

Инструкция по пользованию поисковой программой NoSketch Engine для корпуса текстов бамана Corpus bambara de référence

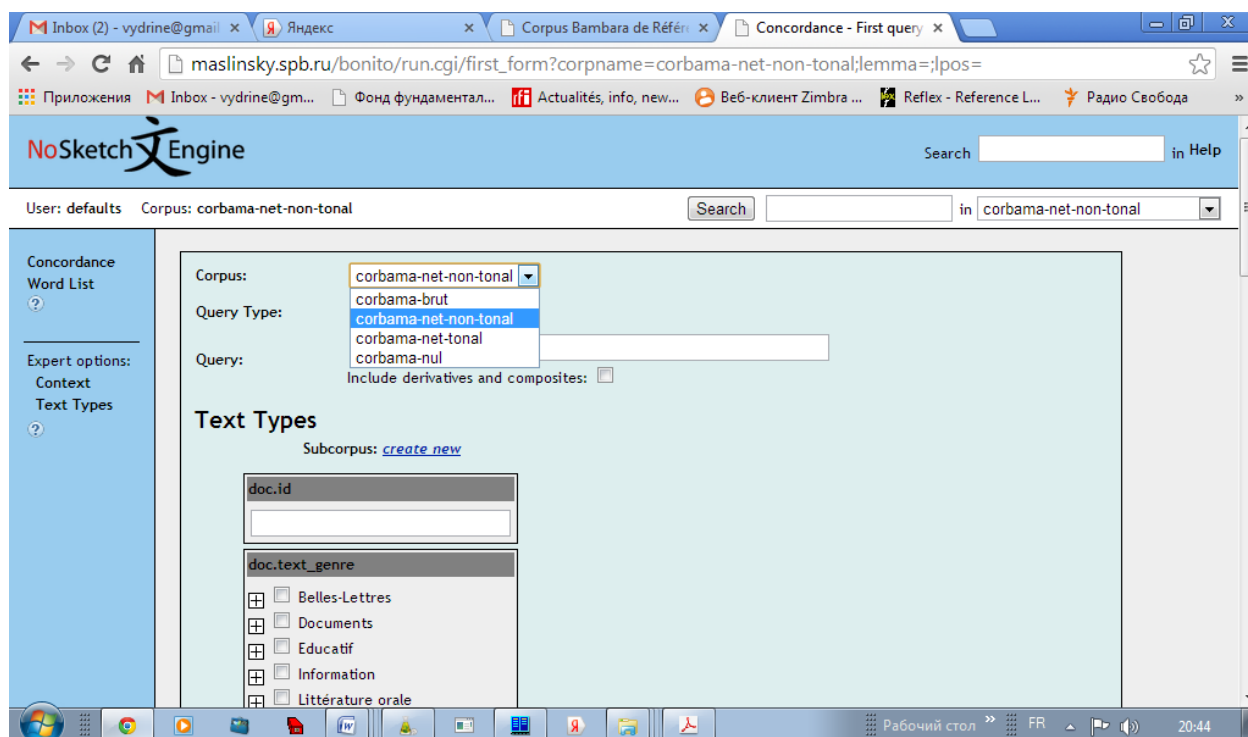
0. Корпус находится в открытом доступе по адресу:

http://maslinsky.spb.ru/bonito/run.cgi/first_form

В верхней части открывающегося интерфейса поисковой программы содержатся три основные опции: *Corpus*, *Query Type*, *Query*.

1. Подкорпуса

В опции *Corpus* можно выбрать один из шести подкорпусов (точнее, четырёх подкорпусов, с разными типами поиска).



- corbama-brut – подкорпус с неснятой омонимией (на апрель 2018 имевший объём около 4 144 000 слов; на самом деле, когда выбирается эта опция, поиск ведётся по обоим подкорпусам – со снятой и с неснятой омонимией);

- corbama-net-non-tonal – подкорпус со снятой омонимией (на апрель 2018 – около 1 076 000 слов), поиск по которому производится без учёта тонов;

- corbama-net-tonal – подкорпус со снятой омонимией (идентичный предыдущему), поиск по которому производится с учётом тонов;

Подкорпус corbama-brut даёт существенно больше употреблений каждого слова, но, в то же время, и много шума: при автоматическом парсинге текста на бамана с опорой на морфологию, порядка 60% всех словоформ текста имеют более одного варианта анализа (впрочем, а февраля 2018 г. проводится частичное автоматическое снятие омонимии при помощи программы, созданной Жан Жаком Мериком, что позволяет снизить уровень омонимии примерно до 30%). По сути дела, поиск по этому подкорпусу эквивалентен поиску в обычном текстовом редакторе (например, Word), отличаясь лишь большей скоростью и удобством представления найденных примеров (в виде конкорданса), а также возможностью сохранения результата поиска в различных форматах.

Поиск по подкорпусу *corbama-net* позволяет исключить шум и задействовать более тонкие параметры.¹ Тонированный поиск (т.е. поиск по *corbama-net-tonal*) даёт возможность ещё больше уменьшить информационный шум, исключая квазиомонимы, отличающиеся от искомой формы тонами.

~~Поиск по *corbama-net* удобен в том случае, если пользователь не знает точно, какой гластный должен быть в нужном слове: например, при наборе формы *te* будут найдены как слова, имеющие форму *te*, так и имеющие форму *te*.~~

Правила тональной нотации в Корпусе изложены в Приложении 2.

- *corbama-ud* – синтаксически аннотированный подкорпус. Синтаксические связи в предложении представлены в формате Universal Dependencies;
- *corbamafara* и *corfarabama* – это один и тот же параллельный бамана-французский корпус. В *corbamafara* основной поиск осуществляется по баманской части, а в *corfarabama* – по французской (при этом в обоих случаях есть возможность вести поиск одновременно и по второй части). Параллельный корпус был опубликован в апреле 2018 г., его объём (около 100 000 слов на момент первой публикации) существенно уступает подкорпусам со снятой омонимией и с неснятой омонимией.

2. Аннотация

Все тексты разбиты на токены. Токен – это словоформа или знак препинания. Каждой словоформе и каждой морфеме в составе словоформы приписана лингвистическая аннотация.

Все корпуса содержат следующие виды лингвистической аннотации словоформ:

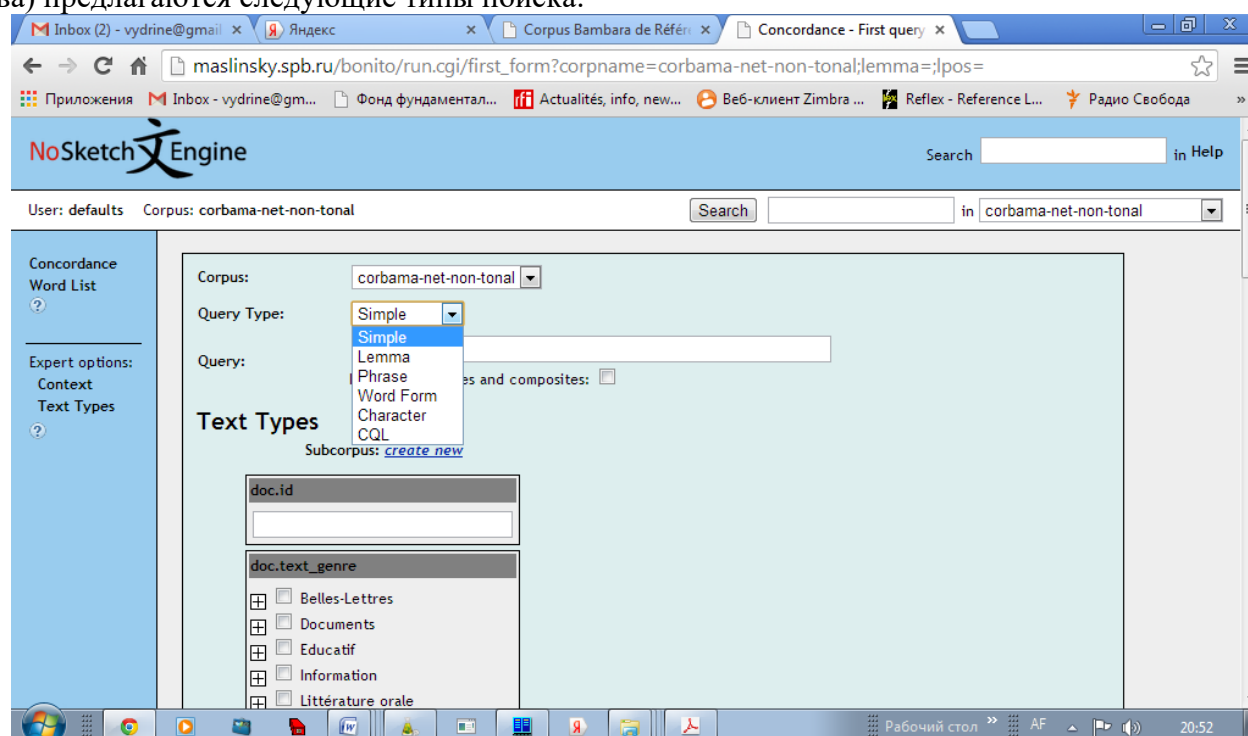
1. **оригинал (word)** — словоформа в том виде, в котором она присутствует в исходном тексте: орфография 1967 года; орфография 1982 года, и др. Если слово в тексте содержит орфографическую ошибку или выступает в своей нестандартной (например, диалектной) форме, эта особенность её написания сохраняется.
2. **лемма (lemma)** (или список лемм) — словарная форма для данной словоформы. В качестве леммы выступают формы без словоизменительных показателей, приведённые к стандартному написанию (орфографические ошибки исправлены, диалектные и прочие «нестандартные» формы заменены на стандартные). Леммами считаются также дериваты, образованные от производящих основ с лексикализацией, и композиты, образование которых сопровождается лексикализацией, т.е. те формы, которые присутствуют в опорной лексической базе данных *Bamadaba* в качестве словарных единиц. В тонированном подкорпусе *corbama-net-tonal* лемма тонирована (т.е. при необозначении тона формы эта при поиске по лемме эта форма найдена не будет). В нетонированных подкорпусах (*corbama-brut*, *corbama-net-non-tonal*) лемма нетонирована (соответственно, при поиске по лемме тон указывать не нужно, иначе формы найдены не будут). При наличии нескольких фонетических вариантов лексем (если это отражено в лексической базе *Bamadaba*) при поиске по одному из них программа находит и те примеры, где эта лексема встретилась в своём втором (а также третьем, четвёртом и т.д.) варианте.
3. **составляющая (part)** — морфема в составе неодноморфемной словоформы.
4. **частеречный тэг (tag)** (см. приложение 3 со списком тегов). В случае неоднозначности возможные частеречные теги записываются через |.
5. **глосса (gloss)** — нормализованный перевод на французский. При создании лексической базы данных *Bamadaba* за основу был взят бамана-французский словарь Шарля Байоля, однако была проведена большая работа по его адаптации с учётом потребности корпусной лексической базы. В частности, каждой лексеме была

¹Подкорпус со снятой омонимией содержит немалое количество ошибок и непоследовательностей, выявлением и устранением которых рабочая группа занимается. Мы будем благодарны пользователям за сообщения о таких ошибках (можно писать В.Ф.Выдрину, vydrine@gmail.com).

приписана французская глосса. Если лексема полисемична, для глоссы выбиралось её наиболее прототипическое значение (разумеется, это было не всегда просто, и какие-то решения могут быть в дальнейшем признаны неудовлетворительными и изменены). Иногда глосса представлена двумя или более французскими словами, разделёнными точками (без пробелов), например: *nèrè* ‘*brisure.de.céréales*’, *ntòmo* ‘*fétiche.des.garçons*’. Для названий биологических видов (особенно – для тех, которые не имеют общепринятых французских названий) в состав глоссы включается латинское название, которому предшествует слово, обозначающее родовую принадлежность. Например: *jénu* ‘*arbre.Hannoa.undulata*’, *ntómi* ‘*serpent.Eryx.muelleri*’.

3. Типы поиска

В опции *Query type* (может быть включена и отключена кликанием на *Query type* в меню слева) предлагаются следующие типы поиска:



3.1. Simple – по словоформе, без учёта регистра. Такой поиск осуществляется по исходной форме текста (строка Word), и одновременно по строке Lemma.

При поиске типа *Simple* есть дополнительная функция – *Include derivates and composites*. Если она не включена, то будут найдены только употребления искомого корня в качестве самостоятельной лексемы (т.е. при поиске по форме *se* будет найдена перфективная форма *sera*, поскольку суффикс *-ra* является словоизменительным, но не *lase*, поскольку *la-* является префиксом деривативным, и не композит *seko*); если включена, то будут найдены все употребления этого корня, в т.ч. в составе дериват и композитов.

3.2. Lemma – по корню (в т.ч. в составе дериват и композитов), точнее по лемме, т.е. исходной форме слова, с учётом регистра. При поиске по лемме, в отличие от поиска *Simple*, не будут отбираться словоформы, содержащие флексии. Так, при поиске *Simple* на стимул *sara* даются и все употребления глагола *sà* ‘умирать’ в перфективе (с суффиксом *-ra*), а при поиске по лемме в результаты попадут только формы слов *sara* (*sàra* ‘paye’, *sàra* ‘payer’, *sàra* ‘avertir’, *sàra* ‘petit tas’, *sàra* ‘charme’), без омонимичной перфективной формы глагола *sà* ‘mourir’. В отношении флексивных форм этот тип поиска имеет смысл лишь для подкорпуса со снятой омонимией; при неснятой омонимии его результаты не отличаются от результатов поиска *Simple*.

Ещё одна особенность поиска по лемме: он позволяет находить и те формы, которые в исходном тексте фигурируют в неправильном виде (с орфографическими ошибками или в нестандартном варианте). Так, если задать поиск по лемме *kunko* ('affaire'), то должны быть найдены и те случаи, когда в тексте эта лексема встретилась в своей диалектной форме *kungo* (это также действительно лишь для подкорпуса со снятой омонимией).

3.3. Phrase – поиск в исходном тексте (уровень аннотации – Word) по последовательности словоформ, разделённых пробелами (в принципе, здесь можно производить и поиск по одной словоформе), с учётом регистра. Этот тип поиска имеет смысл по обоим подкорпусам. **Внимание!** Поскольку поиск типа Phrase производится в исходном (ненормализованном) тексте, он оказывается чувствителен к орфографии источника. Так, при поиске, например, слова *sɔgɔ* будут найдены только написания в новой орфографии, тогда так все случаи употребления этого слова в старой орфографии (*sògò*) будут проигнорированы. Также не будут найдены все случаи, когда в исходном тексте слово записано с ошибкой (даже в том случае, если эта ошибка была исправлена при ручном снятии омонимии).

3.4. Word – поиск в исходном тексте (уровень аннотации – Word) по точной словоформе, без учёта регистра. В отличие от поиска *Simple*, при этом не будут найдены те примеры, где корень, представленный данной последовательностью символов, имеет какие-то аффиксы или входит в состав композитов (при поиске по *mɔgɔ* не будут найдены формы *mɔgɔw*, *dugukɛnɔmɔgɔ*, и т.д.). В то же время, будут найдены словоформы сложной морфологической структуры (так, при поиске на *sara* будет учтена и форма перфектива глагола *sà*). Иначе говоря, этот тип поиска аналогичен поиску в текстовом редакторе Word с включённой опцией «только целые слова», а также поиску закавыченного слова при интернет-поиске. Как и поиск типа Phrase, поиск по Word чувствителен к орфографии исходного текста.

3.5. Character – поиск в исходном тексте (уровень аннотации – Word) по последовательности символов (не разделённой пробелами), не обязательно совпадающий с имеющейся в бамана морфемой (корневой или служебной), с учётом регистра. В принципе, этот тип поиска дублирует поиск по Phrase, отличаясь от него только тем, что не допускает поиск последовательностей, разделённых пробелами.

3.6. CQL – поиск по разным параметрам, а также по комбинациям этих параметров (с учётом регистра). При выборе поиска *CQL* автоматически появляется окно *Default attribute* с опциями *Word*, *Lemma*, *Tag*, *Form*, *Gloss*, *Parts*.

Первые две опции дублируют вышеописанные типы поиска (окно Query Type), но они необходимы для комбинированного поиска, о котором речь пойдёт ниже.

Последние две опции позволяют производить поиск соответственно по частеречной помете и по французской глоссе. Исчерпывающие списки частеречных помет и служебных глосс в Корпусе даны в разделе «Документация» на стартовой странице Корпуса. Заметим, что служебные глоссы в Корпусе помещены вместе с частеречными пометами, т.е. для поиска, скажем, показателя PTCP.RES в качестве опции должен быть указан tag.

Соотношение между типами поиска и уровнями аннотации показано в следующей таблице:

Тип поиска	Задействованный уровень аннотации
Simple	Word, Lemma, Parts
Lemma	Lemma
Phrase	Word
Word form	Word
Character	Word
CQL	Все

4. Ввод искомой формы

4.1. Общие положения

При всех типах поиска, кроме *CQL*, в окно *Query* вводится искомая форма, после чего нужно кликнуть на кнопке *Make Concordance* (внизу экрана) или попросту нажать Enter, после чего программа создаёт конкорданс.

4.2. Обозначение / необозначение тонов во вводимой форме

При поиске по *corbama-brut* и *corbama-net-non-tonal* искомые формы не должны содержать обозначений тонов. При поиске по *corbama-net-tonal* искомая форма должна быть тонированной.²

4.3. Поиск типа *CQL*

При поиске типа *CQL*, в отличие от описанных выше типов, искомая форма заключается в двойные верхние кавычки: “kuma”, “dòn”, “pp”, “serpent”, и т.д.

В новых версиях NoSketchEngine (начиная с 2017 г.) появилась опция *CQLbuilder*, которая существенно облегчает формулировку поисковых задач.

4.3.1. Комбинированный поиск осуществляется одновременно по разным атрибутам лексемы, что позволяет свести до минимума «шум» и получить более прицельную выборку. При таком поиске неважно, какая опция выбрана в окне *Default attribute* (поскольку эти же опции задаются в окне *CQL* «вручную»). Команда, вводимая в окне *CQL*, имеет следующий синтаксис (при этом содержимое одних квадратных скобок соответствует одному токenu):

[опция1="n1" пробел & пробел опция1="n2"]

(n1, n2 – искомые последовательности знаков).

Например, если мы хотим найти все употребления слова *kuma* с частеречной пометой «глагол» (v), запрос выглядит следующим образом:

[word="kuma" & tag="v"]

Возможен и поиск сразу по трём параметрам (или даже четырём, что вряд ли может пригодиться в реальности), например:

[word="kɔ́nɔ" & tag="n" & gloss="oiseau"]

Очевидным образом, комбинированный поиск целесообразен только по подкорпусу со снятой омонимией.

4.3.2. Комбинированный поиск возможен в *CQL* и для многословных выражений. При этом каждое слово (точнее, токен) должен помещаться в квадратные скобки, а между токенами должен быть пробел. Например,

[word="bara" & gloss="calebasse"] [word="kɔ́nɔ" & gloss="à.l'intérieur"]

позволяет найти все сочетания *bàra kɔ́nɔ*, где первое слово – ‘калебаса’ (а не ‘chez’, ‘dancing’, ‘préférè’), а второе – инэссивный послелог (а не ‘attendre’, ‘bouton.de.fleur’, ‘oisezu’, ‘ventre’).

В режиме *CQL* возможен поиск по грамматическому шаблону, который может быть полезен для синтаксических исследований. Например, поиск:

[tag="n"] [tag="adv.p"] [tag="v"]

должен выявить случаи употребления предглагольных наречий с переходными глаголами.³

4.3.3. Режим *CQL* позволяет осуществлять поиск по частям сложной словоформы, в том числе по словоизменительным и деривационным морфемам. Для этого в квадратных скобках должна быть указана соответствующая морфема (напомним, что все служебные морфемы

² Тон обозначается только на первой гласной словоформы, за исключением слов «нестандартных тональных классов». Принципы тональной орфографии, применяемой в Корпусе, изложены в соответствующем документе (стартовая страница, раздел «Документация»).

³ А если такие случаи не находятся, это свидетельствует или о редкости таких наречий в текстах, или (более вероятно) об ошибках операторов снятия омонимии.

помещаются в опции tag!), и поисковик найдёт все словоформы, в составе которых эта морфема содержится.

Например, если нужно найти все формы, содержащие показатель множественного числа, команда формулируется так:

```
[tag="PL"]
```

Если нужно найти последовательность из двух словоформ, в которой первая содержит показатель множественного числа, а вторая – показатель потенциального причастия, команда должна выглядеть так:

```
[tag="PL"] [tag="PTCP.POT"]
```

4.3.4. Режим *CQL* позволяет осуществлять поиск редупликатов (отсутствующих в словаре Bamadaba). Если пользователю нужны все редулицированные глаголы, вводится следующая команда:

```
1:[tag="v"] 2:[tag="v"] & 1.word = 2.word
```

Если он хочет найти все редулицированные слова в корпусе, команда должна иметь следующий вид:

```
1:[] 2:[] & 1.word = 2.word
```

Уточним, что эти команды позволяют найти редулицированные слова, написанные отдельно. Если же нам нужны редулицированные формы, написанные слитно, команда должна выглядеть так:

```
"(.+)\1"
```

Для поиска форм, написанных через дефис, даём такую команду:

```
"(.+)\- \1"
```

Если мы хотим получить сразу и слитные, и дефисные написания, запрашиваем так:

```
"(.+)\-?\1"
```

Чтобы уменьшить шум, можно исключить из поиска ненужные символы (цифры, знак %, и т.п.); они перечисляются без пробелов, помещаясь в квадратные скобки через знак +, при этом им предшествует знак ^. Таким образом, команда «найти все редулицированные формы, написанные слитно или через дефис, исключив из поиска цифры и знак %», выглядит так:

```
"([^\0-9%]+\-?\1"
```

4.4. Ввод нестандартных символов

Ввод нестандартных символов (ɔ, ε, η, ɲ, тональных диакритик) возможен двумя способами:

– при помощи любых клавиатурных раскладок, предназначенных для такого ввода (при этом может быть использована, например, и обычная французская клавиатура – для á, è, é, ÿ... – другое дело, что далеко не всё необходимое можно с её помощью набрать);

– эти нестандартные символы можно заменять следующими комбинациями:

```
;o = ɔ
```

```
;e = ε
```

```
;n = η
```

```
;m = ɲ
```

Знак высокого тона (акут) при этом заменяется запятой, стоящей после соответствующей гласной; знак низкого тона – развёрнутым апострофом. Программа автоматически преобразует эти сочетания в нужные символы, например:

```
k;o, → kó
```

```
su' → sù
```

```
k;e;n;e → kéne
```

```
;m;o' → mò
```

```
;n;o'mi → ñòmi.
```

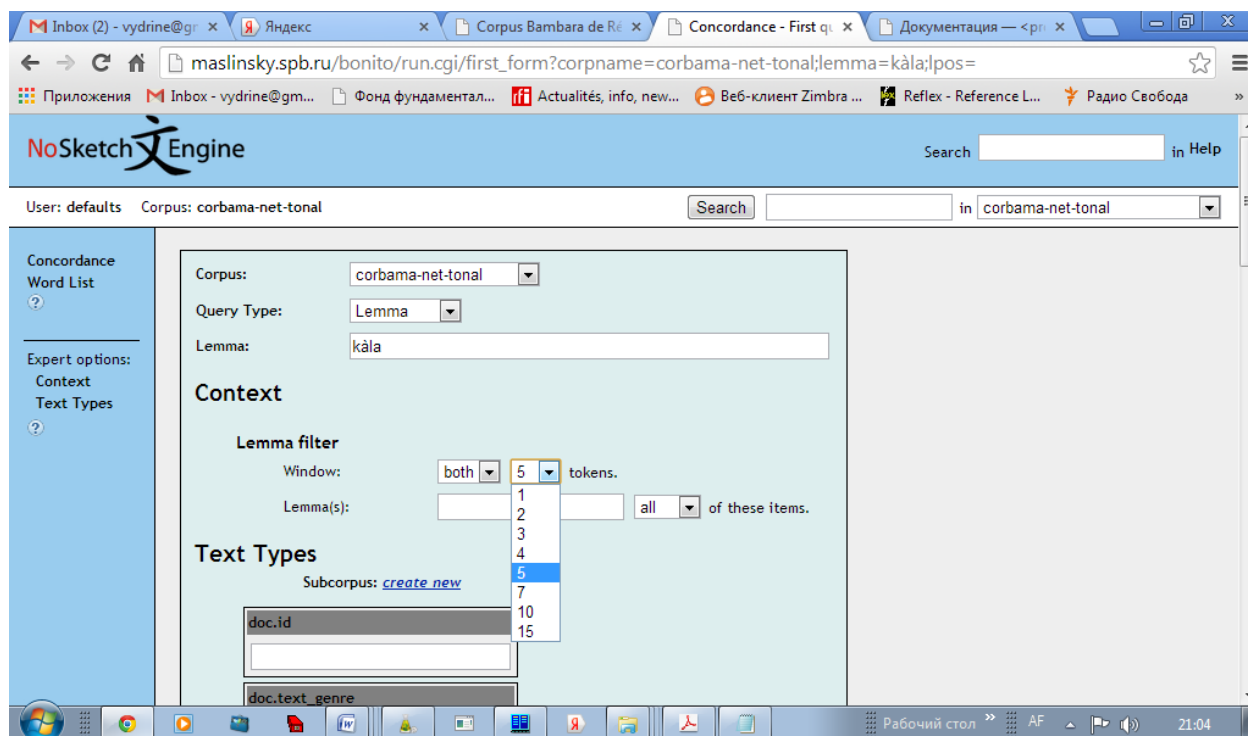

5. Опция Context

Эта опция позволяет осуществлять сочетания дистантно расположенных форм. Она может быть включена и отключена кликанием на *Context* в меню слева

В окне *Query* указывается опорная форма (та, по отношению к которой задаётся контекст).

В *Lemma filter* – *Lemma* указывается интересующая пользователя контекстная форма (т.е. та форма, сочетания с которой с опорной формой требуется найти; тут может быть и более одной формы).

В *Lemma filter* – *Windows* можно указать, какой контекст нас интересует (*left*, *right*, *both* – в последнем случае учитывается и правый, и левый контекст), справа предлагается указать протяжённость контекста, который принимается во внимание (от 1 до 15 форм).



Если задаётся протяжённость 1, то поисковик найдёт только формы, непосредственно прилегающие к опорной (т.е. результат будет аналогичен тому, который мы получим при поиске типа *Phrase*). При протяжённости контекста равной 2, будут найдены формы, как непосредственно примыкающие к опорной, так и те, которые отделены от неё какой-либо одной формой, и т.д. (при этом контекстная форма может быть отделена от опорной и границей предложения).

Справа от окна *Lemma* расположено окно с опциями *All*, *Any*, *None*.

Если выбрана опция *All*, при этом в окне *Lemma* внесены две (или более) контекстные формы, то поисковик найдёт только те примеры, где присутствуют все три формы (опорная и обе контекстные). Например, при опорной форме *ke* и двух контекстных – *yeŕe*, *jaŕɔŕn*, будут найдены такие примеры:

Мəɔ min be a məɔjaŕɔŕn jogin , a ye min ke o tigi la , o **jaŕɔŕn** ka **ke** a **yeŕe** fana la .

O de be cikela ke senyerekorəbaga ye , i n' a fə birokənoəbaarakela ; i n' a fə tapini julabaw , i n' a fə **yeŕe** jamanakuntigi n' a **kokejaŕɔŕnw** , senyerekorə siratege la

Jatigike yeŕe puman na , a **jaŕɔŕn** se kise t' ale denw na .

и т.д.

Такой поиск может быть весьма эффективен для проверки возможности употребления переходных глаголов с различными предикативными показателями (скажем, при изучении акциональных классов), при изучении сочетаний глаголов с послелогоми, и т.п.

При включённой опции *Any*, будут найдены все случаи совместного употребления морфемы *ke* с хотя бы одним из двух контекстных форм (в том числе, разумеется, и те случаи, когда присутствуют обе контекстные формы).

При включённой опции *None* программа выдаст все случаи употребления опорного слова, когда на заданной дистанции ОТСУТСТВУЮТ контекстные формы. Эта опция может быть полезной, когда некая форма обычно употребляется в составе каких-то устойчивых выражений, а пользователя интересуют её употребления вне таких выражений.

6. Опция Text types

Этот раздел позволяет ограничить набор текстов, по которым производится поиск. Опция может быть включена и отключена кликанием на *Text types* в меню слева.

По умолчанию, поисковая программа ищет по всему подкорпусу. В первом окне, *doc.id*, можно задать искомый текст; для этого нужно начать набирать фамилию его автора или первое слово произведения. Если в названиях файлов эти элементы присутствуют, то эти названия будут подсказаны во всплывающей подсказке.

Ниже расположены окна:

- *doc.text_genre*, в котором можно задать ограничения по жанровым характеристикам текстов;
- *doc.source_type*, где можно ограничить выборку по типам источника (периодика, средства аудио- и видеокommunikации, рукописные источники...);
- *doc.source_year*, где можно ограничить поиск по датам создания документов, и некоторые другие (количество признаков, по которым можно задавать ограничения, будет увеличиваться по мере пополнения корпуса и улучшения метатекстовой разметки).

7. Concordance

7.1. Что такое конкорданс корпуса

Неотрицательным результатом поиска по корпусу является конкорданс, т.е. список всех примеров (с их контекстами), найденных в корпусе (или подкорпусе). Справочный корпус бамана не имеет ограничений по количеству предоставляемых пользователю примеров. В верхней белой полосе указано количество найденных примеров (Hits). Под этой полосой указывается количество страниц (если число найденных примеров более 20; по умолчанию, на одной странице даётся по 20 примеров), здесь же расположены кнопки навигации по конкордансу.

Для каждого примера указывается название файла (в названии файла отражены, достаточно прозрачно, имя автора и название текста; о правилах именований файлов см. соответствующий документ в разделе «Документация» на стартовой странице Корпуса).

7.2. Форма представления конкорданса

Для **настройки формы представления конкорданса** в меню имеются две опции — *KWIC/Sentence* и *View Options*.

Кликанием на *KWIC/Sentence* производится простое переключение режима просмотра примеров: *Sentence* — показывается целое предложение («от точки до точки»), содержащее искомую форму; *KWIC* — показывается правый и левый контекст заданного размера (по умолчанию, 40 знаков слева и 40 знаков справа от искомой формы)

Опция *View Options* позволяет регулировать представление конкорданса более детально. В интерфейсе *View Options* можно:

- изменить атрибуты формы (*Attributes*) — если пометить опции *word*, *lemma*, *tag*, *gloss*, то соответствующие атрибуты (лемма, частеречная помета, французская глосса; по

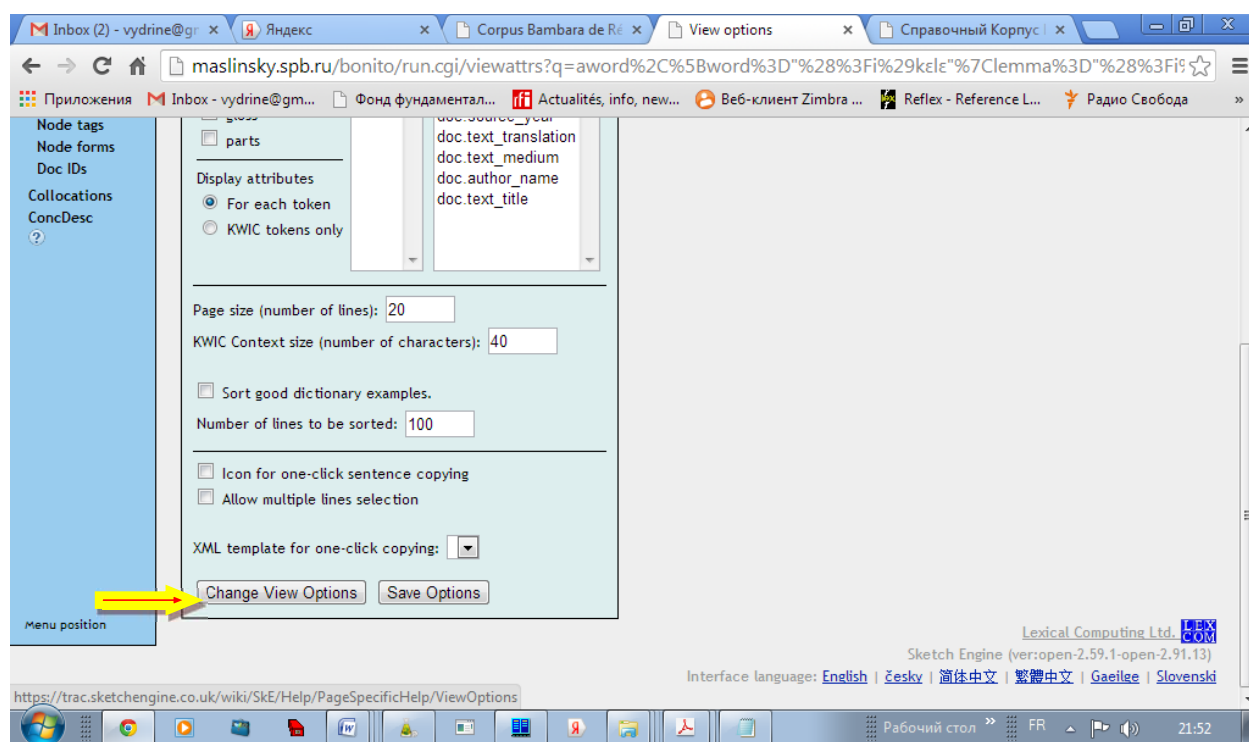
умолчанию, опция *word* помечена всегда) при форме будут указаны. Атрибут *form* представляет собой «нормализованную словоформу» (в новой орфографии, тонированная, с поморфемной разбивкой дефисами). Атрибут *part* даёт доступ к полю, содержащему все знаменательные основы, которые входят в состав словоформы (что актуально для композитов и дериватов);

– указать, должны ли эти атрибуты указываться при каждом слове каждого примера, или только при искомом слове (раздел *Display Attributes*).

Указание атрибутов при каждом слове оказывается несколько громоздким для подкорпуса с неснятой омонимией (corbama.brut), поскольку большинство словоформ имеют более одного варианта разбора. По-видимому, использование этой опции следует признать целесообразной только для подкорпуса со снятой омонимией.

Ниже можно установить, сколько примеров должно быть показано на одной странице (*Page size*; по умолчанию, выставляется цифра 20); каков размер правого и левого контекста (*KWIC Context size*; в принципе, он может быть увеличен до бесконечности, но по умолчанию выставляется цифра 40).

Чтобы активизировать заданную в разделе *View Options* конфигурацию, нужно нажать на кнопку *Change view options* :



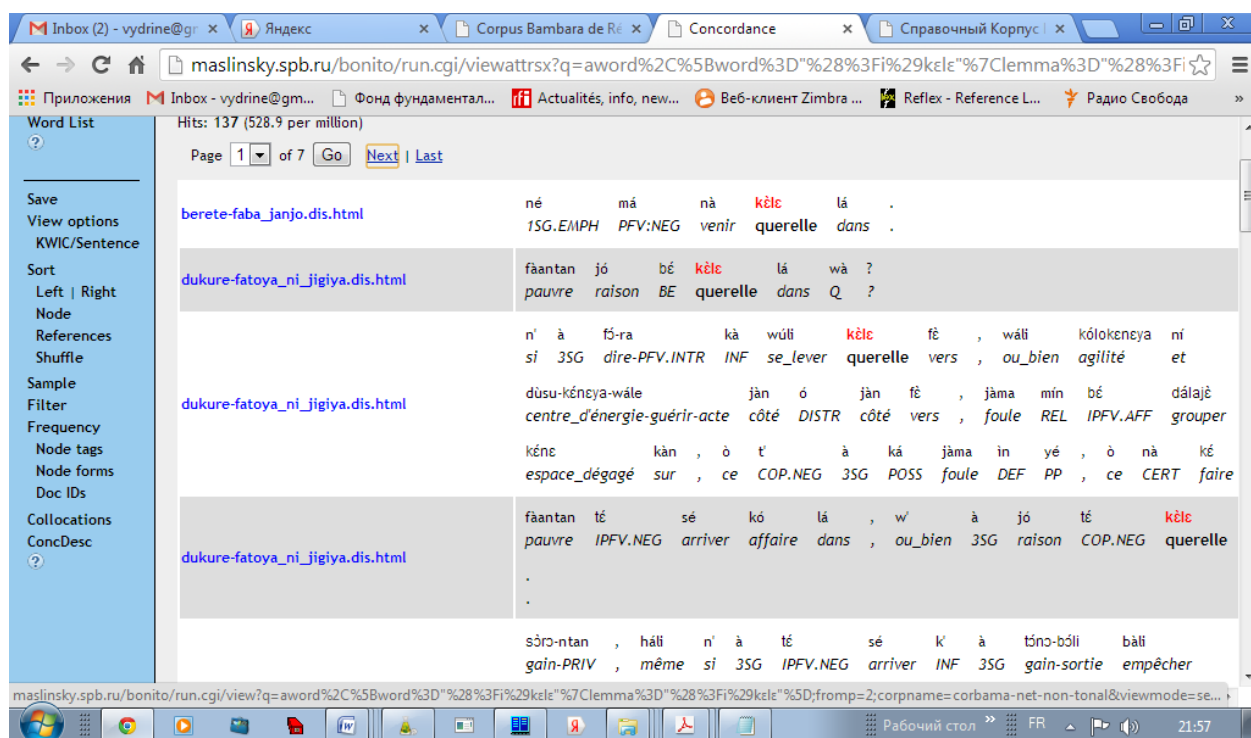
Остальные функции раздела *View Options* (*Sort good dictionary examples* и т.д.) для нашего корпуса нерелвантны.

7.3. Сохранение и экспорт конкорданса

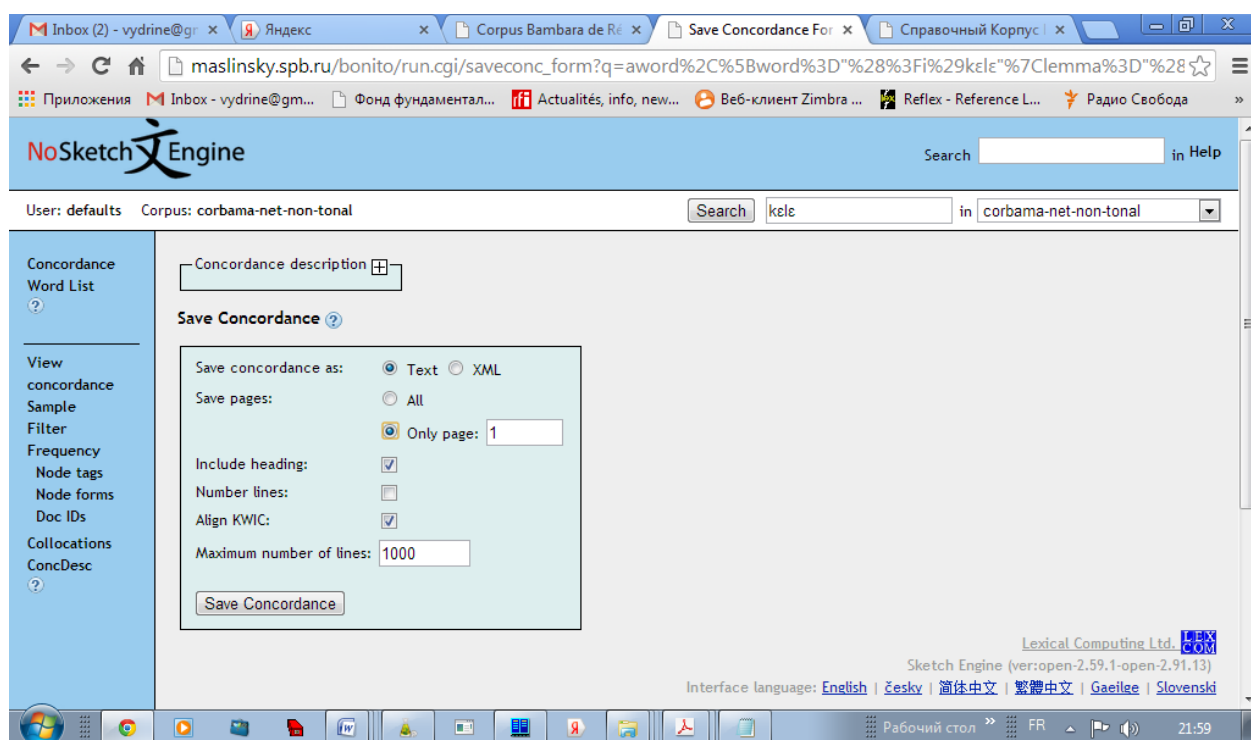
Полученный конкорданс может быть **сохранён** по частями или целиком в текстовом формате. Для сохранения целиком нужно выбрать опцию *Save* в меню слева.

Если пользователь хочет экспортировать глоссированный нормализованный текст (т.е. такой текст, который, после минимальной редактуры, можно использовать, скажем, для примеров в научной статье), то рекомендуется сделать следующее:

- войти в опцию *View options*, выбрать атрибуты *form* и *gloss*, в разделе *Display attributes* : *For each token*, после чего нажать на кнопку *Change view options*. Конкорданс приобретёт примерно такой вид:



Затем нужно выбрать в меню слева опцию *Save* и задать нужные параметры (все страницы или только одна; формат создаваемого файла, и т.д.):



Если выбрана опция *Align KWIC*, то в создаваемом файле все слова и их глоссы будут выровнены пробелами. При отключении этой опции словоформы (а также их глоссы) будут разделены одним табулятором.

7.4. Сортировка примеров

Сортировка примеров регулируется в опции *Sorting*. Примеры могут следовать а алфавитном порядке формы, располагающейся справа от искомого (*Right context*) или слева

от искомого (*Left context*), при этом различие прописных и строчных букв может учитываться или игнорироваться (*Ignore case*). Возможно также выстраивание в обратном алфавитном порядке (*Backward*). Приведение в действие выбранных параметров производится нажатием кнопки *Sort Concordance*.

Многоуровневая сортировка для нашего корпуса пока что неактуальна.

В общем меню содержатся также опции *Sorting – References* (сортировка по именам файлов, в которых найдены искомые формы) и *Sorting – Shuffle* (перемешивание примеров, в результате которого они располагаются в случайном порядке).

7.5. Опция *Sample*

Эта опция позволяет делать случайную выборку заданного размера из числа всех найденных в корпусе примеров.

7.6. Опция *Filter*

Эта опция аналогична по своим функциям опции *Context*, описанной выше (раздел 4).

7.7. Опция *Frequency*

Эта опция даёт доступ к статистике словоформ, в состав которых входит искомый элемент, а также статистики его сочетаемости с соседними формами.

В интерфэйсе этой опции есть два раздела.

6.7.1. *Multilevel frequency distribution*. Для каждого уровня иерархии сортировки по частоте нужно выбрать

- *Node*, и тогда подсчитывается частота словоформ, в которые входит искомый элемент (при этом можно выбрать опцию *Ignore case*, тогда не будет учитываться различие между прописными и строчными буквами),

- элементы из левого контекста (1L, 2L, 3L... – в зависимости от протяжённости контекста) или правого контекста (1R, 2R, 3R...). Тогда будет подсчитана частота сочетаемости с формами слева или справа.

При этом могут задаваться атрибуты опорного или контекстного элемента: *word*, *lemma*, *tag*, *gloss*. Отметим, что при подсчёте частот по корпусу с неснятой омонимией (*Corbama-brut*) выяснение частот опорного элемента по параметрам *lemma*, *tag*, *gloss* нерелевантно.

6.7.2. Раздел *Text Type frequency distribution* позволяет определить частоту встречаемости искомого элемента в:

- разных файлах, опция *doc.id*;
- разных текстах (отметим, что один текст может быть представлен в Корпусе несколькими файлами), опция *doc.text_title*;
- текстах разного жанра, опция *doc.text_genre*.

7.8. Раздел *Collocations*

Опции этого раздела позволяют определить возможные устойчивые сочетания с искомым словом. Возможен поиск сочетаемости по атрибутам (*Attribute*) соседних элементов (*word*, *lemma*, *tag*, *gloss*), при этом можно учитывать только элементы левого контекста (*In the range from -1, -2, etc.*) или правого контекста (*... to 1, 2, etc.*); цифровые значения соответствуют протяжённости контекста (-1/1: учитывается только непосредственно примыкающий сосед; -2/2: учитываются ближайший сосед и непосредственно предшествующее/непосредственно следующее за ним слово, и т.д.).

Нажав кнопку *Make Candidate List*, получаем список кандидатов на устойчивые сочетания. Их можно выстроить в порядке убывания частоты, кликнув по голубой этикетке *Frec*.

8. Word List

Опция *Word List* позволяет создавать частотный словарь. Если войти в эту опцию, то в меню слева появляются этикетки *All words*, *All lemmas*. Нажав на эти кнопки, мы получаем частотный список (в порядке убывания) всех токенов по данному подкорпусу (поскольку знаки препинания тоже имеют статус токенов, они фигурируют в этом списке наряду со словами).

9. Поиск в параллельном бамана-французском корпусе

Поиск по баманской форме осуществляется так же, как и обычный поиск в подкорпусе со снятой омонимией. Различие проявляется в представлении конкорданса, где рядом с баманской фразой показывается соответствующая ей французская:

Home	Query mogo 1,362 > Filter by aligned corpus 1,346 (10,296.03 per million)	
Search	Page 1 of 68 Go Next Last	
Word list		
Corpus info		
My jobs		
User guide		
Save		
Make subcorpus		
View options		
KWIC		
Sentence		
Alignment		
Sort		
Left		
Right		
Node		
References		
Shuffle		
Sample		
Filter		
Sub-hits		
1st hit in doc		
Frequency		
Node tags		
Node forms		
Doc IDs		
Collocations		
Visualize		
0		
Jump position		

#232,baileul-sagesse_bambara_01.dis.html	17) Ala be mogo minnu ke den ni nafolocaman tigi ye , olu ma foyi di	#289,baileul-sagesse_bambara_01.dis.fra.vert	corfarabama
#380,baileul-sagesse_bambara_01.dis.html	26) mogo be se K' a fo mogo ma : = I sen bo n ka ko la = , nka i te se K'	#486,baileul-sagesse_bambara_01.dis.fra.vert	26) On peut dire à quelqu'un de ne pas se mêler de nos affaires , mais on ne peut le dire à Dieu-Créateur .
#386,baileul-sagesse_bambara_01.dis.html	a fo Sabu ma , a K' a sen bo i ka ko la .	#486,baileul-sagesse_bambara_01.dis.fra.vert	26) On peut dire à quelqu'un de ne pas se mêler de nos affaires , mais on ne peut le dire à Dieu-Créateur .
#421,baileul-sagesse_bambara_01.dis.html	26) mogo be se K' a fo mogo ma : = I sen bo n ka ko la = , nka i te se K'	#512,baileul-sagesse_bambara_01.dis.fra.vert	27) On a beau avoir peur , on n' échappe pas à Dieu .
#447,baileul-sagesse_bambara_01.dis.html	a fo Sabu ma , a K' a sen bo i ka ko la .	#556,baileul-sagesse_bambara_01.dis.fra.vert	30) Si une seule personne se met à appeler Dieu : " Mon beau-père " , il montrera aux autres leur père .
#470,baileul-sagesse_bambara_01.dis.html	27) mogo te siran ka Ala dan .	#580,baileul-sagesse_bambara_01.dis.fra.vert	31) L' homme n' a pas de pouvoir , sinon Dieu .
#524,baileul-sagesse_bambara_01.dis.html	30) ni mogo kelen y' a fo : = n buran Ala = i na tow fa jira u la .	#654,baileul-sagesse_bambara_01.dis.fra.vert	35) Ce que Dieu propose à l' homme , qu' il l' accueille , le bonheur comme l' épreuve .
#550,baileul-sagesse_bambara_01.dis.html	31) se te mogo bolo , a be Ala bolo .	#675,baileul-sagesse_bambara_01.dis.fra.vert	36) L' homme ne peut empêcher ce que Dieu a décidé .
#624,baileul-sagesse_bambara_01.dis.html	35) Ala ye ko min da mogo ye , a lamine i la , kojugu fara kojuman	#774,baileul-sagesse_bambara_01.dis.fra.vert	42) Une tête créée par Dieu , l' homme ne peut la briser .
#674,baileul-sagesse_bambara_01.dis.html	kan , a lamine i la .	#831,baileul-sagesse_bambara_01.dis.fra.vert	47) Personne ne peut échapper à son destin .
#790,baileul-sagesse_bambara_01.dis.html	36) Ala ye ko min latige , mogo te se K' o tije .	#999,baileul-sagesse_bambara_01.dis.fra.vert	58) L' origine d' une personne est plus ancienne qu' elle-même .
#801,baileul-sagesse_bambara_01.dis.html	42) Ala ka kunkolo dalen , mogo te se ka o ci .	#1012,baileul-sagesse_bambara_01.dis.fra.vert	59) On ne meurt jamais le jour désiré par un ennemi .
#895,baileul-sagesse_bambara_01.dis.html	47) mogo si te taa K' i dakan dan .	#1148,baileul-sagesse_bambara_01.dis.fra.vert	65) Tant que la fin de la vie n' est pas arrivée , les épreuves ne peuvent pas conduire à la mort .
#914,baileul-sagesse_bambara_01.dis.html	58) mogo soroko ka koro ni i yere ye .		67) On peut réussir au village et ne pas réussir en brousse ou l' inverse .
#927,baileul-sagesse_bambara_01.dis.html	59) mogo te sa i jugu dongodon .		67) On peut réussir au village et ne pas réussir en brousse ou l' inverse .
	65) nagan te mogo faga ni waati ma se .		
	67) so ka di mogo do la , kungo man di i la ; kungo ka di mogo do la ,		
	so man di i la .		
	67) so ka di mogo do la , kungo man di i la ; kungo ka di mogo do la ,		
	on man di i la .		

Поисковый интерфейс параллельного корпуса имеет опцию "Параллельный поиск", который позволяет указать нужную французскую лексему, интересующую пользователя. Например, при поиске по баманской форме *kɔnɔ*, можно задать параллельный поиск по французской лексеме *oiseau*:

Corpus: **corbamafara**

Simple query: **kɔnɔ** Include derivative and composite forms: ☐ Make Concordance

[Query types](#) [Context](#) [Text types](#)

Parallel query

☒ corfarabama (corfarabama) Contains

Simple query: **oiseau** Include derivative and composite forms: ☐ [Query types](#)

☒ filter out empty lines

Context

Lemma filter

Window: **both** 5 tokens.

Lemma(s): all of these items.

PoS filter

Window: **both** 5 tokens.

☐ adjectif ☐ adverbe ☐ adverbe préverbal ☐ conjonction ☐ copule ☐ déterminatif

☐ all of these items.

Make Concordance Clear All

В результате в конкорданс попадут только те баманские предложения с *kɔnɔ*, чьи французские аналоги содержат *oiseau*.

Query kono, oiseau 27 (206.53 per million)	
Page 1 of 2 Go Next Last	
#71,baileul-sagesse_bambara_01.dis.html	corbamafara 5) ni namakoro tun kera jodunfen ye, mogokoroba tun tēna kono gen sanko denmisen .
#1597,baileul-sagesse_bambara_01.dis.html	115) kono min b'i to san fe ka yegoyegeli kiz, n'a jiginna dugu ma, o na kolon sen .
#5727,baileul-sagesse_bambara_01.dis.html	411) sin tē kono fe, kono b'a den balo .
#5730,baileul-sagesse_bambara_01.dis.html	411) sin tē kono fe, kono b'a den balo .
#8511,baileul-sagesse_bambara_02a.dis.html	590) jiri min tulen be, kono be sigi o la .
#11635,baileul-sagesse_bambara_02a.dis.html	807) mogo ye kono min su tōmo, i tē o kasikan don .
#13778,baileul-sagesse_bambara_02b.dis.html	954) kono bee ni i sigiyoro don .
#13973,baileul-sagesse_bambara_02b.dis.html	969) kono min y'a kili da sugufye la, ka a toro sugufye la ka wuli, ni o taara a sigi jiri la, mogo kelen ka taa i jo o jukoro ko : = haa i = i kumana, nka i ma fen fo .
#21136,baileul-sagesse_bambara_02b.dis.html	1453) kono bee be dufon filz nalonma ye, nka kono bee jurulama be teme dufon na .
#21145,baileul-sagesse_bambara_02b.dis.html	1453) kono bee be dufon filz nalonma ye, nka kono bee jurulama be teme dufon na .
#21155,baileul-sagesse_bambara_02b.dis.html	1454) kono bee tigelen duga la, u bee jurulama be teme a la .
#91,baileul-sagesse_bambara_01.dis.fra.vert	corfarabama 5) Si l'hyène mangeait le mil, (même) les adultes ne chasseraient pas les oiseaux , a fortiori les enfants .
#2055,baileul-sagesse_bambara_01.dis.fra.vert	115) L' oiseau qui fait le geste de gratter en restant en l'air creusera un puits quand il sera descendu à terre .
#7614,baileul-sagesse_bambara_01.dis.fra.vert	411) L' oiseau n'a pas de sein, mais il nourrit tout de même son petit .
#7614,baileul-sagesse_bambara_01.dis.fra.vert	411) L' oiseau n'a pas de sein, mais il nourrit tout de même son petit .
#11337,baileul-sagesse_bambara_02a.dis.fra.vert	590) Les oiseaux se posent dans les arbres bien feuillus .
#15570,baileul-sagesse_bambara_02a.dis.fra.vert	807) On ne connaît pas le chant d'un oiseau mort qu'on a ramassé .
#18502,baileul-sagesse_bambara_02b.dis.fra.vert	954) Chaque oiseau a son emplacement (privilégié) pour se poser .
#18763,baileul-sagesse_bambara_02b.dis.fra.vert	969) L' oiseau qui a poudu au marché, et y a fait éclore ses petits, s'il va se poser sur une branche et que quelqu'un vienne se planter au pied de l'arbre en criant ' haaa' .
#28701,baileul-sagesse_bambara_02b.dis.fra.vert	1453) Tous les oiseaux prennent le calao d'Abyssinie pour un imbécile, mais ils passent tous devant lui, attachés à une corde .
#28701,baileul-sagesse_bambara_02b.dis.fra.vert	1453) Tous les oiseaux prennent le calao d'Abyssinie pour un imbécile, mais ils passent tous devant lui, attachés à une corde .
#28727,baileul-sagesse_bambara_02b.dis.fra.vert	1454) Tous les oiseaux prennent le vautour pour un imbécile, mais ils passent tous devant lui, attachés à une corde .

Все описанные выше типы поиска остаются возможными и для параллельного подкорпуса.

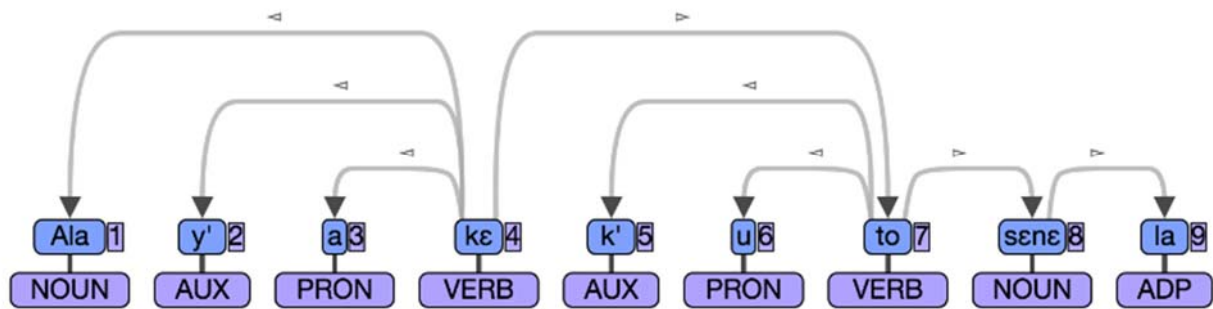
Если выбрать среди подкорпусов corfarabama, основной для поиска становится французская часть параллельного подкорпуса. Эта часть полностью лемматизирована; таким образом, при поиске по форме инфинитива глагола или по форме единственного числа существительного или прилагательного будут найдены все флексивные формы лексемы. Так, поиск по *crier* позволяет найти фразы с формами *crier*, *crie*, *criant*, и т.д.

crier defaults	
Query crier 11 > Filter by aligned corpus 11 (74.72 per million)	
baileul-sagesse_bambara_02b.dis.fra.vert,#18799	corbamafara 969) L'oiseau qui a poudu au marché, et y a fait éclore ses petits, s'il va se poser sur une branche et que quelqu'un vienne se planter au pied de l'arbre en criant ' haaa' .
baileul-sagesse_bambara_02c.dis.fra.vert,#32613	1644) Si on entend l'étranger (imiter le bruit du tam-tam) en criant : ' Pantavu !', il faut que les gens du village lui laissent l'aire de danse .
baileul-sagesse_bambara_02c.dis.fra.vert,#34046	1713) Il aurait mieux valu menacer quelqu'un que de lui crier : ' Relève-toi !' .
baileul-sagesse_bambara_02c.dis.fra.vert,#36944	1853) Quand on crie dans une clairière déserte, et qu'on ne voit personne, on se verra soi-même .
baileul-sagesse_bambara_02c.dis.fra.vert,#47894	2380) Quand tu vas dans un village et que tu trouves les gens à crier comme l'hyène, si tu te mets à bêler comme une chèvre, ils vont te manger .
baileul-sagesse_bambara_02d.dis.fra.vert,#53314	2647) Basiriba crie à tue-tête que sa canne de fer est perdue .
baileul-sagesse_bambara_02e.dis.fra.vert,#73511	3620) (Quelqu'un peut crier) : ' Aldez-moi à maîtriser mon fauve ! ' pourvu qu'il lui tienne la gueule .
baileul-sagesse_bambara_02e.dis.fra.vert,#79144	Il l'a dit, c'est bien, (mais crier) : ' Attrapez-le !' .
kurane002.dis.fra.vert,#101323	Les mécréants ressemblent à [du détail] auquel on crie et qui entend seulement appel et voix confus .
fane-cikela_ni.dis.fra.vert,#135069	et se mit à crier , crier .
fane-cikela_ni.dis.fra.vert,#135071	et se mit à crier, crier .
baileul-sagesse_bambara_02b.dis.html,#13971	corbamafara 969) kono min y'a kili da sugufye la, ka a toro sugufye la ka wuli, ni o taara a sigi jiri la, mogo kelen ka taa i jo o jukoro ko : = haa i = i kumana, nka i ma fen fo .
baileul-sagesse_bambara_02c.dis.html,#24025	1644) ni i ye a men dunanke da ko : = pantavu =, duguleniw ka kenz bila a ye .
baileul-sagesse_bambara_02c.dis.html,#24997	1713) kankarida ka fisa = wuli ka bo n dogo disi la ! = ye .
baileul-sagesse_bambara_02c.dis.html,#27032	1853) n' i kulela fuga la, n' i ma mogo ye, i n' i yere ye .
baileul-sagesse_bambara_02c.dis.html,#34940	2380) e taara dugu min na, n' i y' a soru u be surukulasi ke, n' i ye bakasi ke, u b' i dun .
baileul-sagesse_bambara_02d.dis.html,#38871	2647) Basiriba be woyo ko : = negebere tununen be = .
baileul-sagesse_bambara_02e.dis.html,#53834	= 3620) = a' ye n dmc n ka wara la ! =, a kunkolo ka ye i yere bolo .
baileul-sagesse_bambara_02e.dis.html,#57850	, a b' a da la =, o be ke, nka = a' y' a mine !
kurane002.dis.html,#79578	mogo minnu kafiriyara, olu (welebaa) misali be i n' a fo mogo min be peren (baganw fe) minnu te fosi men fo welekan, ani perenikan, tulogereniw don, bobow don, fiyentow don, u hakili te miiri .
fane-cikela_ni.dis.html,#118412	kabini ne nana ka ji sigi, ko ne be bo yoro min, a sinna ka gorow ke n no, k' a be bin n kan, ee !
fane-cikela_ni.dis.html,#118412	kabini ne nana ka ji sigi, ko ne be bo yoro min, a sinna ka gorow ke n no, k' a be bin n kan, ee !

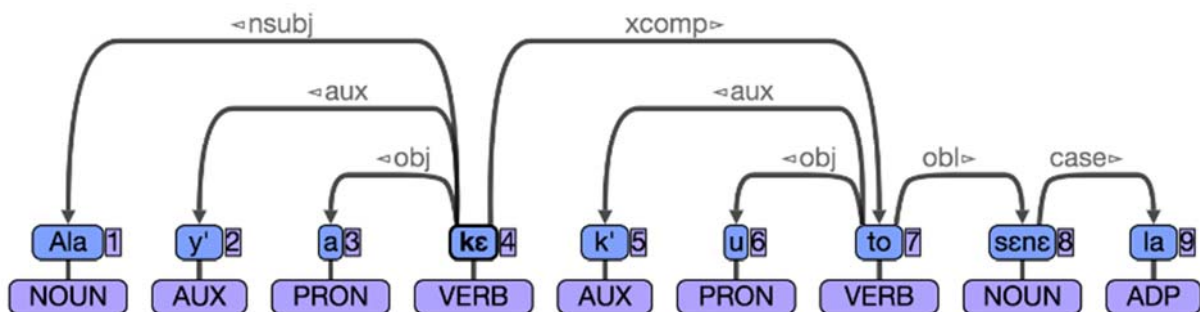
Lexical Computing

10. Поиск в синтаксически аннотированном подкорпусе (corbama-UD)

Синтаксический подкорпус – это трибэнк, размеченный по схеме Universal Dependencies. В текстах из подкорпуса со снятой омонимией были размечены синтаксические зависимости, список которых доступен по ссылке: <http://universaldependencies.org/bm/dep/>. Разметка зависимостей состоит в обозначении связей между парами слов (токенов); в каждой из таких пар одно слово представлено как главное, другое – как зависимое.



Каждое слово в предложении пронумеровано. Таким образом, при создании зависимостей мы получаем информацию о порядковом номере вершины каждого слова. Далее каждой зависимости присваивается ярлык (dependency relation).



В синтаксическом подкорпусе поиск в *simple query* возможен по баманской форме или по типу зависимости. Например, чтобы найти все случаи, когда в клаузе имеется зависимость типа obj (иначе говоря, когда при глаголе есть прямое дополнение), в строке поиска указывается тэг этой зависимости, obj:

The screenshot shows the 'Requête simple' search interface. The 'Corpus' is set to 'corbama-ud'. The 'Requête simple' field contains 'obj'. The 'Type de requêtes' section has 'simple' selected. There are buttons for 'Créer la concordance' and 'Effacer tout'.

В результате такого поиска мы получаем конкорданс, включающий все клаузы, содержащие прямое дополнение:

NoSketch Engine

Page d'accueil
Chercher
Liste de mots
Corpus info
Mes tâches
Guide utilisateur

Enregistrer
Faire des souscorpus
Options de visualisation
KWIC
Phrase
Classement
Contexte gauche
Contexte droit
Node
Références
Mélanger
Échantillon
Filtre
Sub-hits
1ère occurrence dans le doc
Fréquence
Node tags
Node formes
ID de documents
Collocations
Visualiser

Requête **obj** 916 (69,210.43 par million) ⓘ

Page 1 de 46

meyer_gorog-contes_bambara_11namaramatu_denin_karogelen,#21 a bamuso ye **a** fò ko a ka taga a fakè wele ka na .

meyer_gorog-contes_bambara_11namaramatu_denin_karogelen,#28 a bamuso ye a fò ko a ka taga a **fakè** wele ka na .

meyer_gorog-contes_bambara_11namaramatu_denin_karogelen,#49 a tagara ko baba , ko e ka na **dumunni** kè .

meyer_gorog-contes_bambara_11namaramatu_denin_karogelen,#58 ko e na **dumunni** kè !

meyer_gorog-contes_bambara_11namaramatu_denin_karogelen,#66 a tagara **a** fò a bamuso ye ko ne ye baba wele , o ko fònhòn !

meyer_gorog-contes_bambara_11namaramatu_denin_karogelen,#74 a tagara a fò a bamuso ye ko ne ye **baba** wele , o ko fònhòn !

meyer_gorog-contes_bambara_11namaramatu_denin_karogelen,#83 a ye **a** fò ko sisan ni n tagara bilen tugun , ni n ye baba wele , ni a ko fònhòn , ne bè baba bugò !

meyer_gorog-contes_bambara_11namaramatu_denin_karogelen,#96 a ye a fò ko sisan ni n tagara bilen tugun , ni n ye **baba** wele , ni a ko fònhòn , ne bè baba bugò !

meyer_gorog-contes_bambara_11namaramatu_denin_karogelen,#106 a ye a fò ko sisan ni n tagara bilen tugun , ni n ye baba wele , ni a ko fònhòn , ne bè **baba** bugò !

meyer_gorog-contes_bambara_11namaramatu_denin_karogelen,#116 a tagara ko baba , e na **dumunni** kè !

meyer_gorog-contes_bambara_11namaramatu_denin_karogelen,#126 ko e na **dumunni** kè !

meyer_gorog-contes_bambara_11namaramatu_denin_karogelen,#138 a bilen bolila ka na **a** fò a bamuso fè , a ye muruda ge ka a don a yèrè kun .

meyer_gorog-contes_bambara_11namaramatu_denin_karogelen,#146 a bilen bolila ka na a fò a bamuso fè , a ye **muruda** ge ka a don a yèrè kun .

meyer_gorog-contes_bambara_11namaramatu_denin_karogelen,#149 a bilen bolila ka na a fò a bamuso fè , a ye muruda ge ka **a** don a yèrè kun .

meyer_gorog-contes_bambara_11namaramatu_denin_karogelen,#159 ko ni baba ye **a** fò bilen tugun fònhòn , ko ale bè baba kun tigè !

meyer_gorog-contes_bambara_11namaramatu_denin_karogelen,#169 ko ni baba ye a fò bilen tugun fònhòn , ko ale bè baba **kun** tigè !

meyer_gorog-contes_bambara_11namaramatu_denin_karogelen,#177 a ko baba , na **dumunni** kè !

meyer_gorog-contes_bambara_11namaramatu_denin_karogelen,#187 baba , na **dumunni** kè !

meyer_gorog-contes_bambara_11namaramatu_denin_karogelen,#197 a ye baba **kunkoto** tigè .

Все описанные в предыдущих разделах типы поиска остаются возможными и для синтаксического подкорпуса. Кроме того, в *CQL* можно осуществлять поиск по частям речи, по порядковому номеру вершины и по глоссе.

Corpus:

Requête simple:

Include derivative and composite forms: ☐

[Types de requêtes](#) [Contexte](#) [Types de texte](#) ⓘ

Type de requêtes ☐ simple ☐ lemme ☐ phrase ☐ mot ☐ caractère ☒ CQL

Lemme:

Phrase:

Forme du mot:

Caractère:

CQL:

[Résumé de tags CQL builder](#)

NoSketch Engine

Page d'accueil
Chercher
Liste de mots
Corpus info
Mes tâches
Guide utilisateur

Enregistrer
Faire des souscorpus
Options de visualisation
KWIC
Phrase
Classement
Contexte gauche
Contexte droit
Node
Références
Mélanger
Échantillon

Requête **DET, 5, Gloss=REL** 11 (831.13 par million) ⓘ

meyer_gorog-contes_bambara_11namaramatu_denin_karogelen,#873 n ye nstirinin sòrò yòrò **min** , n ye a bila yen .

meyer_gorog-contes_bambara_05sinamuso_jugu,#1247 a ye dònkili la tuma **min** , muso nin ye a bolo la a kun .

meyer_gorog-contes_bambara_12ala_ani_sonsannin,#1857 a ye nugurè sigi tuma **min** , u bè tagalen , a ko : Mama Ngala ko n ka a sanba siginònòkèné dó .

meyer_gorog-contes_bambara_12ala_ani_sonsannin,#2197 n ye a ta yòrò **min** , n ye a bila yen .

meyer_gorog-contes_bambara_08dogonin_ani_koroke,#4293 n ye o ta yòrò **min** , n ye o bila yen .

meyer_gorog-contes_bambara_09cenin_kegunman,#5862 ko : hén , den **min** kumana , a be a ba kòrò , ko o tè se ka i yèrè wolowolo wa ?

meyer_gorog-contes_bambara_01syefan_ani_syedennin,#7386 n ye nstirinin sòrò yòrò **min** , n ye a bila yen .

meyer_gorog-contes_bambara_02mansake_denmuso_furuli,#8779 o ye a ci tuma **min** na , mansakè tagara ka denmuso miné , ka a ko ka santiya , ka sanu di a ma , ka wari di a ma , ka a di a ma .

meyer_gorog-contes_bambara_06subagamuso_ani_a_biranmuso,#12576 dennin bè dogò tigè yòrò **min** , a désèra ka dogò tigè .

meyer_gorog-contes_bambara_06subagamuso_ani_a_biranmuso,#13104 o dogò bèna bin yòrò **min** , Jasun nana ka i jò a kòfè .

meyer_gorog-contes_bambara_06subagamuso_ani_a_biranmuso,#13216 ne ye o sòrò yòrò **min** na , n ye o da yen .