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

Alexei Lavrentiev[1] & Dominique Stutzmann[2]

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

Oriflamms project

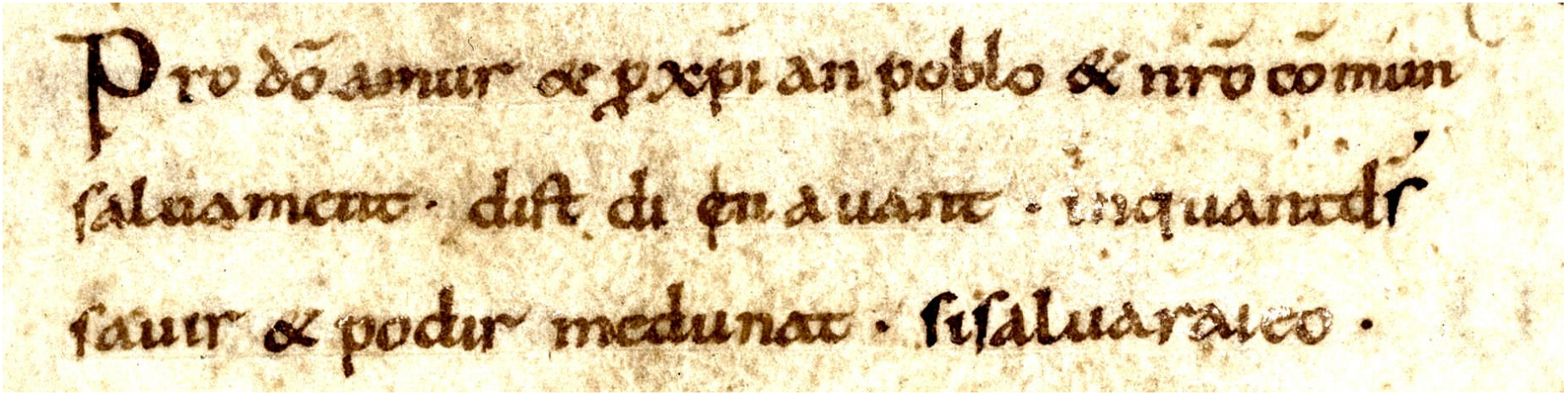
- Ontology Research, Image Features, Letterform Analysis on Multilingual Medieval Scripts
  - http://oriflamms.hypotheses.org
  - Funded by the French ANR Agency (2013-2015)
  - resp.: Dominique Stutzmann
- Researchers and engineers in
  - the humanities
    - palaeography, epigraphy, linguistics
  -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

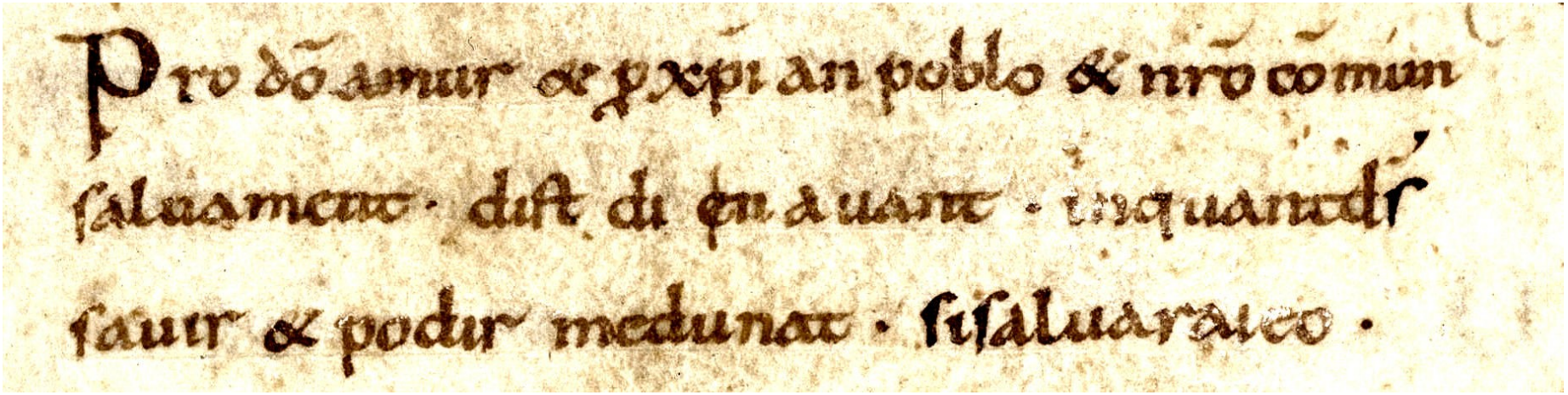


Ms. Paris, BnF, lat. 9768 (*circa* 1000)

1 Pro dõ amur & ꝑ xp̃ian poblo & nrõ cõmun | ſalua- 6
ment. diſt di ẽn auant: inquantdſ | ſauir & podir medunat.
8 ſiſaluaraieo. | ciſt meon fradre karlo. & in aḍ iudha. |

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

Oaths of Strasbourg

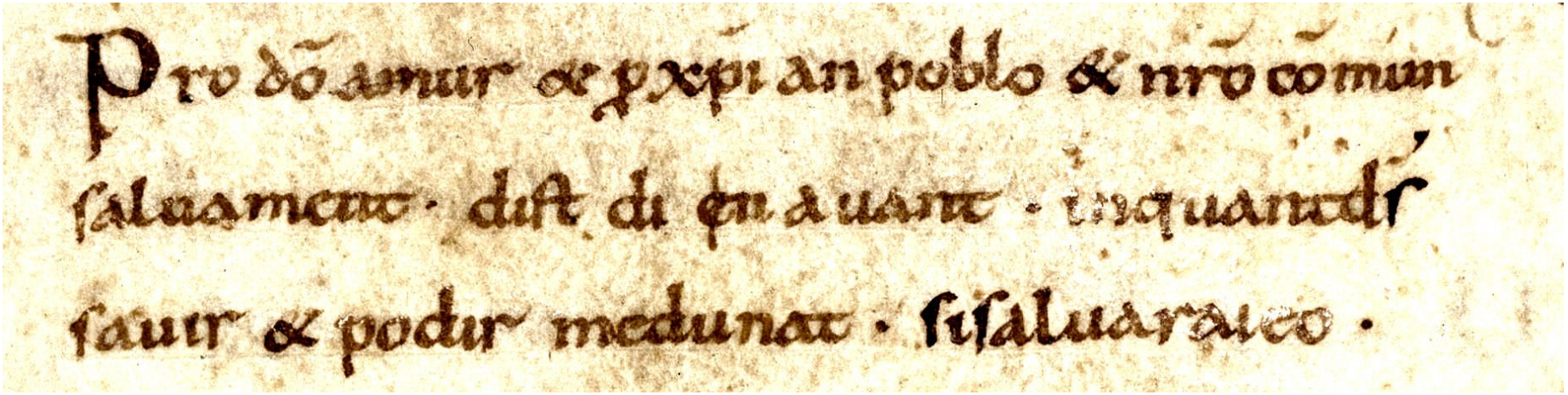


Ms. Paris, BnF, lat. 9768 (*circa* 1000)

« Pro Deo amur et pro christian poblo et nostro com-
« mun salvament, d'ist di in[b] avant, in quant Deus savir
« et podir[2] me dunat, si salvarai[c] eo cist meon fradre

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

Oaths of Strasbourg



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 dunat, si saluarai eo

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

Allographetic transcription

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

Pro dō amur & ꝑxpı an poblo & nr̄o cōmun
ſaluament · dıſt dı [e + I]n auant · ınquantdſ̄
ſauır & podır medunat · ſıſaluaraıeo ·

- major letter variants (ſ / s, ꝛ / r) preserved
- abbreviation markers preserved
- original punctuation and word segmentation

- Relatively easy to align to image
- Rare in practice

Diplomatic transcription

AKA “documentary”

Pro d*e*o amur *et* p*ro chris*tian poblo *et* no*s*tro co*m*mun
saluament. d’ist di In auant. in quant d*eu*s
sauir *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 ... »

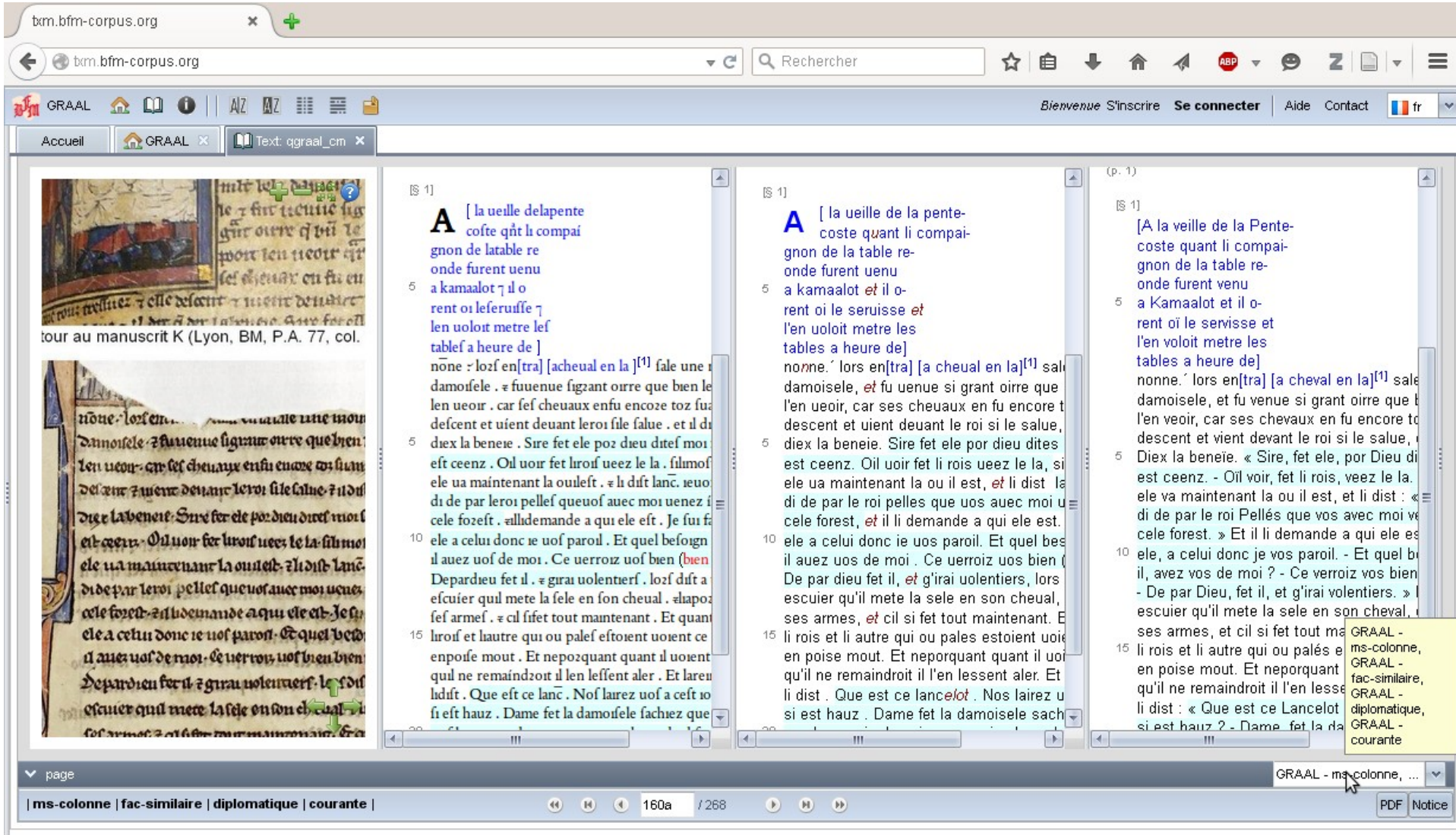
- *u/v* and i/*j* letters normalized
- abbreviation expansions are unmarked
- punctuation normalized, proper nouns capitalized

- Hardly alignable to images
- The most widespread in literary text editions

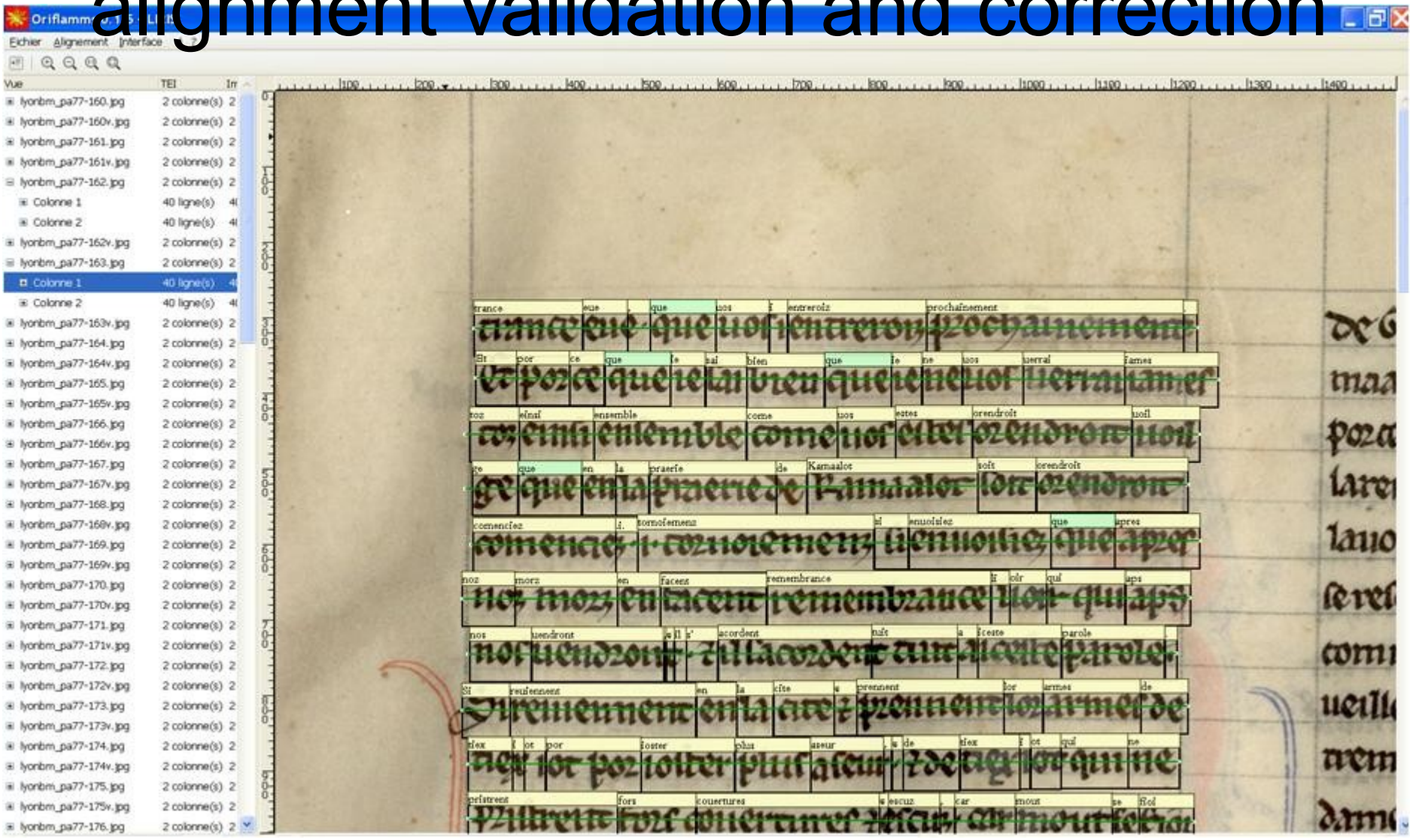
Hybrid Digital Transcriptions

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

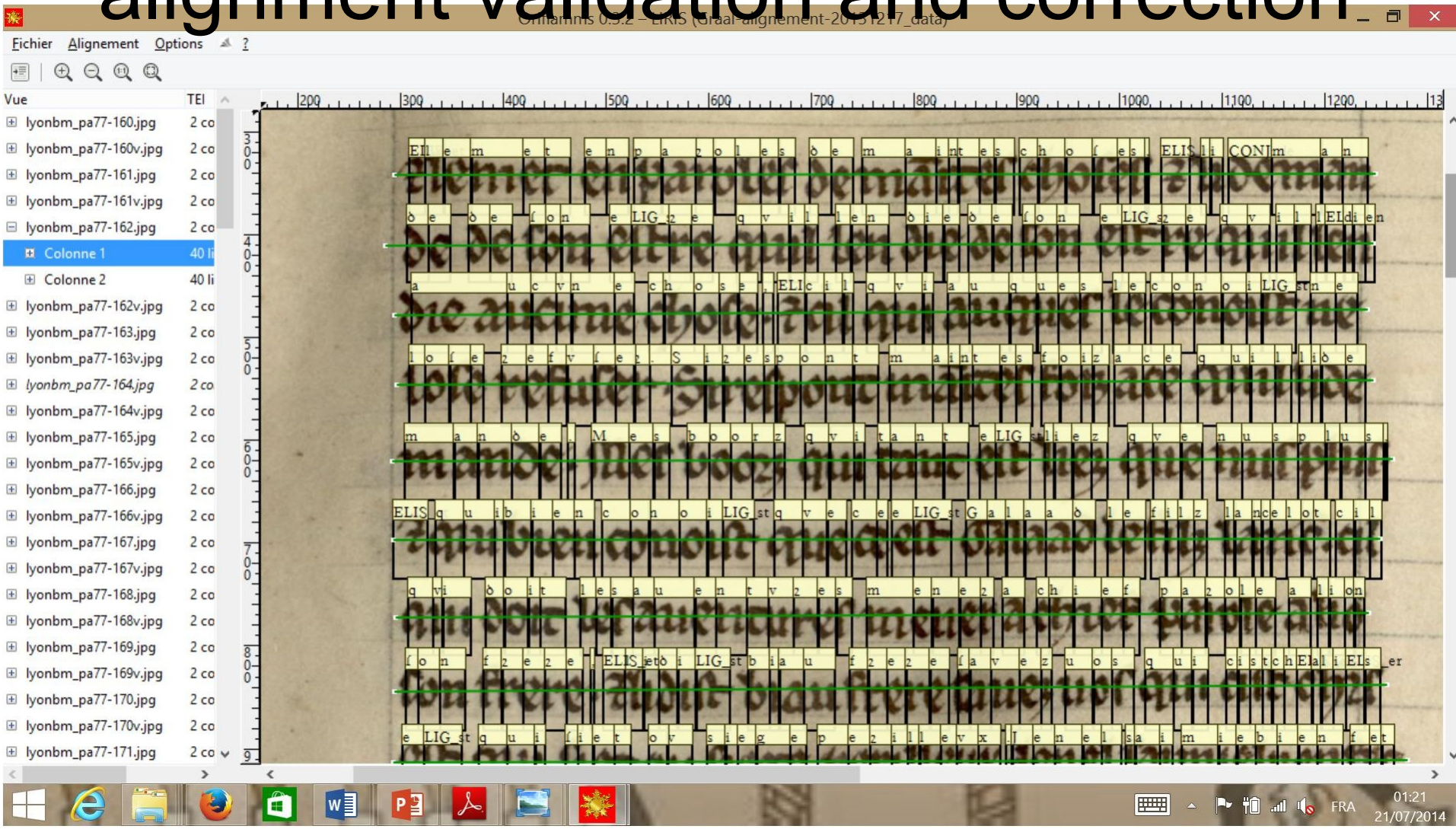
Hybrid Digital Transcriptions



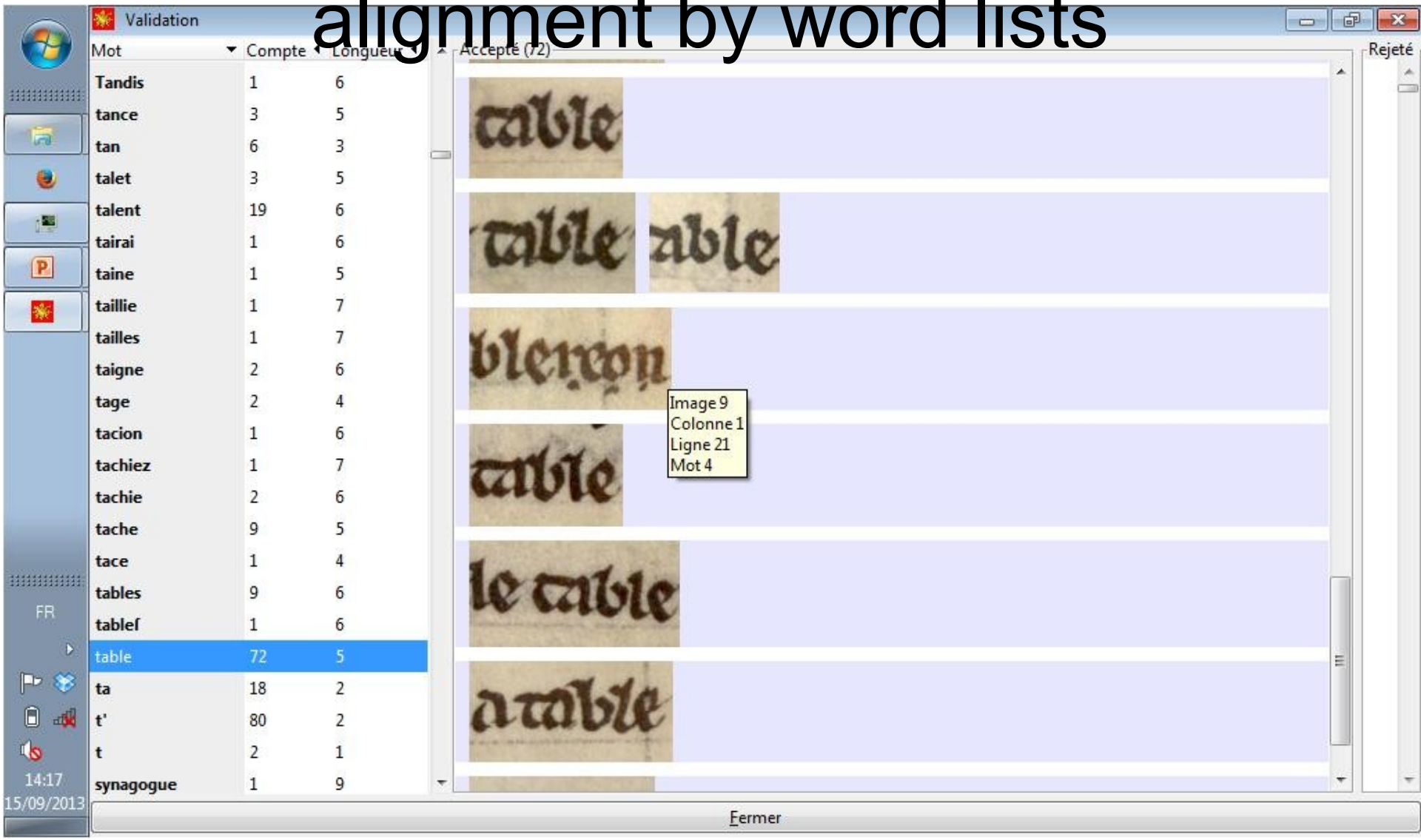
Oriflamms software: word alignment validation and correction



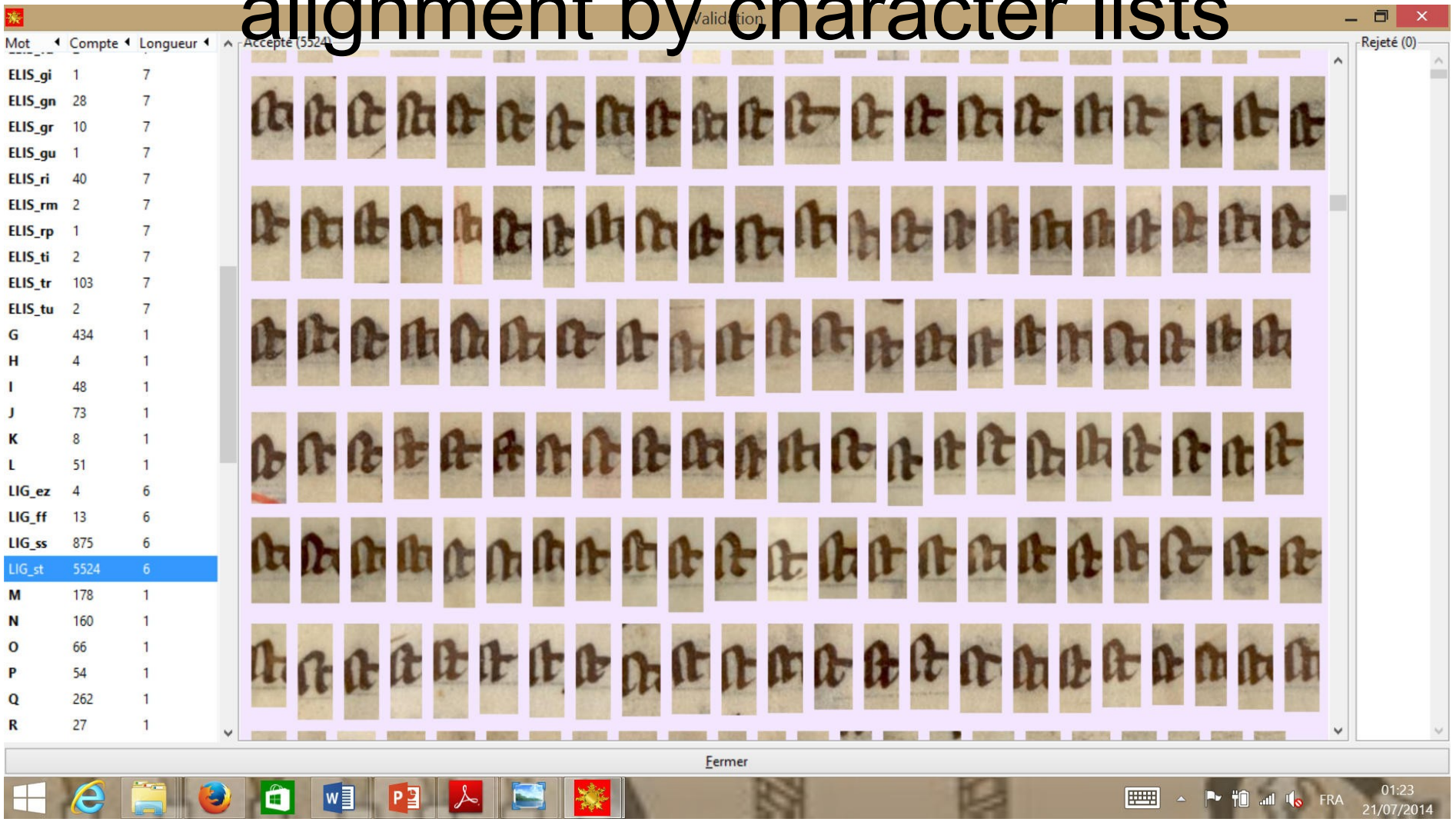
Oriflamms software: character alignment validation and correction



Oriflamms software: validating alignment by word lists



Oriflamms software: validating alignment by character lists



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 / ꝛ
  - 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 ſ 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!