Problems in Linguistic Annotation and Analysis of XML TEI Editions of Textual Heritage Works

ICAR Research laboratory
Université de Lyon, ENS de Lyon, CNRS

El'Manuscript 2012
Petrozavodsk, 3 – 9 September 2012

Outline

- Reasons for using TEI in digital editions and in corpus building
- Possible conflicts between philological markup and linguistic annotation
- Analyzing and enriching critical editions with TXM platform
  - ☐ TEI-BFM customization
  - ☐ tokenization and applying NLP tools
  - ☐ TEI-TXM extension
  - ☐ corpus query and analysis

Reasons for using TEI XML

- Detailed documentation
  - elaborate text structure and markup theory
- Wide community (TEI < XML)
- Constant evolution
  - relatively long backward compatibility
  - tools for markup upgrade
- Opportunities for interchange and using external tools

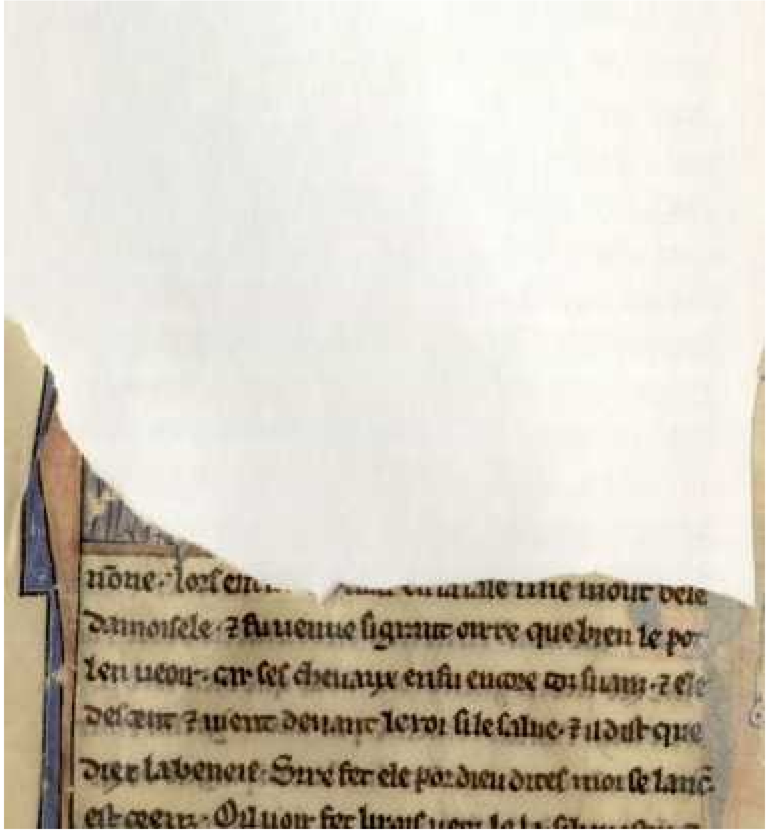
Possible conflicts

- XML does not allow overlapping structures:
  - elements must nest
- Only one hierarchy can be marked up directly
- "Competing" hierarchies can be marked up using
  - joining mechanisms
  - "milestones" (empty tags)
  - stand-off mechanisms
- Sophisticated tools may be necessary to re-build elements of alternative hierarchies

Multiple hierarchies

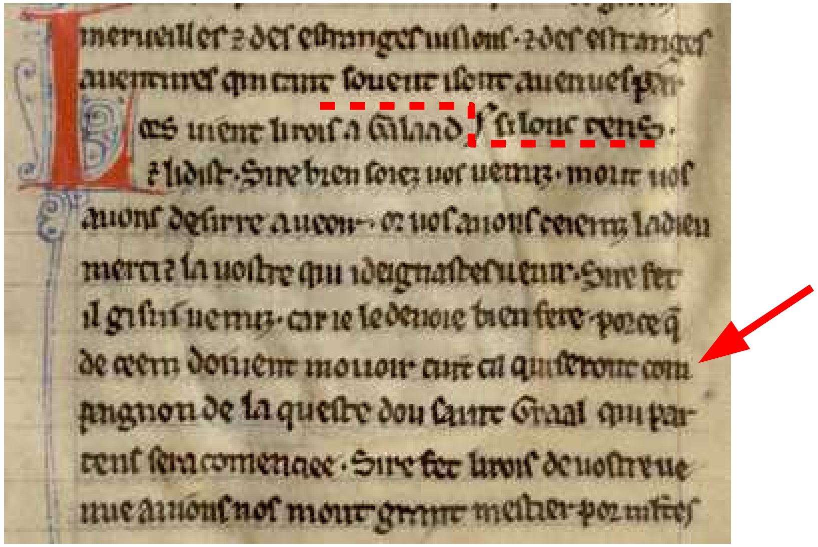
- Physical (geometrical)
  - book > folio > page > column> line...
- Logical (semantic)
  - work > part > paragraph / verse group...
- Rythmic
  - verse...
- Linguistic
  - sentences > phrases > words > morphemes...
- Genetic
  - additions, deletions...

Examples of competing hierarchies



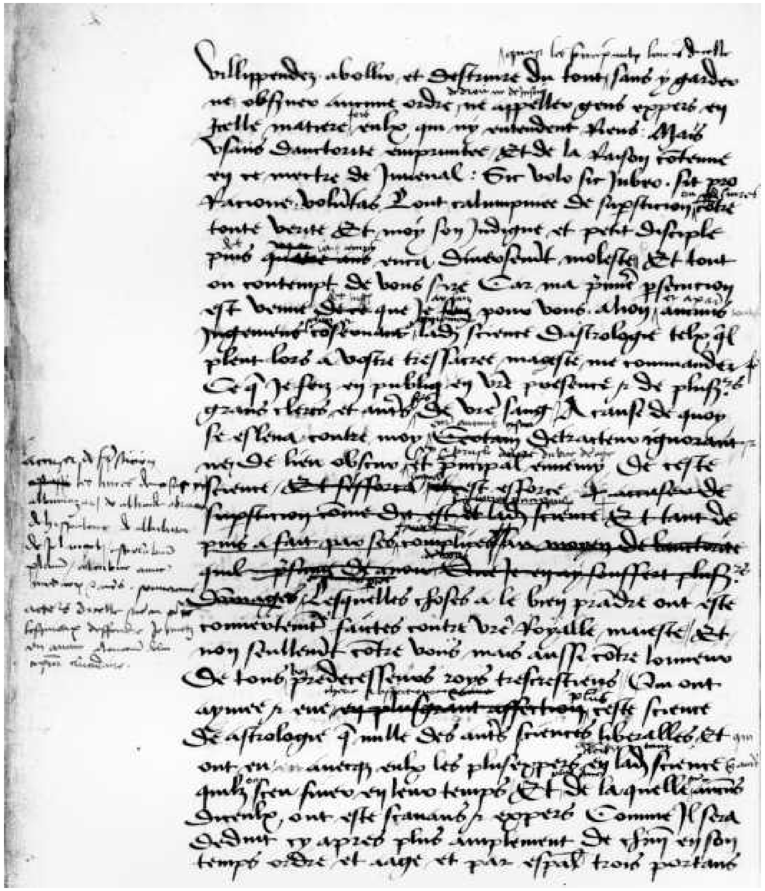
- Linguistic vs. editorial structures
  - <supplied>A la veille de la Pentecoste […] tables a heure de</supplied> nonne lors en<supplied>tra a cheval</supplied> <unclear>en la s</unclear>ale

Compatibility problems



- Linguistic vs. physical structures
  - “overlapping” lines in the end of divisions
  - line/pagebreaks “inside” words
  - “irregular” word segmentation
    - lirois, idegnastesuenir
    - es paule, en mena

Manuscript transcriptions for linguistic research



- If a manuscript contains additions, deletions, corrections, etc. it may be interesting to analyse different states of the text (genetic edition)

Ms. BnF fr. 1357, f° 2v, Simon de Phares, *Recueil des astrologues*

Processing texts with NLP tools

- Lexical items recognition (tokenisation)
  - w
- Sentence splitting
  - s
- POS annotation and lemmatization
  - e.g. TreeTagger
  - export / import
- Inline / standoff annotation

Compatibility with NLP tools

- Tokenizers usually rely on “separator” characters
  - xml editing tools can easily reformat text (insert line breaks and tabs) to make the xml structure more visible ==> spacing characters around xml tags are not reliable
- Sentence parsers rely on “strong” punctuation
  - they can also use structural tags to force sentence breaks (e.g. end of heading)
  - it may be tricky to place s tags correctly in a text with a rich editorial markup

Compatibility with NLP tools

- Rich manuscript transcriptions are likely to be regularly (constantly?) updated
- Standoff markup solutions require stable sources
  - character position for simple text alignment
  - identifier for xml alignment

Corpus query tools

- IMS Open corpus workbench (CWB)
  - CQL query language ← index tables
- TXM
  - TEI (extended)
  - CQP concordances and indexes
  - statistical analysis tools
- TigerSearch
  - querying syntactic structures ← index tables

BFM TEI customization

- Define tag classes with respect to their position in linguistic hierarchy (project-specific)
  - word-internal (c, g, am, ex...)
  - = word (num, abbr)
  - one or more words within a sentence
  - at least one sentence (div, p, head...)
  - out-of-text for NLP tools (note, foreign...)
  - floating (most editorial tags)
    - but may have a “privileged” position
    - additional markup for “unusual” positions (see next slide)

BFM TEI customization

- Deal with the floating tags
  - split, if a tag starts in the middle of a word and ends after a few other words
    - ☐ en<supplied>tra</supplied> <supplied>a cheval</supplied>
  - use special attribute values if a tag is not in its default position (e.g. “word_part” and “multi_s”)
    - ☐ en<supplied txm:range="word_part">tra</supplied>
    - ☐ cf. lb/@type
      - The type attribute may be used to characterize the line break in any respect, but its most common use is to specify that the presence of the line break does not imply the end of the word in which it is embedded. A value such as inWord or nobreak is recommended for this purpose, but encoders are free to choose whichever values are appropriate. (TEI Guidelines)
  - or pre-tag words containing floating tags
    - ☐ <w>en<supplied>tra</supplied></w> <supplied>a cheval</supplied>

TEI TXM Extension

- Systematic tagging of words and sentences
  - <s> / <w> with @xml:id
- Inline markup
  - <txm:form> +
  - <txm:ana> *
- Standoff
  - referring to w/@xml:id
  - updating
    - “inject” annotation in-line before updates (if possible)
    - use diff algorithms to re-align text and markup

Generalization

- Check the tag usage in the document
- Use an XSLT filter
  - before the tokenization
  - maybe also after the tokenization
- Remove xml elements unlikely to used in corpus query
- Deal with typical problems
  - e.g. full stop in the end of a sentence-internal element

Generalization

- Tested on
  - Bouvard et Pécuchet digital edition
    - http://dossiers-flaubert.ish-lyon.cnrs.fr/
  - Bibliothèques virtuelles des humanistes
    - http://www.bvh.univ-tours.fr/
  - Frantext
    - http://www.frantext.fr

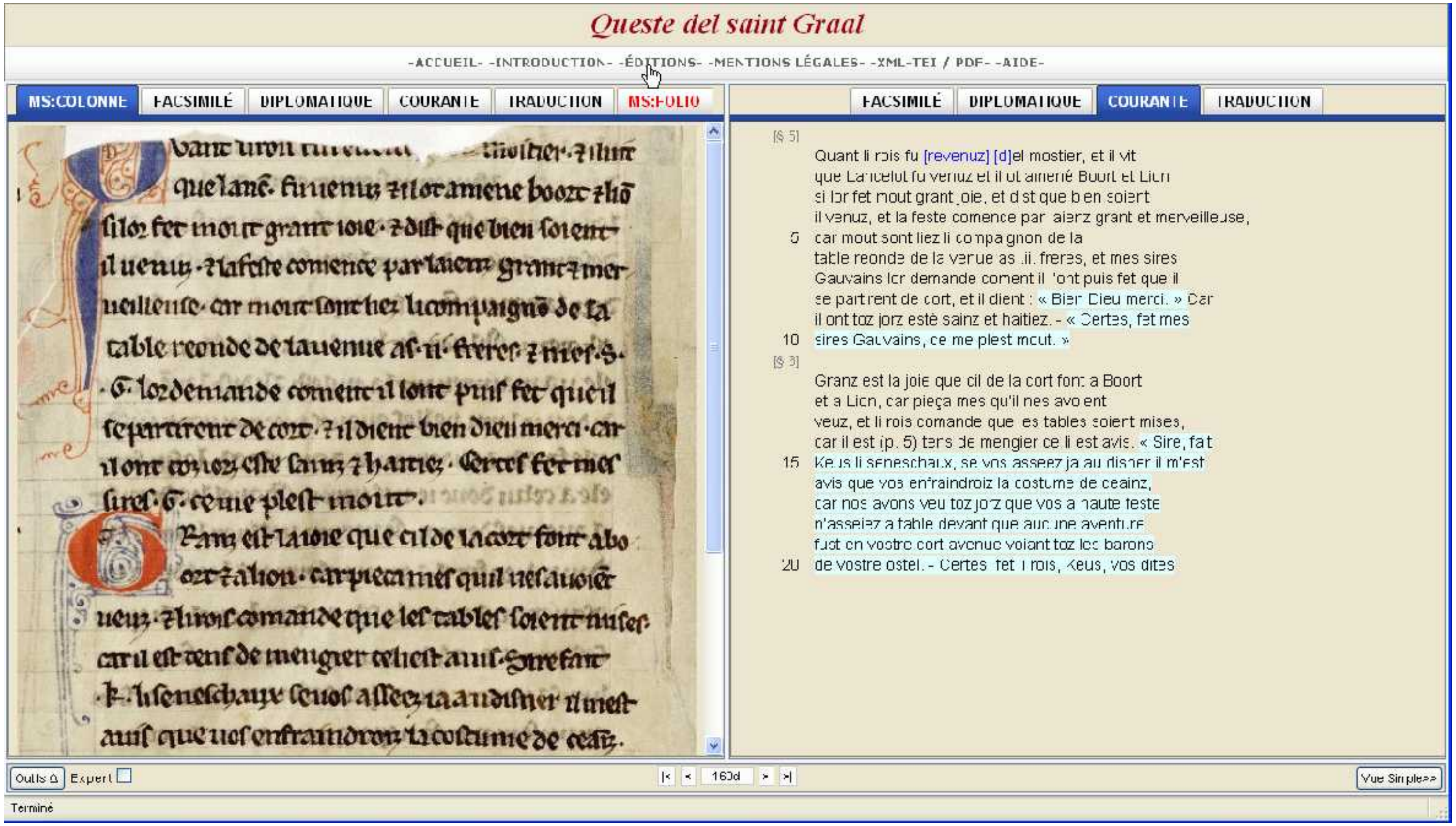
Queste del saint Graal digital edition

- XML markup (tag usage)

Tag	Occurs
w	118 885
lb	10 751
me:facs, me:dipl; me:norm, choice, s	2000 - 3500
q	622
milestone, orig, pb, seg, p	200 - 350
corr, sic, ex, bfm:mdvAbbr	150 - 199
note, bfm:hyphen, supplied	10 - 50
del, hi, bfm:sb, bfm:lettrine	5 - 8
text, subst, damage	1

Queste del saint Graal digital edition

- Interface http://txm.bfm-corpus.org/txm/



[§ 5]

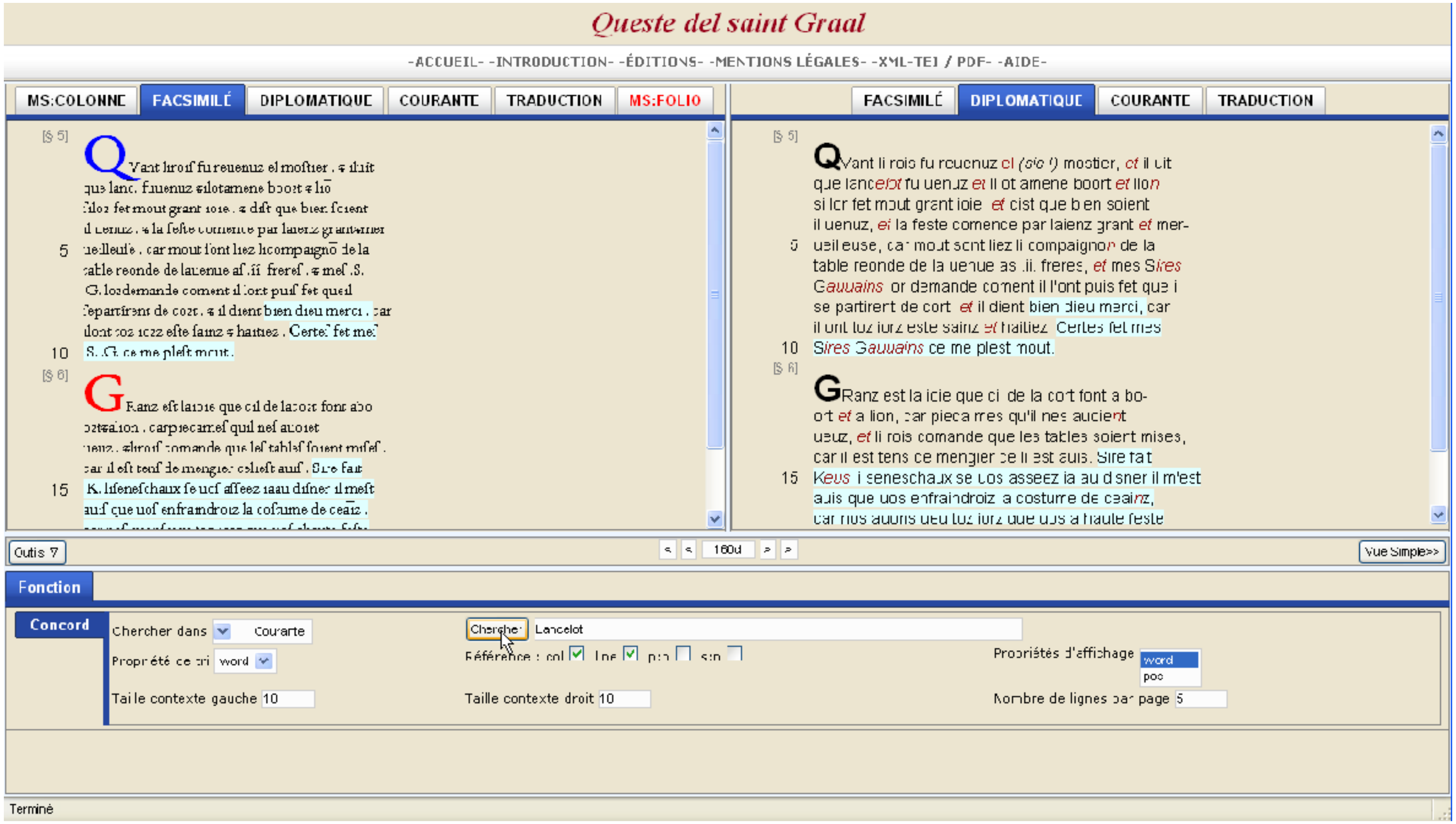
Quant li rois fu [revenuz] [d]el mostier, et il vit que Lancelot fu venuz et il ot amené Boort et Lion si lor fet mout grant joie, et dist que bien soient il venuz, et la feste comence par laienz grant et merveilleuse, car mout sont liez li compagnon de la table reonde de la venue as .ii. freres, et mes sires Gauvains lor demande coment il l'ont puis fet que il se partrent de cort, et il dient : « Bien Dieu merci. » Car il ont toz jorz esté sainz et haitiez. - « Certes, fet mes sires Gauvains, ce me plest mout. »

[§ 3]

Granz est la joie que cil de la cort font a Boort et a Lion, car pieça mes qu'il nes avoient veuz, et li rois comande que les tables soient mises, car il est (p. 5) tens de mengier ce li est avis. « Sire, fait Keus li seneschaux, se vos asseez ja au disner il m'est avis que vos enfraindroiz la costume de ceainz, car nos avons veu toz jorz que vos a haute feste n'asseiez a table devant que aucune aventure fust en vostre cort avenue voiant toz les barons de vostre ostel. - Certes, fet li rois, Keus, vos dites

Queste del saint Graal digital edition

- Interface http://txm.bfm-corpus.org/txm/



[§ 5]

QVant li rois fu reuenuz el *(sic !)* mostier, et il uit que lancelot fu uenuz et il ot amene boort et lion si lor fet mout grant ioie et cist que bien soient il uenuz, et la feste comence par laienz grant et merueilleuse, car mout sont liez li compaignon de la table reonde de la uenue as .ii. freres, et mes Sires Gauuains or demande coment il l'ont puis fet que il se partirent de cort et il dient bien dieu merci, car il ont toz iorz este sainz et haitiez. Certes fet mes Sires Gauuains ce me plest mout.

[§ 6]

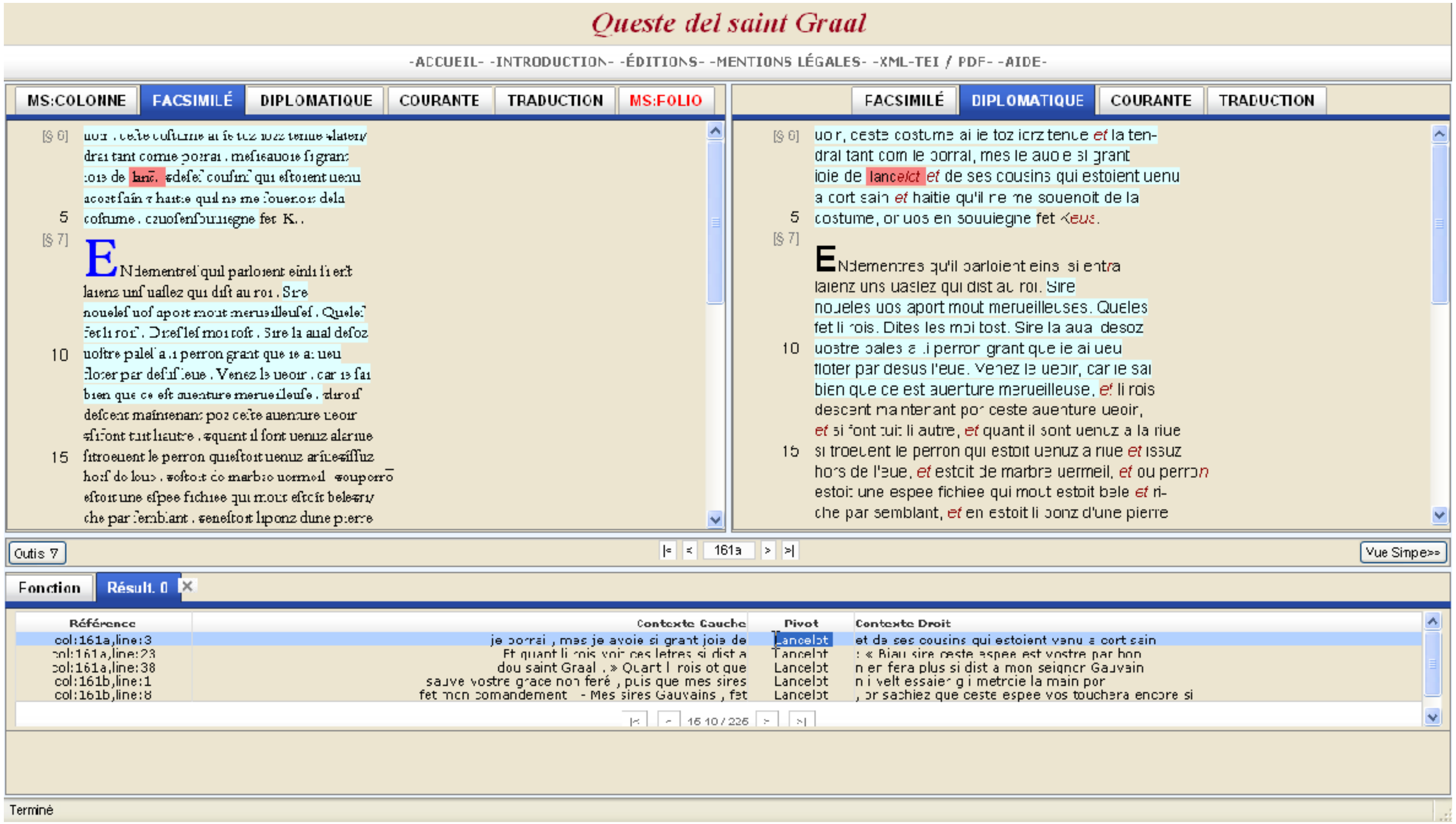
GRanz est la ioie que cil de la cort font a boort et a lion, car pieca mes qu'il nes auoient ueuz, et li rois comande que les tables soient mises, car il est tens de mengier se li est auis. Sire fait Keus li seneschaux se uos asseez ia au disner il m'est auis que uos enfraindroiz la costume de ceainz, car nos auons ueu toz iorz que uos a haute feste

Chercher dans: Courante
Propriété de tri: word
Taille contexte gauche: 10
Lancelot
Référence : col ☑ lne ☑ pid ☐ sid ☐
Taille contexte droit: 10
Propriétés d'affichage: word, pos
Nombre de lignes par page: 5

Terminé

Queste del saint Graal digital edition

- Interface http://txm.bfm-corpus.org/txm/



[§ 6] uoir, ceste costume ai ie toz iorz tenue *et* la tendrai tant com ie porrai, mes ie auoie si grant ioie de lancelot *et* de ses cousins qui estoient uenu a cort sain *et* haitie qu'il ne me souenoit de la costume, or uos en souuiegne fet Keus.

[§ 7] ENdementres qu'il parloient einsi si entra laienz uns uaslez qui dist au roi. Sire noueles uos aport mout merueilleuses. Queles fet li rois. Dites les moi tost. Sire la aual desoz uostre pales a .i. perron grant que ie ai ueu floter par desus l'eue. Venez le ueoir, car ie sai bien que ce est auerture merueilleuse, *et* li rois descent maintenant por ceste auenture ueoir, *et* si font tuit li autre, *et* quant il sont uenuz a la riue si troeuent le perron qui estoit uenuz a riue *et* issuz hors de l'eue, *et* estoit de marbre uermeil, *et* ou perron estoit une espee fichiee qui mout estoit bele *et* riche par semblant, *et* en estoit li ponz d'une pierre

Référence	Contexte Gauche	Pivot	Contexte Droit
col:161a,line:3	je porrai , mes je avoie si grant joie de	Lancelot	et de ses cousins qui estoient venu a cort sain
col:161a,line:23	Et quant li rois voit ces letres si dist a	Lancelot	: « Biau sire ceste espee est vostre par bon
col:161a,line:38	dou saint Graal . » Quant li rois ot que	Lancelot	n en fera plus si dist a mon seignor Gauvain
col:161b,line:1	sauve vostre grace non feré , puis que mes sires	Lancelot	n i velt essaier g i metrcie la main por
col:161b,line:8	fet mon comandement - Mes sires Gauvains , fet	Lancelot	, or sachiez que ceste espee vos touchera encore si

Thank you!