Скоба Мария, 2 группа

JRC-Acquis corpus

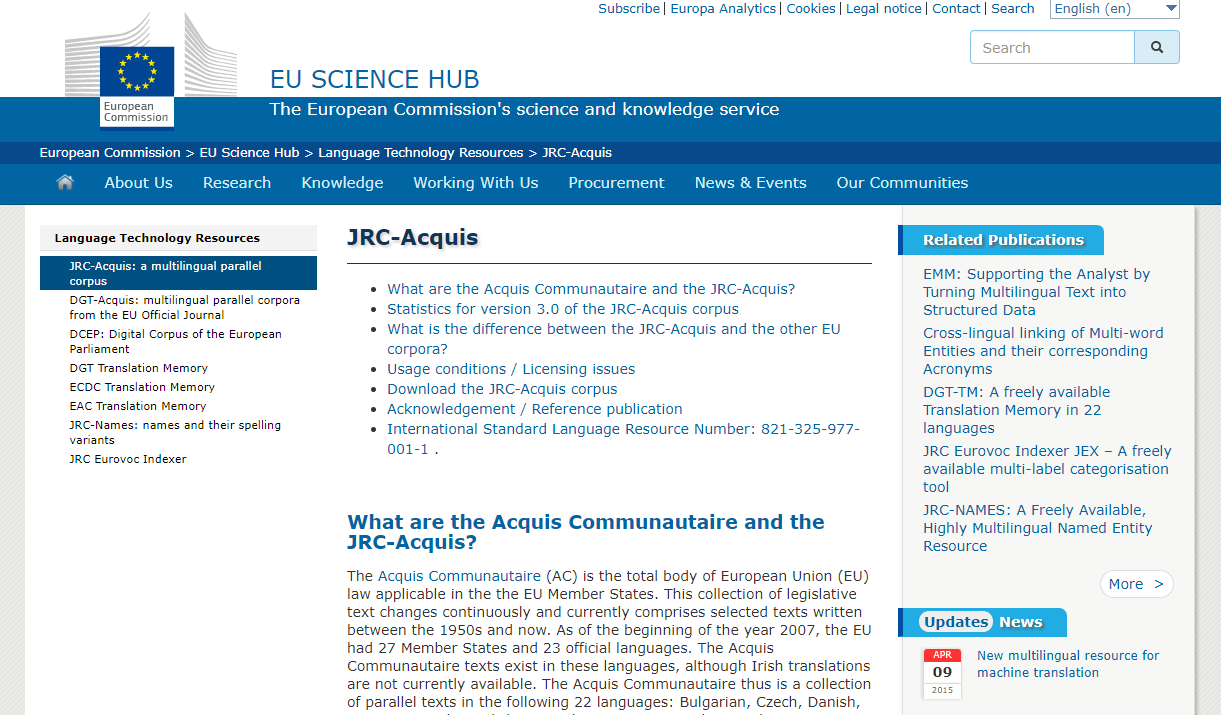
Введение

JRC-Acquis corpus – параллельный корпус, доступный на 22 из 23 официальных языках Европейского союза: болгарский, чешский, датский, немецкий, греческий, английский, испанский, эстонский, финский, французский, венгерский, итальянский, литовский, латвийский, мальтийский, голландский, польский, португальский, румынский, словацкий, словенский и шведский. Корпус на данный момент состоит из более чем 460 тыс. текстов и 1 млрд. слов.

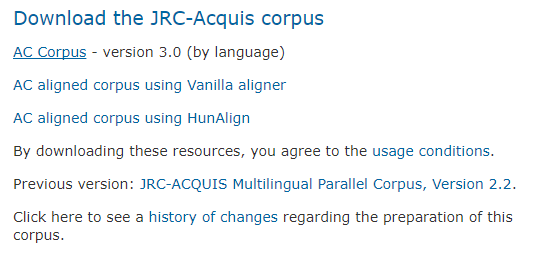
Дизайн

Сам корпус не представлен онлайн и сам состоит из файлов и архивов, доступных для скачивания. После всех манипуляций можно получить XML-файл, который в дальнейшем может использоваться с помощью Python и NLTK.

Говорить можно только о дизайне страницы, на которой всё это представлено. Страница (<https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis>) находится на сайте Еврокомиссии.



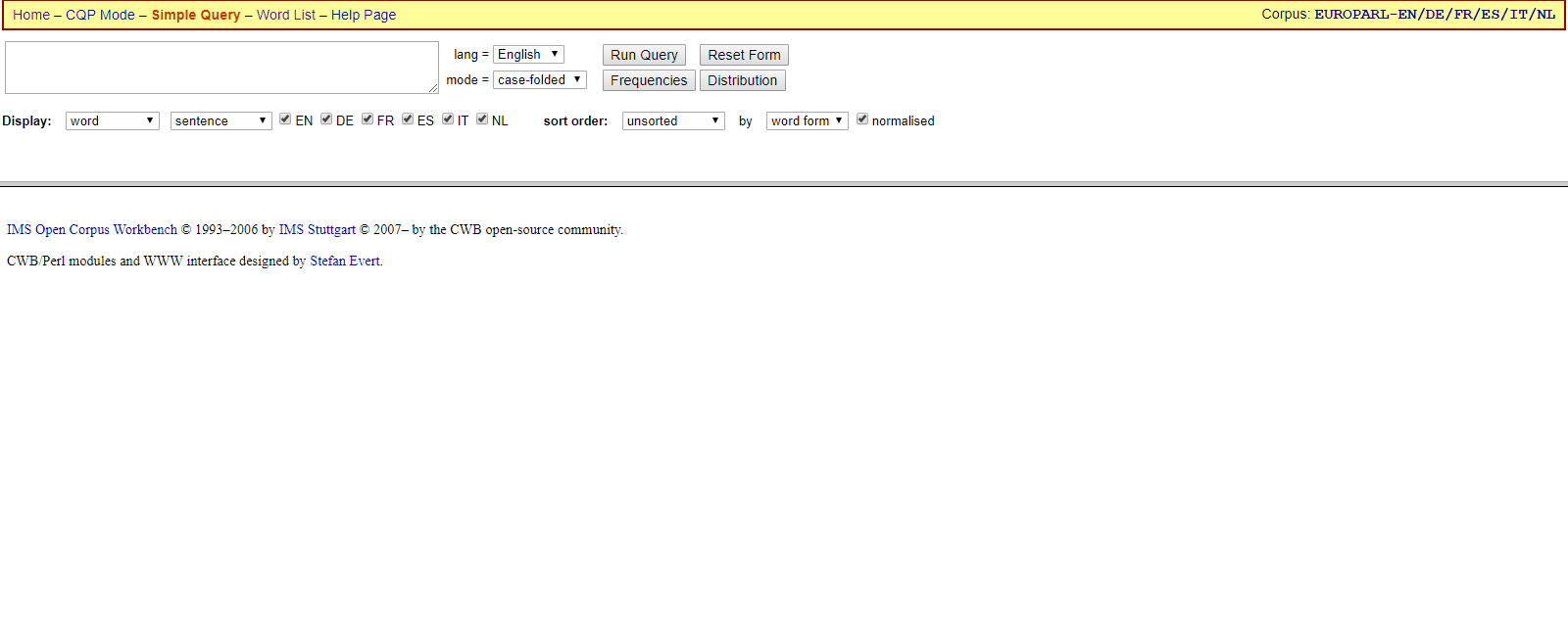
Из списка в начале страницы можно сразу же перейти к разделу про скачивание корпуса.



По ссылкам находятся архивы с XML-файлами для каждого языка, файлы с парами языков и гайд по созданию собственной пары для пользователей Windows.

Т.к. сам корпус не предлагает какого-либо интерфейса, было решено изучить NLTK, но задача оказалась невыполнимой в кратчайшие сроки, зато были найдены демо-версии онлайн-поиска по корпусу. Поэтому мы рассмотрим одну из из них.

<https://corpora.linguistik.uni-erlangen.de/demos/CQP/Europarl/frames-simple.html>



Здесь дизайн такой же простой, как и страница сайта Еврокомиссии.

* Смущает нагромождение кнопок, их можно было бы расположить по ширине всей страницы и снабдить пояснениями.
* Вверху есть переключение между тремя режимами запросов, которые переключаются сразу на странице, но было не совсем очевидно, что их вообще можно переключать.
* Остаётся очень много пустого места, было бы хорошо разместить там примеры запросов и результатов поиска.

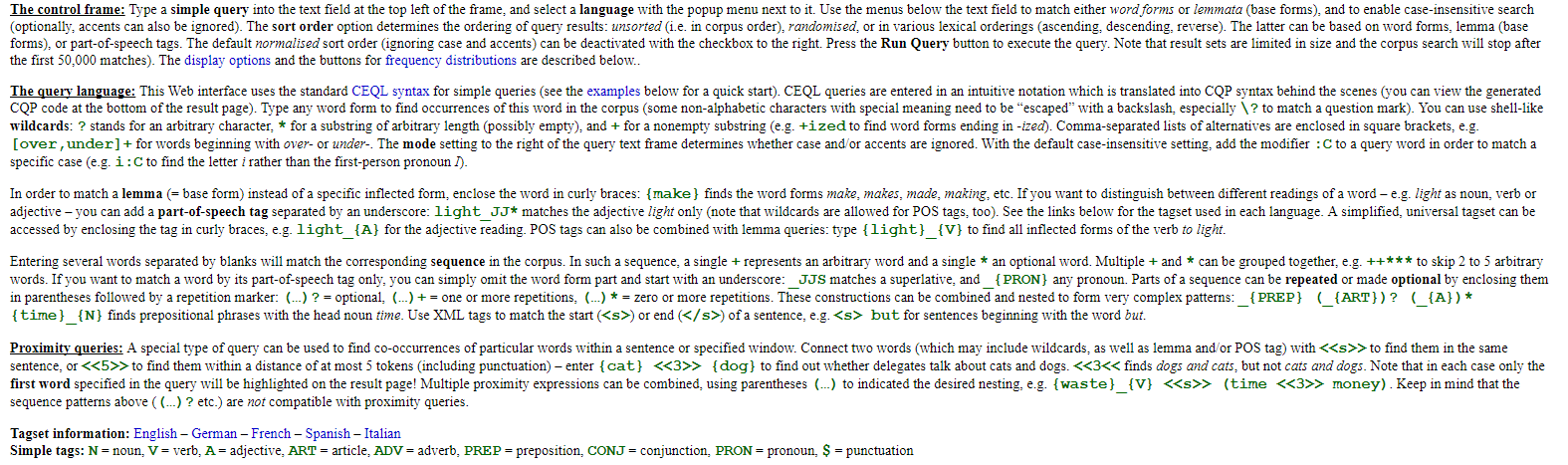
Ресурс глазами новичка

В Гугле ресурс находится на четвёртой позиции по запросу «europarlament online corpus», но первые две ссылки нерабочие, а на третьем месте находится сайт Еврокомиссии, так что найти его совсем несложно.

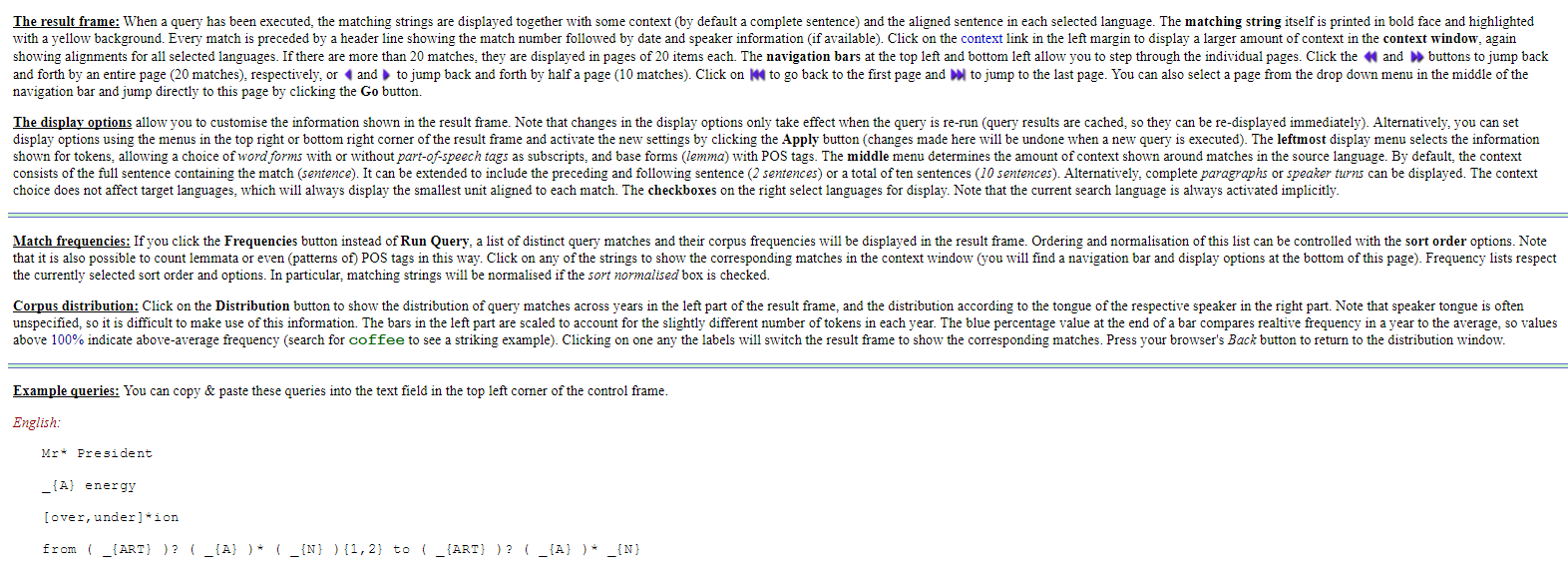
Назначение всех кнопок интуитивно понятно, но к кнопкам выдачи результатов хотелось бы видеть некоторые пояснения.

Неудобно, что каждый новый запрос открывается не в отдельном окне, а вместо предыдущей выдачи в том же окне.

Вверху есть ссылка help page, на которой находится полное руководство по поиску в корпусе. Оно не очень удобное, т.к. нет никаких иллюстраций, а текст слишком мелкий и в огромном количестве.



В руководство включены также примеры запросов, но опять же нет пояснений и иллюстраций с выдачами.

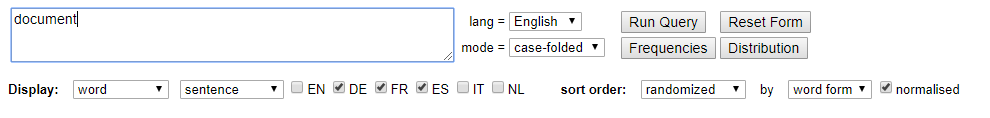


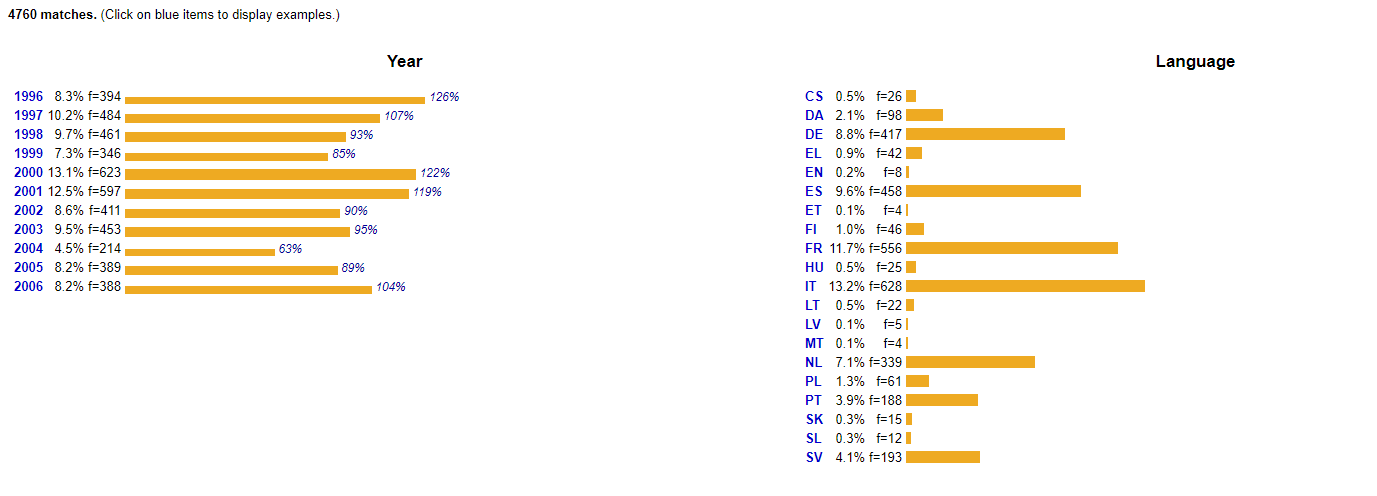
Помощь пользователю

Как было выше сказано, подсказок и пояснений нет нигде кроме страницы с руководством. Она в свою очередь очень неудобная из-за того, что нет иллюстраций к использованию, а сам пользователь не может одновременно смотреть и в руководство, и на страницу для запросов, т.к. они в разных окнах. Руководство плохо поделено на главы и воспринимается сплошным текстом, что затрудняет поиск ответа на конкретный вопрос.

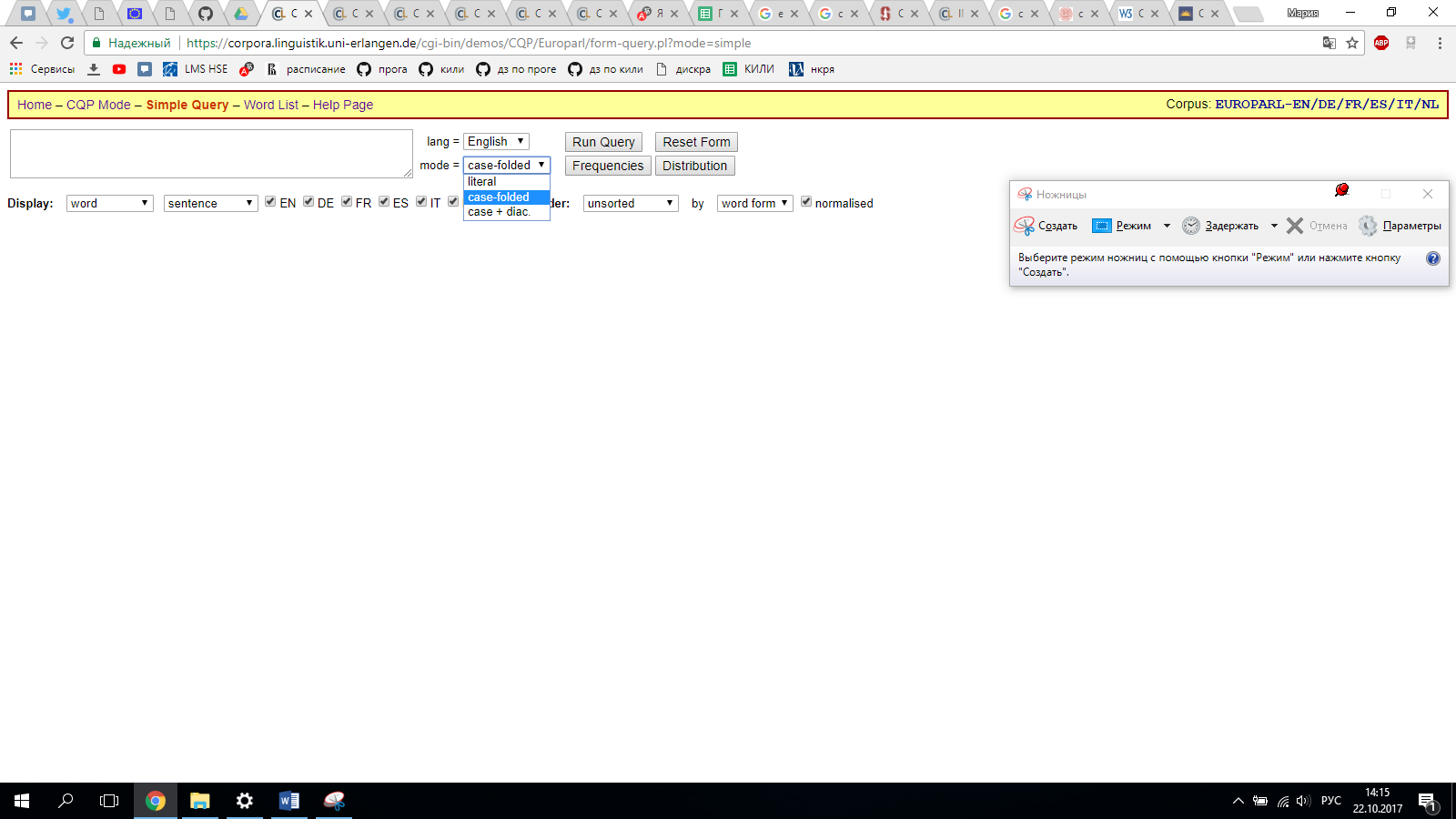
Продвинутый функционал

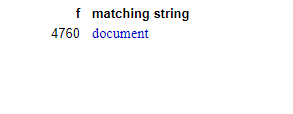
Ресурс предлагает параллельный поиск только по шести языкам из доступных двадцати двух, но кнопка «distribution» позволяет посмотреть количество вхождений слова в каждый из доступных языков корпуса (вне зависимости от языка, выбранного для запроса) и распределение количества слов по годам в выбранном языке.





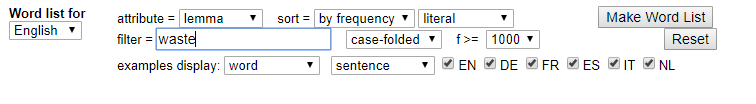
Также окно запросов позволяет выбрать значимость регистра.

С помощью кнопки «Frequencies» можно узнать количество вхождений в язык.



Режимы «Word List» и «CQP» остаются загадкой, т.к. судя по заголовку руководство дано только для режима «Simple Query». Методом проб и ошибок было выяснено, что второй режим вообще не работает, а первый позволяет посмотреть варианты перевода.

Пример работы режима с переводом:

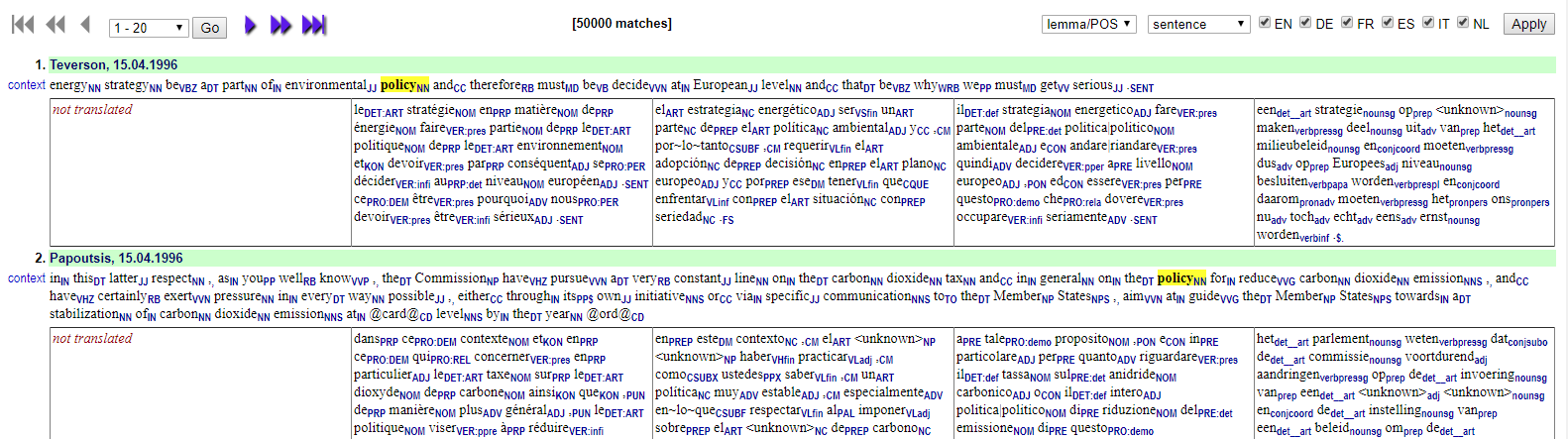




Также существуют несколько способов сортировки и отображения выдачи. Помимо обычной выдачи слов есть выдача «слово + часть речи»:



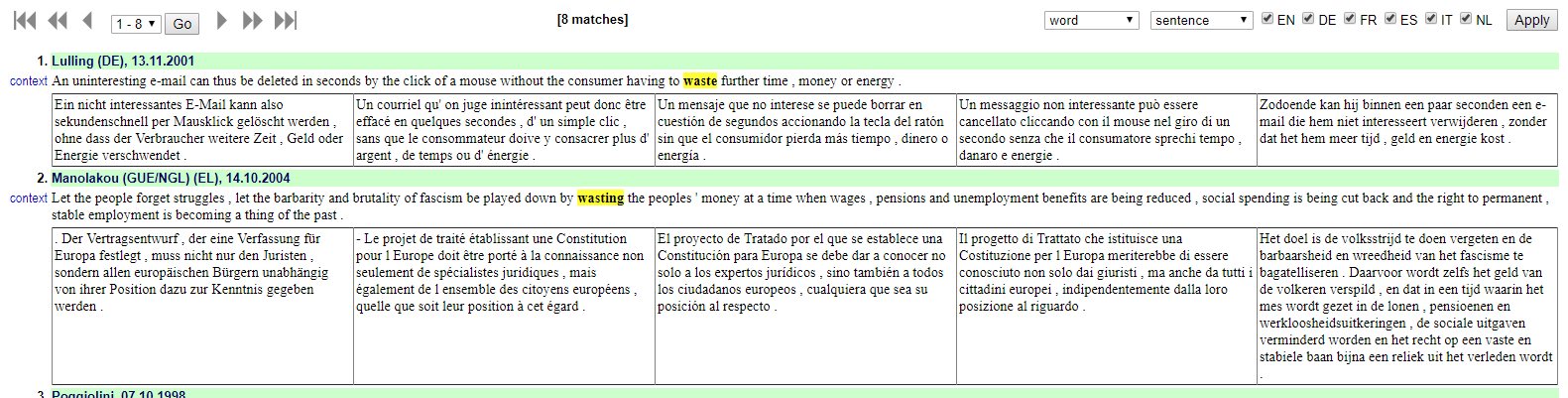
и «лемма + часть» речи:



Так же есть функция случайной сортировки и сортировки по словоформам, частям речи и леммам.

Примеры сложных запросов

{waste}\_{V} <<s>> (time <<3>> money) – глагол waste в любой форме в одном предложении с time и money:



\_{PREP} (\_{ART})? (\_{A})\* {time}\_{N} – предложные конструкции, где главное слово – time:

