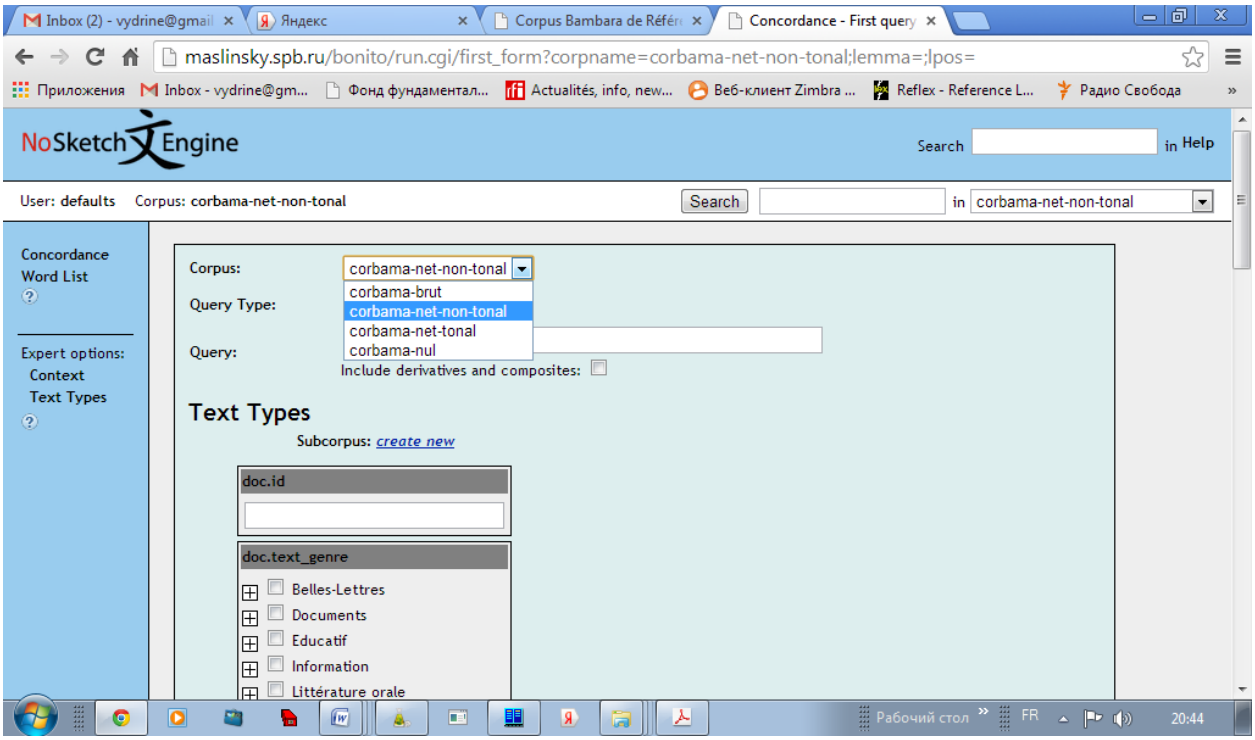
Инструкция по пользованию поисковой программой NoSketch Engine для корпуса текстов бамана Corpus bambara de référence

0. Корпус находится в открытом доступе по адресу:
http://maslinsky.spb.ru/bonito/run.cgi/first_form

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

1. **Подкорпуса**

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



- corbama-brut – подкорпус с неснятой омонимией (на март 2017 имевший объём около 3 146 000 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 март 2017 – около 700 000 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. 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

[1]Подкорпус 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).

возможность ещё больше уменьшить информационный шум, исключая квазиомонимы, отличающиеся от искомой формы тонами.

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

Правила тональной нотации в Корпусе изложены в Приложении 2.

1.2. **Аннотация**

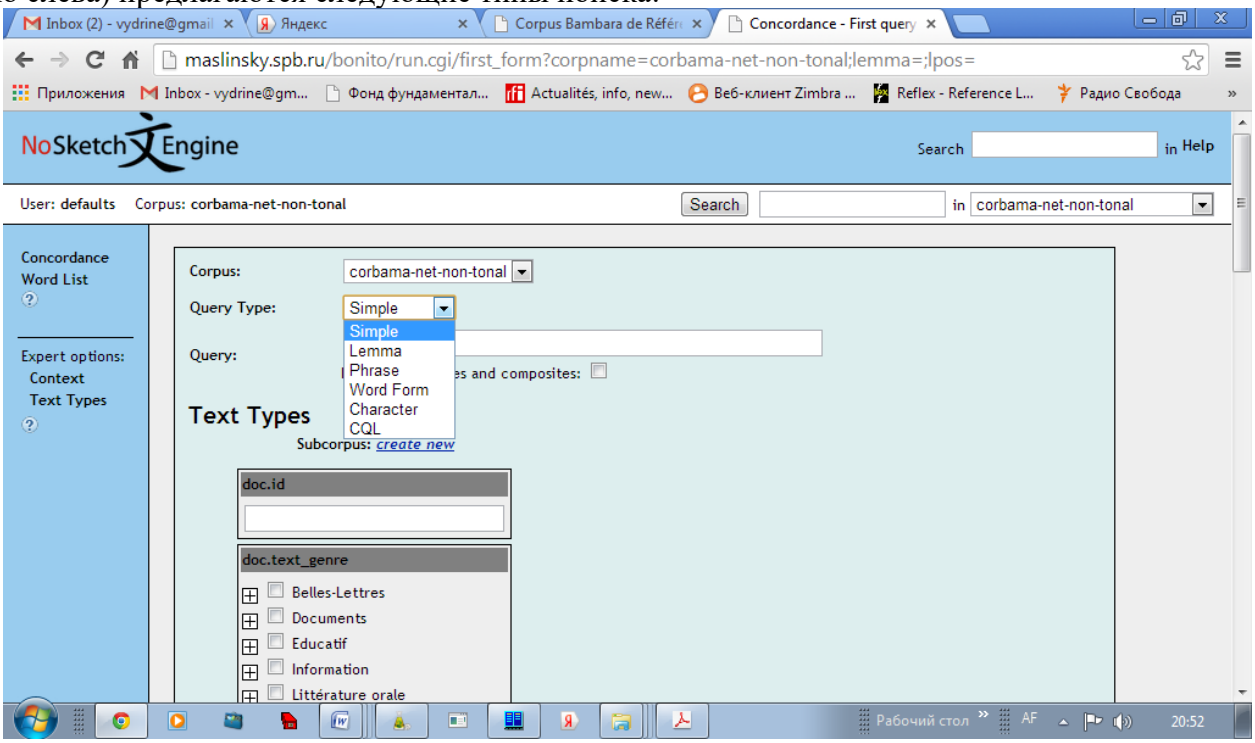
Все тексты разбиты на токены. Токен – это словоформа или знак препинания. Каждой словоформе и каждой морфеме в составе словоформы приписана лингвистическая аннотация.

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 исправленны, 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 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: ɲɛ̀ɲɛ ‘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: ɲɛ́nu ‘arbre.Hannoa.undulata’, ntómi ‘serpent.Eryx.muelleri’.

Типы поиска

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



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.

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’, *sára* ‘petit tas’, *sára* ‘charme’), без омонимичной перфективной формы глагола *sà* ‘mourir’. В отношении флективных 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 поиске 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).

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 будет найдены все случаи, когда в исходном тексте слово записано с ошибкой (даже в том случае, если эта ошибка была исправлена при ручном снятии омонимии).

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, dugukɔnɔmɔgɔ,* 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.

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.

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

2. **Ввод искомой формы**

3.1. 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.

3.2. 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]

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

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

3.3.1. 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 токену):

[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="kɔnɔ" & tag="n" & gloss="oiseau"]

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

3.3.2. 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="kɔnɔ" & gloss="à.l’intérieur"]

позволяет найти все сочетания *bàra kɔ́nɔ,* где первое слово – ‘калебаса’ (а не ‘chez’, ‘dancing’, ‘préféré’), 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]

3.3.3. 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 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 потенциалисного причастия, команда должна выглядеть так:

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

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

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

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

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

---

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

Уточним, что эти команды позволяют найти редупликанды, написанные раздельно. Если же нам нужны редуплицированные 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 перез знаком +, при этом им предшествует знак ^. Таким образом, команда «найти все редуплицированные формы, написанные слитно или через дефис, исключив из поиска цифры и знак %», выглядит так:

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

3.4. **Ввод нестандартных символов** (*ɔ, ɛ, ŋ, 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 *à, è, é, ù…* – другое дело, что далеко не всё необходимое можно с её помощью набрать);

– эти нестандартные символы можно заменять следующими комбинациями:

;o = ɔ

;e = ɛ

;n = ŋ

;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, → kɔ́

su` → sù

k;e,n;e → kɛ́nɛ

;m;o` → ɲɔ̀

;n;o`mi → ŋɔ̀mi.

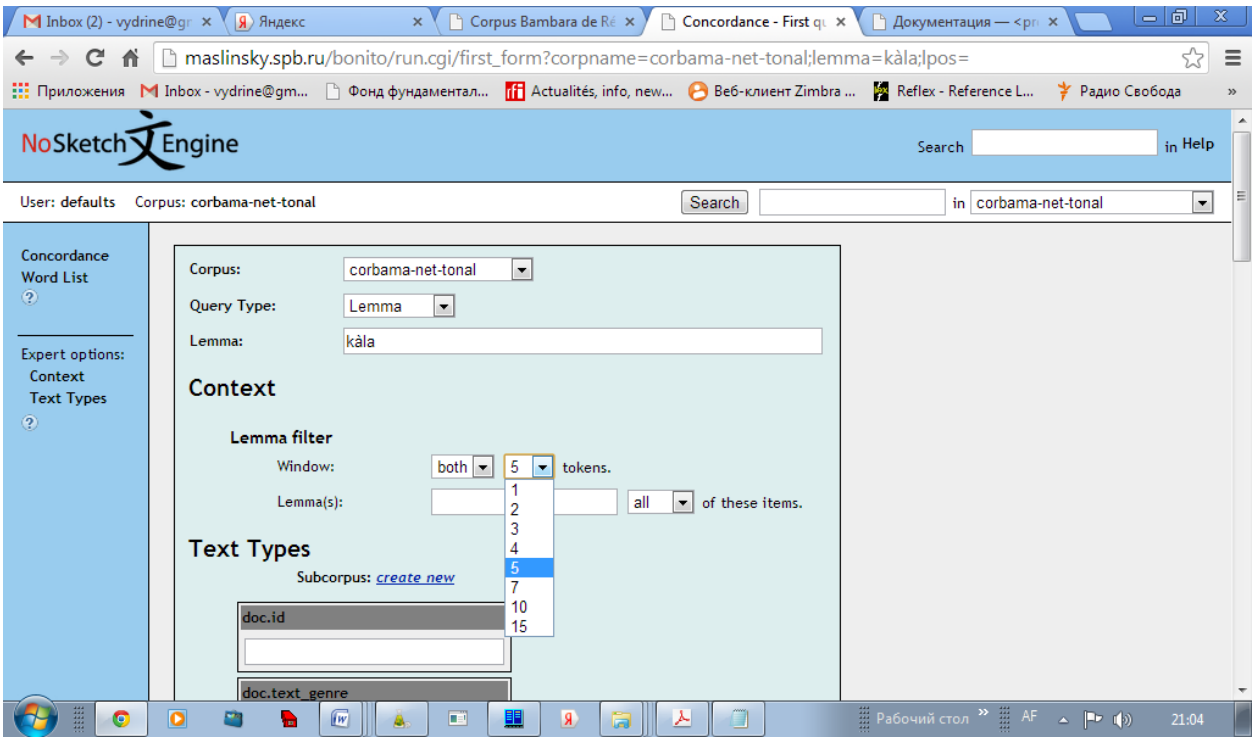
3. **Опция Context**

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 *kɛ* и двух контекстных – *yɛrɛ, ɲɔgɔn.* будут найдены такие примеры:

Mɔgɔ min bɛ a mɔgɔɲɔgɔn jogin , a ye min kɛ o tigi la , o ***ɲɔgɔn*** ka **kɛ** a ***yɛrɛ*** fana la .

O de bɛ cikɛla kɛ senyɛrɛkɔrɔbaga ye , i n' a fɔ birokɔnɔbaarakɛla ; i n' a fɔ taɲini julabaw , i n' a fɔ ***yɛrɛ*** jamanakuntigi n' a **kokeɲɔgɔnw** , senyɛrɛkɔrɔ siratɛgɛ la

Jatigikɛ ***yɛrɛ*** ɲuman na , a ***ɲɔgɔn*** cɛ kisɛ 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 послелогами, и т.п.

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

При всключённой 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.

4. **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 видеокоммуникации, 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).

5. **Concordance**

6.1. 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).

6.2. 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:

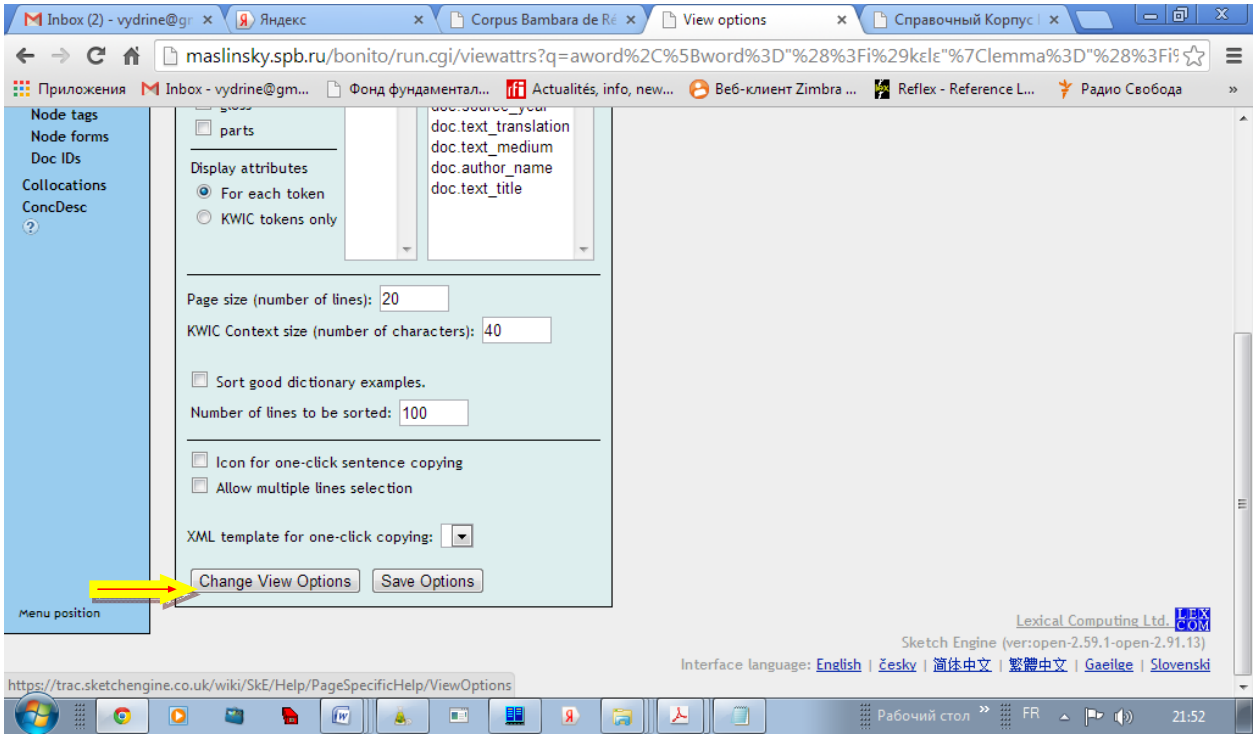
– изменить атрибуты формы (*Attributes*) – если пометить опции *word, lemma, tag, gloss,* то соответствующие атрибуты (лемма, частеречная помета, французская глосса; по умолчанию, опция *word* помечена всегда) при форме будут указаны. Атрибут *form* представляет собой «нормализованную словоформу» (в новой орфографии, тонированная, с поморфемной разбивкой дефисами). Атрибут *part* даёт доступ к поле, 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 умочанию выставляется цифра 40).

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

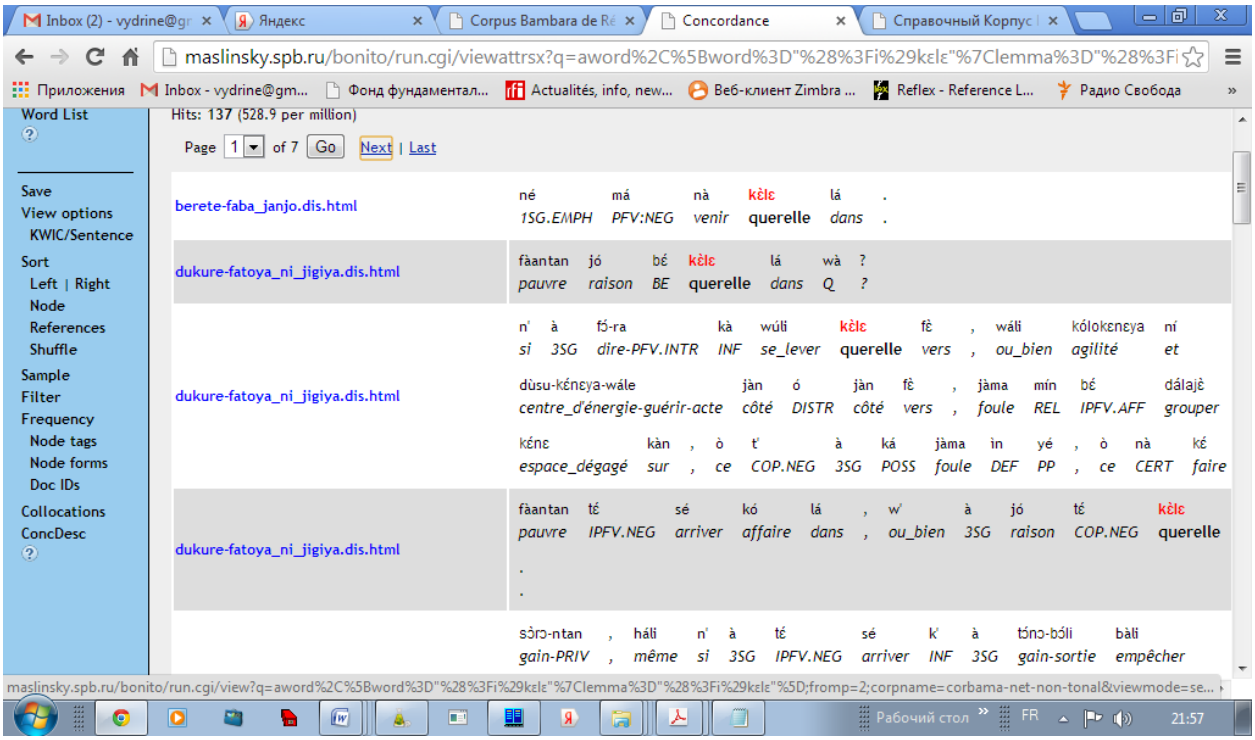


Остальные функции раздела *View Options* (*Sort good dictionary examples* и т.д.) для нашего корпуса нерелвантны.

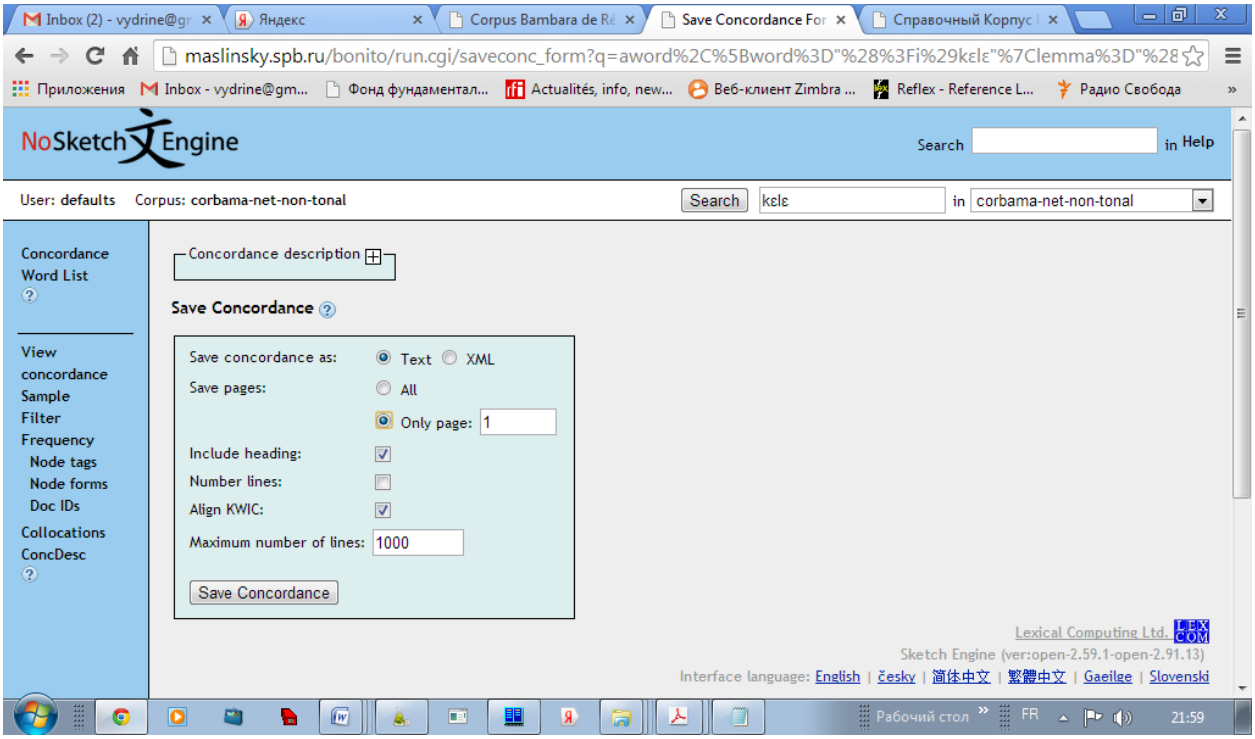
6.3. Полученный конкорданс может быть **сохранён** по частями или целиком в текстовом формате. Для сохранения целиком нужно выбрать опцию *Save* в меню слева.

Если пользователь хочет экспортировать глоссированный номализованный 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 кнопу *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.

6.4. **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).

6.5. Опция *Sample* позволяет делать случайную выборку заданного размера из числа всех найденных в корпусе примеров.

6.6. Опция *Filter* аналогична по своим функциям опции *Context,* описанной выше (раздел 4).

6.7. Опция *Frequency* 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* нерелевантна.

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.*

6.8. Раздел *Collocations* позволяет 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.*

7. Word List

Опция *Word List* позволяет создавать частотный словарь. Если войти в эту опцию, то в меню слева появляются эктикетки *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).