Школа "Письменное наследие и современные информационные технологии"
Ижевск, 15 октября 2009 г.

Исследование корпуса с помощью Weblex и работа с электронным изданием «Поисков Святого Грааля»

Преподаватели: А.М. Лаврентьев и С. Эйден (Serge Heiden)
(Национальный центр научных исследований и Лионский университет)

1. Weblex

1.1. Введение

Weblex – это прототип онлайновой исследовательской платформы, предоставляющей доступ к различным корпусам, включающий поисковую машину и статистические инструменты для проведения текстометрических исследований. Его адрес в Интернете:

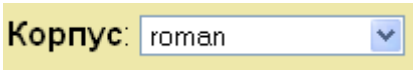
http://weblex.ens-lsh.fr/wlx

Со страницы доступа к Weblex перейдите по ссылке на Русский интерфейс, затем наберите ‘roman’ в поле [Корпус] и нажмите на кнопку [Начало].

Мы будем работать с демо-корпусом, состоящим из 49 отрывков текстов XVIII-XX вв. Среди их авторов – Вальтер Скотт, Чарльз Диккенс, Редьярд Киплинг, Вирджиния Вульф, Эрнест Хемингуэй, Джордж Оруэлл и др. Цель этого корпуса – сравнительное исследование британских и американских текстов в прозе.
В начале занятия мы рассмотрим язык запросов поисковой машины, позволяющей исследовать словарный состав и морфологические характеристики слов. Далее мы представим статистический инструмент «Специфичность» (Specificity), который используется для сравнения текстовых свойств подкорпусов, а затем познакомимся с несколькими статистическими инструментами, позволяющиит исследовать соупотребление (collocation) слов и строить лексико-семантические сети.

1.2. Первые шаги

1.2.1. Выбираем корпус и изучаем его свойства

Кнопка «Корпус»: 

Запросы могут обращаться ко всему корпусу, к отдельному тексту, к одному подкорпусу или к ряду подкорпусов для сопоставительного анализа.

Кнопка «Размер»
Эта кнопка позволяет выводить на экран информацию о размере корпуса или подкорпуса (число текстоформ, различных словоформ и предложений).

Кнопка «Словарь»
Эта кнопка позволяет вывести на экран состав словоформ текста в алфавитном и частотном порядке. В лемматизированных и морфологически размеченных корпусах имеется возможность просмотра состава лемм и морфологических категорий.

1.2.2. Что мы ищем?

- **Определенную словоформу**

Если набрать *this* или *look* в поле [Искать], машина будет искать употребления словоформ *this* и *look*, но не *This* (Weblex учитывает регистр), ни *looks*, ни различные орфографические варианты, которые можно встретить в старых текстах.

- **Лемму** («начальную форму»)

Автоматические лемматизаторы показывают хорошие результаты на материале современных текстов с нормированной орфографией, в старых текстах с вариативной орфографией их эффективность может быть значительно ниже. Тем не менее, язык запросов CQP позволяет осуществлять своего рода «лемматизацию на лету» (см. 1.3.1), что в какой-то мере снимает остроту проблемы.

- **Часть слова или леммы** (корень, префикс, суффикс…)

- **Несколько слов, следующих подряд или разделенных несколькими другими словами, например** *young woman* или *meet (with a) woman*.

1.2.3. В каком формате получить результат?

Результаты запроса могут быть представлены в виде конкорданса, т.е. списка всех текстоформ, отвечающих запросу (отдельное слово или синтагма), и их контекстов. Чтобы получить конкорданс, нажмите на кнопку [Составить Конкорданс]. Вы можете задать размер контекста в печатных знаках (150 по умолчанию).
Вы можете получить список частотности форм, отвечающих вашему запросу, нажав на кнопку [Индекс].

1.3. Как это делается? Использование выражений CQP

1.3.1. Ищем лемму

Выражения CQP составляются на «формальном» языке, основанном на использовании «специальных знаков», подобных тем, что применяются в «регулярных выражениях» (regular expressions). Они позволяют указать существенные для запроса свойства словарной единицы корпуса. Далее мы представим несколько образцов выражений CQP, наиболее востребованных при работе с Weblex. Более подробную информацию можно получить на сайте:
http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPSyntax.html.

Символ `?` используется для указания на то, что «предшествующий знак факультативен». Если вы введете запрос `looks?`, вы получите как «базовую» форму английского глагола, так и форму 3-го л. ед. ч. настоящего времени на *-s*. `%c` означает «без учета регистра», а кавычки обозначают границы словарной формы. Таким образом, выражение `"looks?"%c` позволит извлечь из корпуса все формы «простого настоящего», начинающиеся со строчной или с заглавной буквы.

> Следует заметить, что корпус, используемый для настоящего занятия, был отформатирован с заменой всех прописных букв строчными.

Выражение `[sz]` означает `s` или `z`, таким образом, запрос `enterpri[sz]es?` позволит отыскать в корпусе все употребления следующих форм: *enterprise*, *enterprises*, *enterprize* и *enterprizez*.

При поиске слов в корпусе следует вспомнить все формы словоизменительной парадигмы, а также вероятные орфографические варианты, а затем написать выражение CQP, позволяющее их извлечь.

С помощью кнопки [Индекс] вы можете получить указатель частотности форм, отвечающих заданному выражению CQP в исследуемом корпусе. Таким способом можно проверить, позволяет ли заданное выражение извлечь все представленные в корпусе интересующие вас формы и не создает ли оно слишком много «шума» (извлекая формы, вас не интересующие). Так, вы можете запросить различные формы слова ENTERPRISE в корпусе, задав «расширенное» запросное выражение, содержащее *точку*, специальный символ, соответствующий любому знаку. Набрав `"ent.*i.*e.*"` (значение «звездочки» мы рассмотрим ниже), вы получите все возможные формы леммы, но также и такие формы, как *entertained* или *entreaties*, которые нас не интересуют.

Использование оператора «любой знак» весьма целесообразно при работе со старыми текстами с их высокой вариативностью написания слов, поскольку может быть очень сложно заранее «вычислить» все возможные графические варианты.

1.3.2. Ищем часть слова или леммы

Если вы наберете `acquaint.*`, вы не только отыщете все формы глагола *acquaint*, но также и образованное от него существительное *acquaintance*. Точка означает «любой знак», а звездочка – «предшествующий знак факультативен и может повторяться сколько угодно раз».

Аналогичным способом можно искать существительные, оканчивающиеся на *-ism(s)* или на *-tion(s)/-sion(s)/-cion(s)*, используя выражения `".+isms?"` и `".+[cst]ions?"` соответственно (`x+` означает ‘не менее одного знака *x*’).

1.3.3. Выражение альтернативы: оператор |

Мы можем одновременно искать в корпусе употребления слов *dog(s)* и *cat(s)*, используя следующее выражение: `(dog|cat)s?`. Скобки определяют границы действия оператора `|`.
Следующие два выражения эквивалентны: `enterpri[sz]e` и `enterpri(s|z)e`.

1.3.4. Ищем последовательности слов

Young woman
`"young" "wom[ae]n"`: кавычки используются для указания границ отдельного слова.

Тот же самый запрос можно сформулировать более эксплицитно:
`[word="young"] [word="wom[ae]n"]`

NB 1. Если набрать в выражении пробел без кавычек, для CQP это будет означать, что речь идет об одном составном слове (как английское *look for* или русский союз *потому что*). В зависимости от предварительной подготовки корпуса, составные слова в нем могут быть или не быть размечены. Запрос `[word=".* .*"]` позволяет проверить наличие в корпусе составных слов (т.е. слов, содержащих пробел).
NB 2. Обратите внимание на два «уровня» употребления квадратных скобок: обозначение словарных единиц (1) и выражение списка буквенных знаков (2).

Meet ... (a) woman
`"meet" []* "woman" within 6`.
Здесь `[]*` означает любое количество любых слов, а `within 6` позволяет ограничить размеры искомого выражения (в числе слов).

1.3.5. Выражаем то, что необходимо «отфильтровать»

Формы глагола и отглагольные существительные на -ing

```
[word=".+ing" & word!="(some|any|no|every)?thing"%c]
```

Это выражение позволяет получить все слова на *–ing* и «отфильтровать» такие частотные и не интересующие нас в данном случае слова, как *thing* и его дериваты. Оператор `&` между двумя выражениями `word="…"` внутри одной словарной единицы, отмеченной квадратными скобками, означает сочетание двух условий поиска. Оператор `!=` во втором выражении создает негативное условие. Иными словами, мы ищем текстоформы, оканчивающиеся на *-ing*, и исключаем из результата те, что отвечают поисковому запросу `"(some|any|no|every)?thing"%c`.

1.3.6. Ищем свойства слова, например, к какой части речи оно относится

Если наш корпус содержит лингвистическую разметку (например, указание частей речи), этой разметкой можно воспользоваться для уточнения поискового запроса.
Все слова используемого на нашем занятии корпуса были автоматически размечены с помощью программы LTPOS (http://www.ifi.uzh.ch/arvo/cl/broder/tttdoc/c1040.htm) с использованием классификации, принятой в UPenn TreeBank (http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html).

С помощью следующего выражения получаем все употребления междометий:
`[P2="UH"]`.

Данное выражение позволяет выявить все случаи употребления слова *man* в качестве междометия («слова-паразита»):
`[P2="UH" & word="man"]`.

А такое выражение позволяет отыскать все примеры, в которых за словом *man* следует глагол в пределах одного предложения:

```
"man"%c []* [P2="V.*"] within s
```

Следует заметить, что при поиске категорий можно использовать те же специальные символы, что и при поиске форм слов.

1.4. Резюме

- Выберите корпус.
- Составьте предварительный список словоформ или выражений, которые вас интересуют.
- Составьте выражение CQP.
    - **Совет:** используйте выражение `.*`, чтобы на первом этапе не ограничивать жестко список форм и обнаружить графические варианты, которые было трудно предусмотреть заранее.

- Используйте кнопку [Индекс], чтобы проверить, какие словоформы корпуса соответствуют запросу.
- При необходимости сформулируйте более «жесткое» выражение CQP, чтобы сократить «шум».
- Нажмите на кнопку [Составить Конкорданс], чтобы получить конкорданс KWIC.

Напоминание: поисковая система Weblex учитывает регистр и диакритические знаки. Используйте операторы %c, %d или %cd, чтобы изменить эти настройки.

NB 3. Некоторые знаки препинания «ведут себя» как «специальные символы» в языке запроса. Если вы хотите искать знаки препинания в корпусе, используйте перед ними символ \.

NB 4. Пробелы учитываются «внутри» словоформы (поиск составных слов), но не имеют значения между текстоформами.

NB 5. Некоторые символы (такие как кавычки или точка с запятой) не допускаются в выражениях CQP. Если необходимо найти их в корпусе, следует использовать восьмеричный код с предшествующей тройной обратной косой чертой: \\\.
Например, `"\\\042"` означает двойные кавычки, а `"\\\073"` следует использовать для поиска точки с запятой.

1.5. Изучаем частотность употребления частей речи (необходим полный интерфейс на французском)

См. перевод терминов в словаре.

- Corpus : roman-sdal
- Source A : `[P2="POS"]` *Possessive ending*
- Source B : `[P2="PRP$"]` *Possessive pronoun*

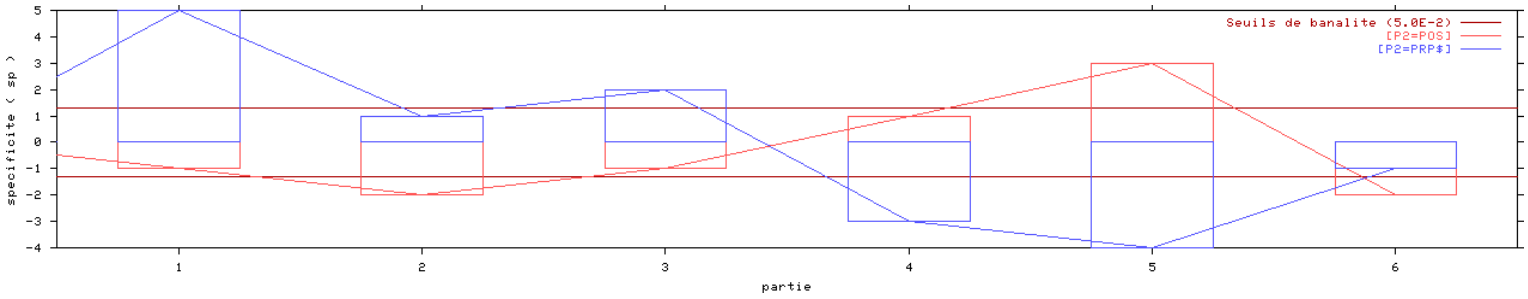
Tableau des spécificités des expressions :

[P2="POS"] , [P2="PRP$"]

du corpus roman-sdal

Partie		1		2		3		4		5		6	
T	58023	5988		1372		11307		7875		11891		18337	
	F	f	sp	f	sp	f	sp	f	sp	f	sp	f	sp
[P2="POS"]	250	22	-1	2	-2	48	-1	36	1+	69	3+	68	-2
[P2="PRP$"]	1648	219	5+	46	1+	353	2+	187	-3	281	-4	508	-1

1. roman-sdal-18GB
2. roman-sdal-18US
3. roman-sdal-19GB
4. roman-sdal-19US
5. roman-sdal-20GB
6. roman-sdal-20US



Index des occurrences de [][P2="POS"] dans le corpus roman-sdal-20GB

ord	f	événement
1	6	. '
2	4	one 's
3	3	gudrun 's
4	3	kenton 's
5	3	man 's
6	3	other 's
7	3	ursula 's
8	2	butler 's
9	2	farraday 's
10	2	people 's
11	2	ricardo 's
12	1	'see '
13	1	bookmakers '
14	1	castle 's
15	1	claridge 's
16	1	davis 's
17	1	durtnall 's
18	1	eileen 's
19	1	else 's
20	1	employer 's
21	1	england 's
22	1	father 's
23	1	friend 's
24	1	gentleman 's
25	1	hour 's
26	1	james 's
27	1	jim 's
28	1	jones 's
29	1	ladies '
30	1	lady 's
31	1	lucy 's
32	1	major 's
33	1	men 's

```
34     1 name 's
35     1 newscaster 's
36     1 novelist 's
37     1 party 's
38     1 peter 's
39     1 pie '
40     1 rumplemayer 's
41     1 sailor 's
42     1 shepherd 's
43     1 signora 's
44     1 watney 's
45     1 wives '
46     1 world 's
47     1 years '

      69 au Total
```

Index des occurrences de *[P2="PRP$"]*

```
ord     f événement

  1  102 his
  2   52 her
  3   49 their
  4   39 my
  5   17 our
  6   14 its
  7    8 your
```

1.6. «Соупотребления» слова woman в подкорпусе 'roman-sdal-20US'

- Corpus : roman-sdal-20US
- Source A : woman
- minimum frequency (минимальная частотность) : 2
- minimum meeting (минимальная встречаемость) : 2
- Кнопка: Lexicogramme

Seuils : **f** 2, **cf** 2, **p** 9.0E-1, $\mathbf{d_m}$ 1000.0

woman

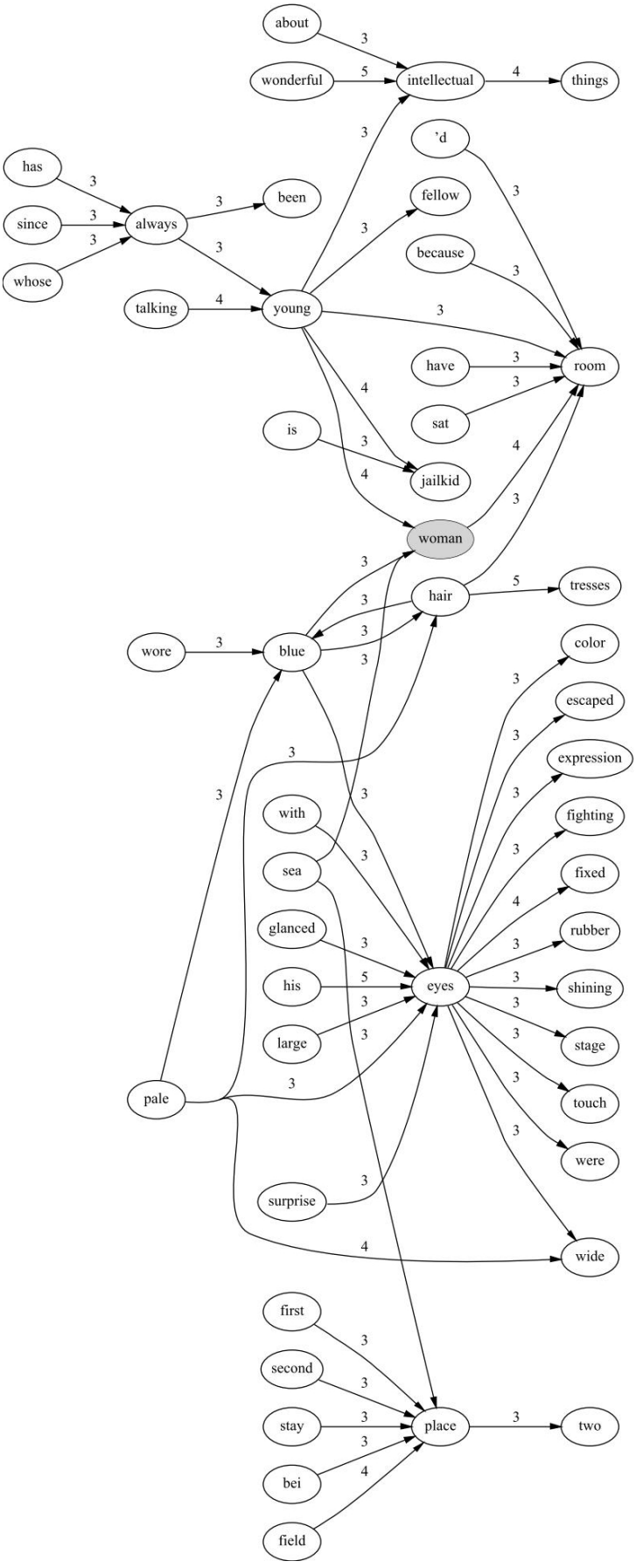
(9)

cooccurrents gauches					cooccurrents droits				
	f	cf	p	$\mathbf{d_m}$		f	cf	p	$\mathbf{d_m}$
young	22	3	6e-04	0.0	room	6	2	1e-03	3.0
sea	7	2	1e-03	37.5	an	47	2	5e-02	11.0
blue	14	2	5e-03	21.5	's	94	2	2e-01	3.0
little	33	2	3e-02	15.5	in	330	3	4e-01	8.3
that	200	4	4e-02	10.2	her	194	2	4e-01	3.0
an	47	2	5e-02	16.5					
she	128	2	3e-01	21.0					
on	149	2	3e-01	27.5					
with	172	2	4e-01	21.5					

1.7. График соупотреблений словоформы woman в подкорпусе 'roman-sdal-20US'

- Corpus : roman-sdal-20US
- Source A : woman
- minimum frequency : 2
- minimum meeting : 2
- Кнопка : Lexicogramme récursif

Seuils : **p** 9e-03, **r** 2, **f** 2, $\mathbf{d_m}$ 1000.0, **pl** 3


about
wonderful
intellectual
things
has
since
whose
always
been
talking
young
'd
fellow
because
have
room
sat
is
jailkid
woman
hair
tresses
wore
blue
color
escaped
expression
fighting
fixed
rubber
shining
stage
touch
were
wide
with
sea
glanced
his
large
eyes
pale
surprise
first
second
stay
bei
field
place
two

2. Грааль / Graal

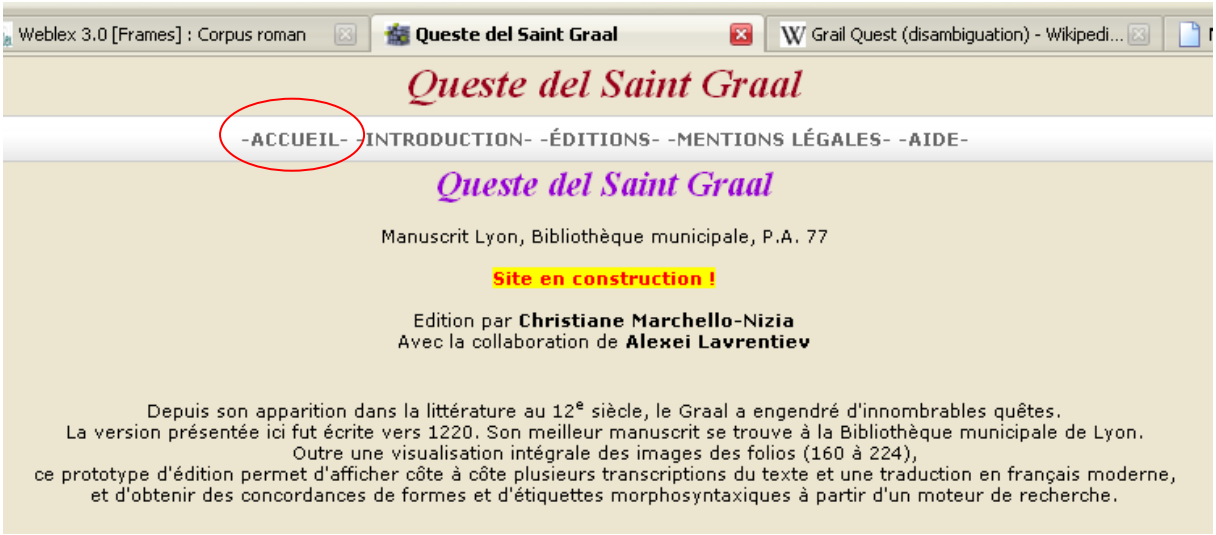
2.1. Общая информация

«Грааль» - это электронное интернет-издание романа «*La queste del saint Graal*» («Поиски святого Грааля»), на материале рукописи Лионской муниципальной библиотеки P.A. 77, под общей редакцией К. Маркелло-Низья (Christiane Marchello-Nizia).
Издание включает:

- введение;
- цветные фотографии 418 колонок рукописи;
- «многоуровневое» издание старофранцузского текста, состоящее из:
  - «традиционного» (нормализованного) представления всего текста;
  - «дипломатического» (графематического) представления всего текста (в процессе подготовки);
  - «имитативного» (аллографического) представления нескольких страниц;
- частеречную разметку;
- перевод на современный французский язык;
- текстометрический инструментарий (в настоящий момент действует генератор конкордансов).

Прототип издания доступен по адресу http://textometrie.risc.cnrs.fr/txm/ (временный адрес издания в процессе разработки, в настоящий момент действует только французский интерфейс).
Прототип протестирован с использованием веб-браузера Mozilla Firefox (v. 3.5.3).

2.2. Accueil (Заглавная страница)



Эта страница содержит общую информацию об издании.

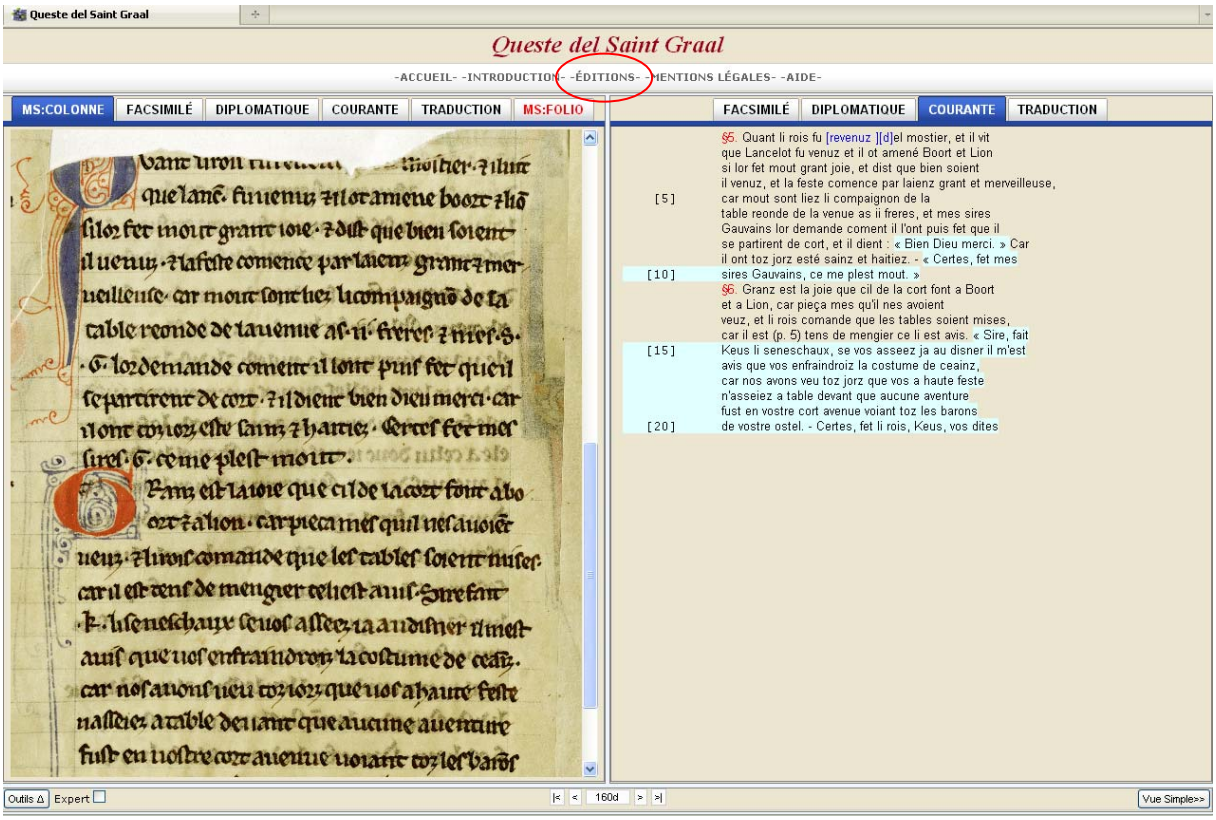
2.3. Introduction (Страница введения)

На этой странице отображается научное введение в издание.

2.4. Editions (Многоуровневое издание)

2.4.1. Просмотр и навигация по различным уровням издания

Закладка Editions предоставляет доступ к основной части (интерфейсу) издания.

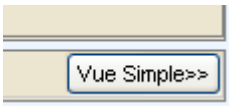


Перед нами две колонки с двумя параллельными уровнями представления текста. Вы можете сопоставить любые два представления:

- MS:COLONNE – цветная **фотография** колонки рукописи;
- FACSIMILÉ – детальная (**аллографическая**) транскрипция рукописи (включает варианты букв, сокращения, средневековые знаки препинания);
- DIPLOMATIQUE – упрощенная (**графематическая**) транскрипция рукописи (с выделенной курсивом расшифровкой сокращений и употреблением букв *i/j* и *u/v*, как в источнике);
- COURANTE – традиционное (**нормализованное**) представление издания (расшифровки сокращений не выделяются, используются современные диакритика и пунктуация, «фонетическое» употребление букв *i/j* и *u/v*);
- TRADUCTION – **перевод** на современный французский;
- MS:FOLIO – цветная фотография целой **страницы** рукописи (занимает весь экран).

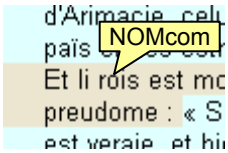
NB 6. В настоящий момент аллографическая и графематическая транскрипции представлены только для колонок с 160d по 161d.

NB 7. Для отображения «нестандартных» знаков средневековой графической системы в аллографической транскрипции необходимо установить шрифт Andron Scriptor Web (v. 3), который можно загрузить с сайта MUFI (http://www.mufi.info/fonts/#Andron).



В любое время можно перейти к отображению только одного уровня представления, нажав на кнопку [Vue Simple>>] в правом нижнем углу окна.

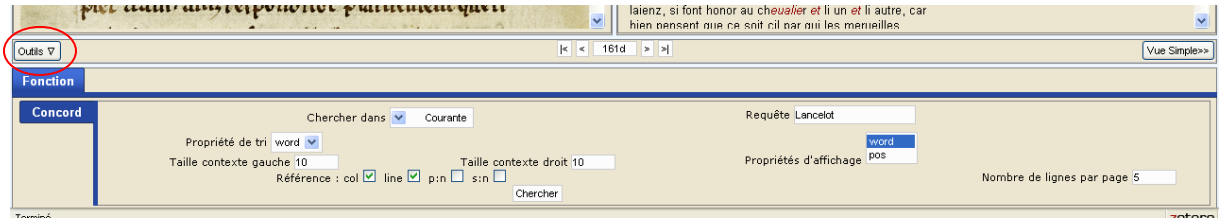
Вы можете перемещаться по тексту, используя панель навигации, расположенную в середине нижней части окна.



Часть речи высвечивается форме «всплывающей» подсказки при наведении мыши на то или иное слово.

2.4.2. Использование текстометрического инструментария

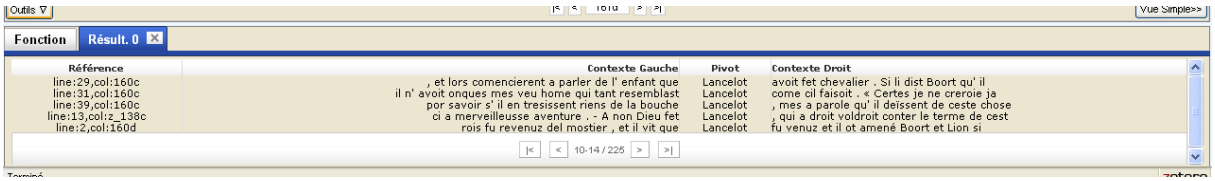
Текстометрическую поисковую форму можно открыть, нажав на кнопку [Outils] в нижнем левом углу окна.



Эта форма напоминает упрощенный интерфейс Weblex.

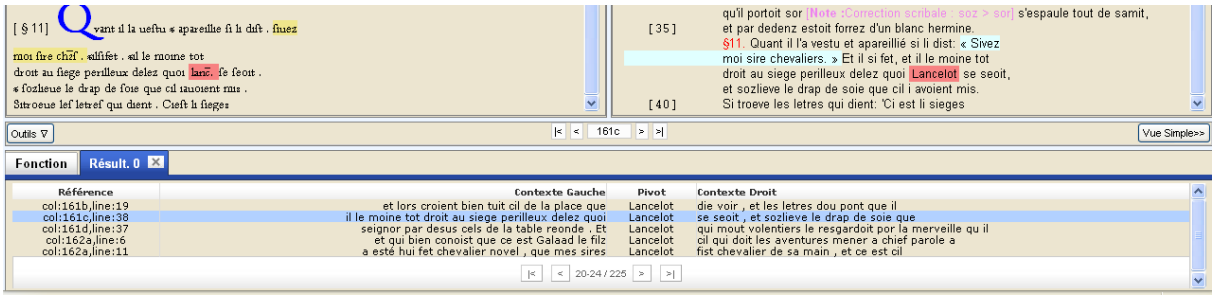
Вы можете вести поиск на одном из уровней представления текста издание (выбрав его в поле [Chercher dans]), выбрать параметры сортировки (поле [Propriété de tri]), задать размер контекста (поле [Taille de contexte…]), число строк, которые выводятся на экран на одной странице конкорданса (поле [Nombre de lignes par page]) и формат ссылок (поля [Référence]).

Вы можете ввести запрос CQP в поле [Requête] и получить соответствующий конкорданс, нажав на кнопку [Chercher].



Конкорданс будет выведен на экран в новой закладке рядом с запросной формой.

Если щелкнуть мышкой по строчке конкорданса, соответствующая колонка текста появится в верхней части окна, а представленная в конкордансе форма будет выделена (розовым фоном):



2.5. Mentions légales (Авторские права)

На этой странице приводится информация об авторах проекта, спонсорах и условиях использования издания (лицензия).

2.6. Aide (Помощь)

Эта страница содержит инструкции по установке шрифта Andron Scriptor Web. В будущем здесь будут собраны ответы на часто задаваемые вопросы и разного рода педагогические материалы.