- Encoding annotations at character level
 - text-to-image alignment
 - allogrpahs, normalization, capitalization, diacritics
- TEI solutions
 - <c> / <g> + <glyph>
- Conclusion

Oriflamms project

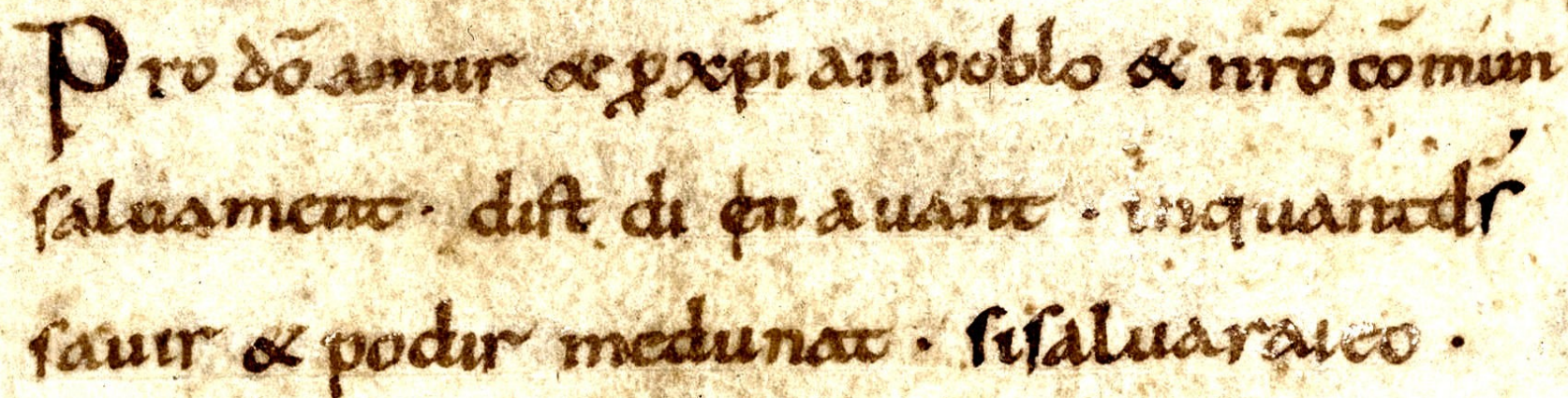
- **Ontology Research, Image Features, Letterform Analysis on Multilingual Medieval Scripts**
 - <http://oriflamms.hypotheses.org>
 - Funded by the French ANR Agency (2013-2015)
 - resp.: Dominique Stutzmann
- **Researchers and engineers in**
 - the humanities
 - palaeography, epigraphy, linguistics
 - information technologies
 - image analysis, corpus design and query

Objectives and deliverables

- Establish an ontology of letter (character) forms in the Medieval scripts (Latin and vernacular)
- Analyze the graphical structures in the M.S.
- Provide training data for HTR (handwritten text recognition) software

- ==> a corpus of transcriptions
 - aligned to facsimile image zones
 - at word and character level
- standard, interchangeable format
 - XML-TEI

Oaths of Strasbourg



Pro dō amur & p xpi an poblo & nrō cōmun
saluament. dist di en auant. inquantd' r'
saur & podir medunat. sisaluarai eo.

Ms. Paris, BnF, lat. 9768 (circa 1000)

1 Pro dō amur & p xpi an poblo & nrō cōmun | salua- 6
ment. dist di en auant: inquantd' r' | saur & podir medunat.
8 sisaluarai eo. | cist meon fradre karlo. & in ad iudha. |

Ed. Eduard Koschwitz (*Les plus anciens monuments de la langue française...*, Leipzig:
O.R. Reisland, 1920)

Oaths of Strasbourg

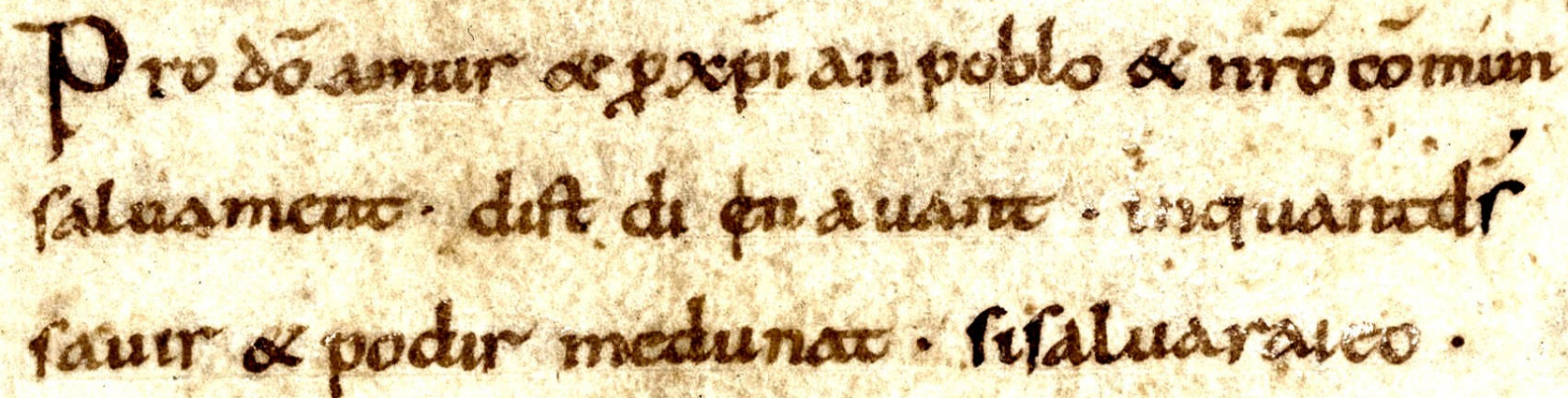
Pro dō amur & p xpi an poblo & nro comun
saluament. dist di en avant. inquant d'
sauir & podir medunat. si salvarai eo.

Ms. Paris, BnF, lat. 9768 (*circa* 1000)

« Pro Deo amur et pro christian poblo et nostro com-
« mun salvament, d'ist di in^b avant, in quant Deus savir
« et podir² me dunat, si salvarai^e eo cist meon fradre

Ed. Philippe Lauer (*Nithard, Histoire des fils de Louis le Pieux*, Paris: Champion, 1926)

Oaths of Strasbourg



Pro dō amur et pro xpian poblo & nro comun
saluament. d'ist di en auant. in quant
saur & podir me dumat. si saluarai eo.

Ms. Paris, BnF, lat. 9768 (*circa* 1000)

Pro Deo amur et pro christian poblo et nostro
commun saluament, d'ist di in auant, in quant
Deus sauir et podir me dumat, si saluarai eo

Ed. Robert-Leon Wagner & Olivier Collet (*Textes d'étude (Ancien et moyen français)*,
Genève: Droz, 1995)

Allographic transcription

AKA “palaeographic”, “imitative”, “diplomatic” (in epigraphy)

Pro $\bar{d}\bar{o}$ amur & $\text{p}\bar{x}\bar{p}\bar{i}$ an poblo & $\bar{n}\bar{r}\bar{o}$ $\bar{c}\bar{o}\bar{m}\bar{u}\bar{n}$
 faluament · dist d₁ [e + I]n auant · inquant $\bar{d}\bar{f}$
 fauir & podir medunat · fifaluar₁eo ·

- major letter variants (f / s, z / r) preserved
 - abbreviation markers preserved
 - original punctuation and word segmentation
- Relatively easy to align to image
 - Rare in practice

Diplomatic transcription

AKA “documentary”

Pro deo amur *et pro christian* poblo *et nostro commun*
saluament. d’ist di In auant. in quant *deus*
saur *et* podir me dunat. si saluarai eo

- *u/v* and *i/j* letters not normalized, no capitals added
 - abbreviation expansions are marked (italics)
 - original punctuation partly preserved
- Can be aligned to image if properly tagged

Normalized transcription

AKA “critical”

« Pro Deo amur et pro christian poblo et nostro commun salvament, d'ist di in avant, in quant Deus savir et podir me dunat, si salvarai eo ... »

- *u/v* and *i/j* letters normalized
 - abbreviation expansions are unmarked
 - punctuation normalized, proper nouns capitalized
- Hardly alignable to images
 - The most widespread in literary text editions

Hybrid Digital Transcriptions

- Using markup to combine features of various transcription types
 - `<tei:choice>`
 - `<abbr>` + `<expan>`
 - `<orig>` + `<reg>`
 - `<sic>` + `<corr>`


Hybrid Digital Transcriptions

bzm.bfm-corpus.org

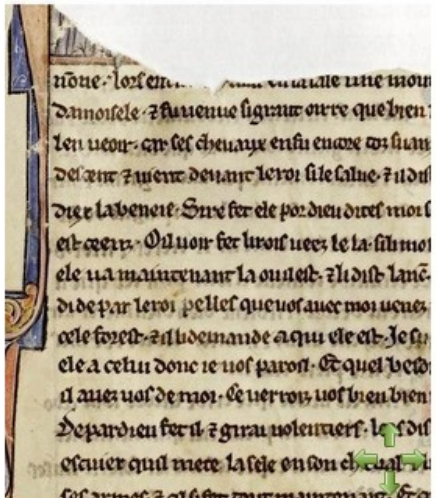
Rechercher

Bienvenue S'inscrire Se connecter Aide Contact fr

Accueil GRAAL Text: qgraal_cm



tour au manuscrit K (Lyon, BM, P.A. 77, col.



§ 1

A [la ueille delapente
coste qnt li compai
gnon de latable re
onde furent uenu
5 a kamaalot 7 il o
rent oi leseruisse 7
len uoloit metre lef
tablef a heure de]
nōne . lors en[tra] [acheual en la]^[1] fale une
damoisele . ɛ fuueneue figrant oirre que bien le
len ueoir . car fes cheuaux enfu encoze toz fua
descent et uient deuant le roi si le salue . et il di
5 diex la beneie . Sire fet ele por dieu dites
est ceenz . Oil uoir fet li rois ueez le la . filmos
ele ua maintenant la oule est . ɛ li dist lanc . ueo
di de par leroi pelles queuof avec moi uenez i
ele a celui donc ie uof paroil . Et quel besoign
10 il auez uof de moi . Ce uerroz uof bien (bien
Depardieu fet il . ɛ gra uolentierf . lozf dist a
escuier quil mete la sele en son cheual . ɛ lapoz
fef armesf . ɛ cil lifet tout maintenant . Et quan
15 liroif et lautre qui ou pales estoient uoient ce
enpoise mout . Et nepozquant quant il uoient
quil ne remaindrait il len lessent aler . Et lareu
ladist . Que est ce lanc . Nos lairez uof a cest so
fi est hauz . Dame fet la damoisele fachiez que

§ 1

A [la ueille de la pente-
coste quant li compai-
gnon de la table re-
onde furent uenu
5 a kamaalot et il o-
rent oi le seruisse et
l'en uoloit metre les
tables a heure de]
nonne . lors en[tra] [a cheual en la]^[1] sale
damoisele , et fu uenue si grant oirre que
l'en ueoir , car ses cheuaux en fu encore t
descent et uient deuant le roi si le salue ,
diex la beneie . Sire fet ele por dieu dites
est ceenz . Oil uoir fet li rois ueez le la , si
ele ua maintenant la ou il est , et li dist le
di de par le roi pelles que uos avec moi u
cele forest , et il li demande a qui ele est .
10 ele a celui donc ie uos paroil . Et quel bes
il auez uos de moi . Ce uerrois uos bien
De par dieu fet il , et g'irai uolentiers , lors
escuier qu'il mete la sele en son cheual ,
ses armes , et cil si fet tout maintenant . E
15 li rois et li autre qui ou pales estoient uoi
en poise mout . Et neporquant quant il uoi
qu'il ne remaindrait il l'en lessent aler . Et
li dist . Que est ce lance/lot . Nos lairez u
si est hauz . Dame fet la damoisele sach

(p. 1)

§ 1

[A la veille de la Pente-
coste quant li compai-
gnon de la table re-
onde furent uenu
5 a Kamaalot et il o-
rent oi le seruisse et
l'en uoloit metre les
tables a heure de]
nonne . lors en[tra] [a cheual en la]^[1] sale
damoisele , et fu uenue si grant oirre que
l'en ueoir , car ses cheuaux en fu encore t
descent et uient deuant le roi si le salue ,
diex la beneie . « Sire , fet ele , por Dieu di
est ceenz . - Oil uoir , fet li rois , ueez le la .
ele va maintenant la ou il est , et li dist : «
di de par le roi Pellés que uos avec moi ve
cele forest . » Et il li demande a qui ele es
10 ele , a celui donc je uos paroil . - Et quel bi
il , auez uos de moi ? - Ce verrois uos bien
- De par Dieu , fet il , et g'irai uolentiers . »
escuier qu'il mete la sele en son cheual ,
ses armes , et cil si fet tout ma
15 li rois et li autre qui ou palés e
en poise mout . Et neporquant
qu'il ne remaindrait il l'en lessé
li dist : « Que est ce Lancelot
si est hauz ? - Dame , fet la da

GRAAL - ms-colonne, ...

page

ms-colonne | fac-similaire | diplomatique | courante

160a / 268

PDF Notice

Oriflamm's software: word alignment validation and correction

The screenshot displays the Oriflamm software interface for word alignment validation and correction. The main window shows a medieval manuscript page with text in Gothic script. The text is overlaid with a grid of colored boxes (yellow and green) representing word boundaries and alignment. The text is organized into two columns, with the left column being the primary focus.

The text in the left column is as follows:

france que que uos ientreroy prochainement
 Et por ce que ie lai bien que ie ne uos ierral fames
 coz enu ensemble come uos estes iorendroit uoil
 ge que en la praere de l'amaalor iorendroit
 comencez i coruolement lienuoie que apres
 nos moz en facez remembrance li ior qui aps
 nos uendront a li l' accordent tuit a celle parole
 Diremment en la cite i prennent ior armes de
 nex iot por ioster iobu ieur la de nex iot qui ne
 pristrest fors iouertures iueuz i car mouz sa iod

The right column of text is partially visible and reads:

de G
 ma
 por
 la re
 lauo
 le re
 com
 ueill
 trem
 dame

The software interface includes a file list on the left, a toolbar at the top, and a ruler at the bottom. The file list shows various image files (e.g., lyonbm_pa77-160.jpg) and their corresponding column and line counts.

Oriflamm's software: character alignment validation and correction

The screenshot displays the Oriflamm software interface for character alignment validation. The main window shows a manuscript page with a timeline at the top ranging from 1200 to 1300. The text is segmented into individual characters, with some characters highlighted in yellow boxes. Green horizontal lines indicate alignment points across the text. The interface includes a menu bar (Fichier, Alignement, Options), a toolbar with zoom and pan icons, and a sidebar on the left with a file list and a column selection menu (Colonne 1, Colonne 2). The Windows taskbar at the bottom shows various application icons and the system clock (01:21, 21/07/2014).

Oriflamms software: validating alignment by word lists

The screenshot shows the Oriflamms software interface. On the left, a table lists words and their statistics. The main area displays a document image with a grid overlay for validation. A tooltip is visible over the document image.

Mot	Compte	Longueur
Tandis	1	6
tance	3	5
tan	6	3
talet	3	5
talent	19	6
tairai	1	6
taine	1	5
taillie	1	7
tailles	1	7
taigne	2	6
tage	2	4
tacion	1	6
tachiez	1	7
tachie	2	6
tache	9	5
tace	1	4
tables	9	6
tablef	1	6
table	72	5
ta	18	2
t'	80	2
t	2	1
synagogue	1	9

Document image content (from top to bottom):

- table
- table able
- blexon
- table
- le table
- a table

Tooltip text:

```
Image 9
Colonne 1
Ligne 21
Mot 4
```

System tray: 14:17, 15/09/2013, Fermer

Oriflamms software: validating alignment by character lists

The screenshot displays the Oriflamms software interface. The main window, titled "Validation", shows a grid of handwritten characters in a medieval script. The grid is organized into rows and columns, with each cell containing a single character. The characters are arranged in a regular pattern, likely for validation purposes. On the left side, there is a list of characters and their corresponding counts. The list is as follows:

Mot	Compte	Longueur
ELIS_gi	1	7
ELIS_gn	28	7
ELIS_gr	10	7
ELIS_gu	1	7
ELIS_ri	40	7
ELIS_rm	2	7
ELIS_rp	1	7
ELIS_ti	2	7
ELIS_tr	103	7
ELIS_tu	2	7
G	434	1
H	4	1
I	48	1
J	73	1
K	8	1
L	51	1
LIG_ez	4	6
LIG_ff	13	6
LIG_ss	875	6
LIG_st	5524	6
M	178	1
N	160	1
O	66	1
P	54	1
Q	262	1
R	27	1

The "LIG_st" entry is highlighted in blue. On the right side of the validation window, there is a section labeled "Rejeté (0)", which is currently empty. The software is running on a Windows operating system, as indicated by the taskbar at the bottom, which shows various application icons and the system tray with the date and time (21/07/2014, 01:23).

Tokenization

■ Word level

- `<w>` or `<pc>` on every alignable word or punctuation mark
- `@xml:id` required (anchor for alignment)

■ Character level

- `<c>` on every alignable character
- `@xml:id` required



TEI solutions for allographs

- Functional approach
- <choice> approach
- <g> approach

Functional approach

- “WYSIWYG” at cdata level
- use “semantic markup” to normalize when processing:
 - first letter of <persName> → capitalize
 - first letter in <s> → capitalize
 - <c function="consonant">u</c> → v ???

Functional approach


- “WYSIWYG” at cdata level
- use “semantic markup” to normalize when processing:
 - first letter of <persName> → capitalize
 - first letter in <s> → capitalize
 - <c function="consonant">u</c> → v ???

 - What to do with “non-functional” allographs?
 - initial R, r / ʀ
 - What is <s> in Old French?
 - What if we don't want complex semantic markup?

<choice> approach (1)

- <c>
 - <choice>
 - <orig>u</orig>
 - <reg>V</reg>
 - </choice>
- </c>

<choice> approach (1)

- <c>
 - <choice>
 - <orig>u</orig>
 - <reg>V</reg>
 - </choice>
- </c>
-  Not allowed!
- Very verbose
 - 46 code characters for 1 letter in the source!
- Limited representation layers (S / s / ſ)

<choice> approach (2)

- <choice>
 - <orig><c>u</c></orig>
 - <reg><c>V</c></reg>
- </choice>

<choice> approach (2)

- <choice>
 - <orig><c>u</c></orig>
 - <reg><c>V</c></reg>
- </choice>

- Even more verbose
 - 52 code characters for 1 letter in the source!
- Duplication of <c>
 - problem with identifiers and linking
 - intellectually problematic
 - the choice is within a character, not between 2 characters!

<g> approach

- **<g>** (character or glyph) represents a glyph, or a non-standard character (<http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-g.html>)
- The gaiji module... adds a further way of representing specific characters and glyphs in a document... This allows the encoder
 - to distinguish characters and glyphs which Unicode regards as identical,
 - to add new nonstandard characters or glyphs,
 - and to represent Unicode characters not available in the document encoding by an alternative means.
(<http://www.tei-c.org/release/doc/tei-p5-doc/en/html/WD.html#D25-20>)

<g> approach

- S, s and f or V, v and u
 - different Unicode characters
 - but variants of the same letter at a certain point of writing history
- Why not to use <g> to say:
 - this is a letter variant (=allograph) used at a certain transcription/normalization level
 - at a different transcription/normalization level it corresponds to a different allograph
 - it does not matter whether both allographs have there own Unicode codepoint

Character annotations

- `g/@ref` → `charDecl` / `char` | `glyph`
 - powerful mechanism
 - name, description, gloss, note
 - property/value characterization
 - mapping (Unicode, MUFI...)
 - figure
 - overkill for simple cases?
 - capitalization
 - u/v, i/j distinctions

Character annotations

- `g/@ana` → interp
 - general TEI annotation mechanism
 - project specific (but generalizable?) interpretation model
 - 2 transcription/interpretation levels
 - `dipl(omatic)`
 - `norm(alized)`
 - the basic transcription may be either
 - annotation provides the other level
 - `<g ana="ori:dipl-small">D</g>ieu`
 - `<g ana="ori:norm-u ori:norm-caps">v</g>ne`

Conclusion

- Questions to the TEI
 - Should the content model of <c> be revised?
 - allow <choice> and family
 - allow <am>
 - Is <g> the right element for encoding allographs?
 - even those in the Unicode standard
 - What annotation mechanism to use for “simple” normalization cases?



Thank you!