NATIONAL RESEARCH UNIVERSITY

Елизавета Смирнова
esmirnova2@hse.ru
cmelizaveta@yandex.ru

ОБЗОР СУЩЕСТВУЮЩИХ ИНСТРУМЕНТОВ ДЛЯ КОРПУСНОГО АНАЛИЗА ТЕКСТОВ

Национальный исследовательский университет «Высшая школа экономики», Пермь

Пермь, 2019

Elizaveta Smirnova
esmirnova2@hse.ru
cmelizaveta@yandex.ru

ПЛАН

- Наиболее известные инструменты корпусного анализа и их особенности
- Основные функции одного из популярных инструментов
- Применение корпусного анализа в педагогической практике

Elizaveta Smirnova
esmirnova2@hse.ru
cmelizaveta@yandex.ru

КОРПУСНАЯ ЛИНГВИСТИКА: ОСОБЕННОСТИ

- эмпирический подход; анализируются языковые конструкции в реальных контекстах использования
- в качестве материала для анализа используется репрезентативная выборка языковых средств, которые хранятся в виде электронной базы данных (корпуса)
- часть анализа – изучение языковых моделей с помощью специализированных компьютерных программ
- результаты интерпретируются с применением количественных и качественных методов (Biber et al., 1998).

Elizaveta Smirnova
esmirnova2@hse.ru
cmelizaveta@yandex.ru

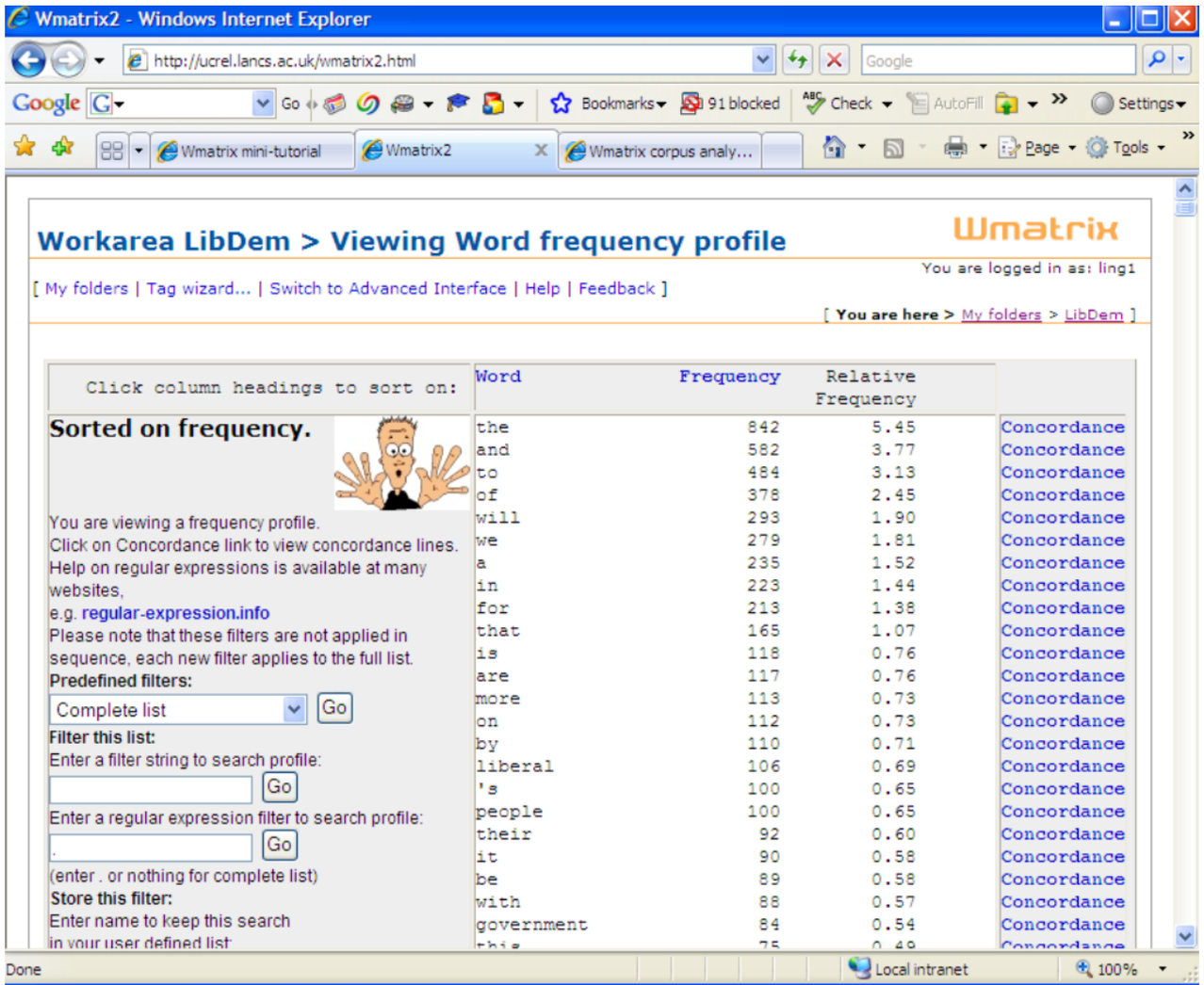
НАИБОЛЕЕ ПОПУЛЯРНЫЕ ИНСТРУМЕНТЫ

- AntConc
- WMatrix
- SketchEngine
- WordSmith

Elizaveta Smirnova
esmirnova2@hse.ru
cmelizaveta@yandex.ru

WMATRIX

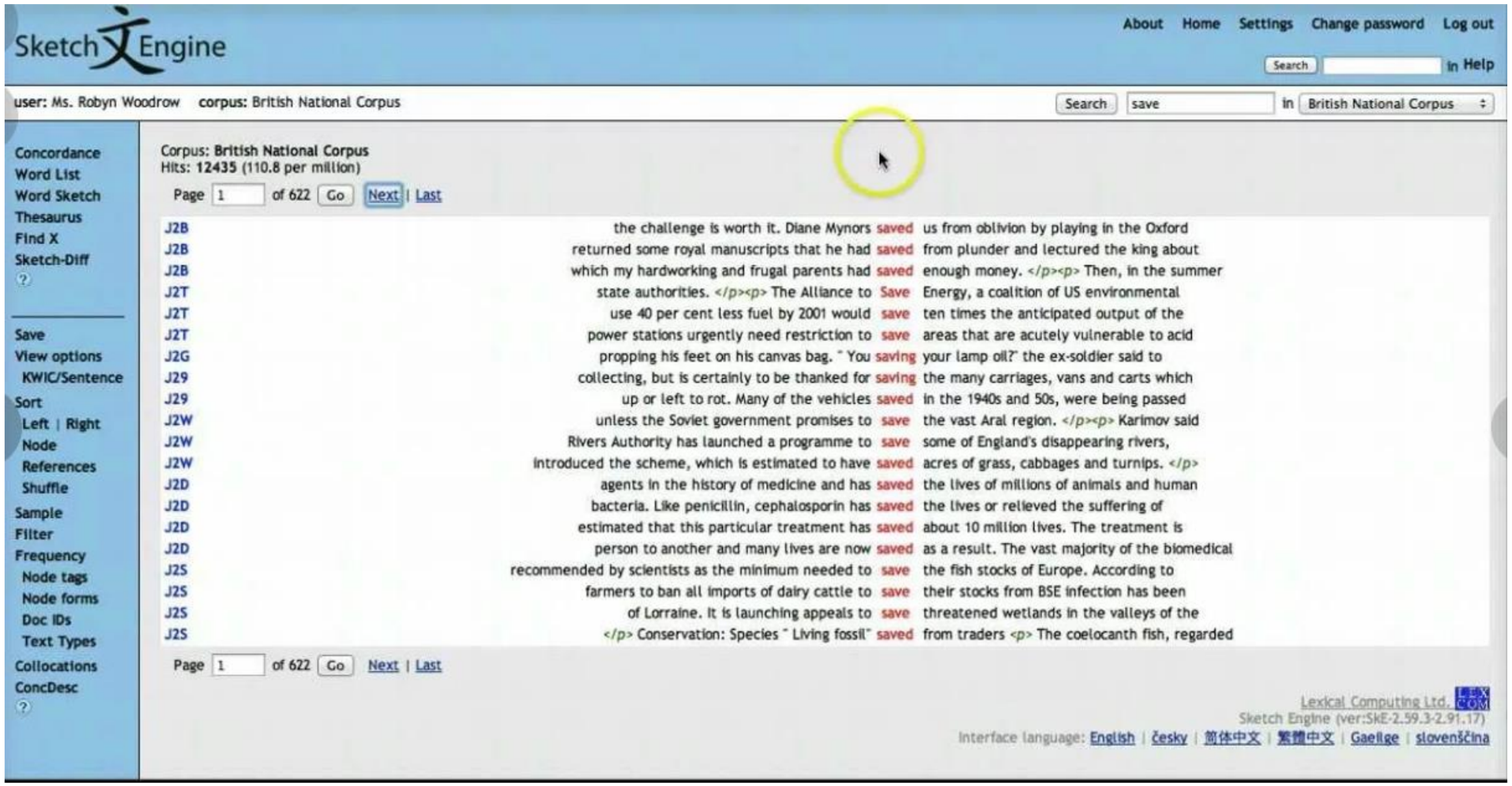
- разработчик – Пол Рейсон, Университет Ланкастера (Великобритания)
- сайт http://ucrel.lancs.ac.uk/wmatrix/
- интерфейс



Elizaveta Smirnova
esmirnova2@hse.ru
cmelizaveta@yandex.ru

SKETCH ENGINE

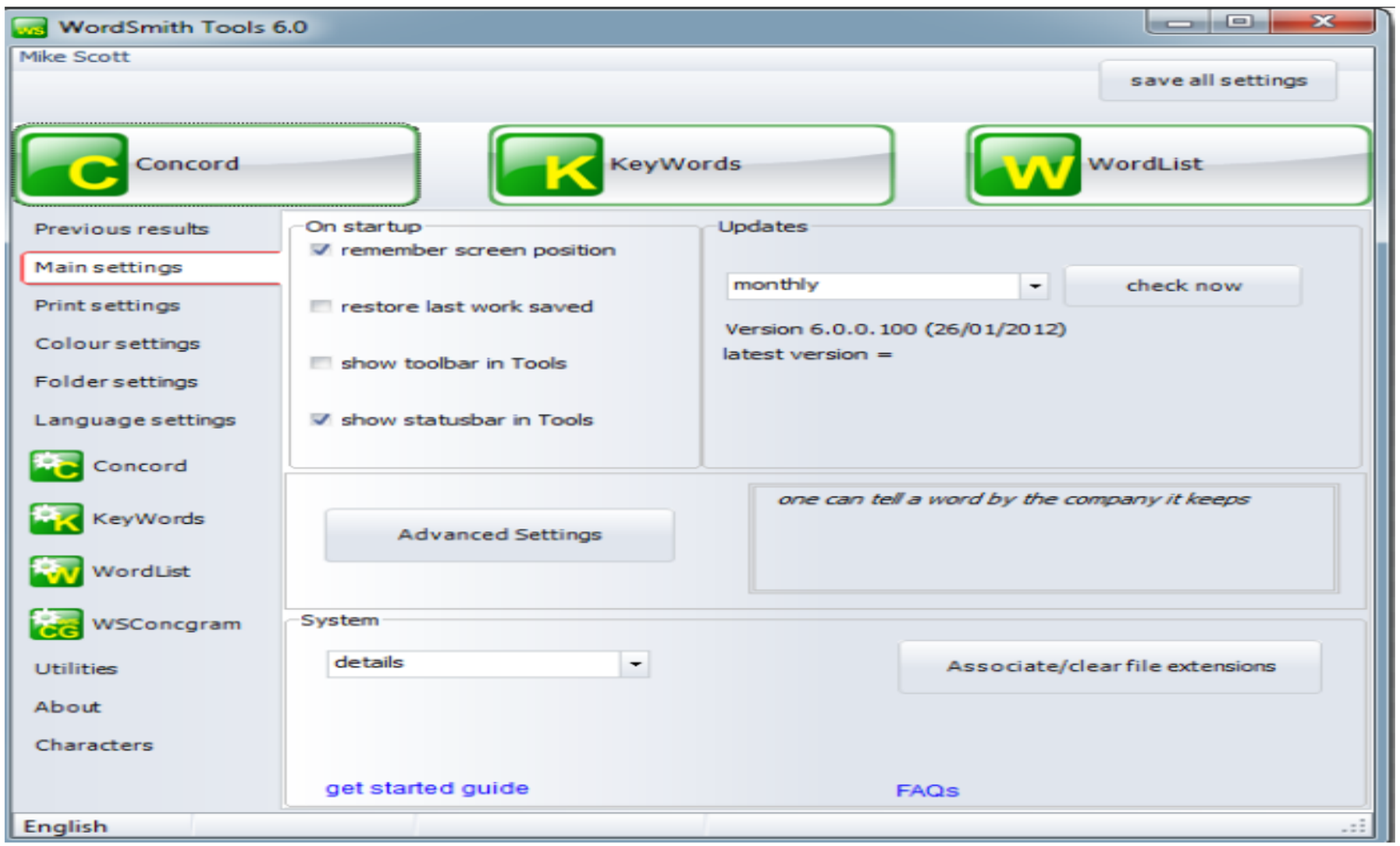
- разработчик – Lexical Computing Limited (Adam Kilgarriff & Pavel Rychlý)
- сайт https://www.sketchengine.eu
- интерфейс



Elizaveta Smirnova
esmirnova2@hse.ru
cmelizaveta@yandex.ru

WORDSMITH

- разработчик – by Lexical Analysis Software and Oxford University Press
- сайт https://lexically.net/wordsmith/
- интерфейс

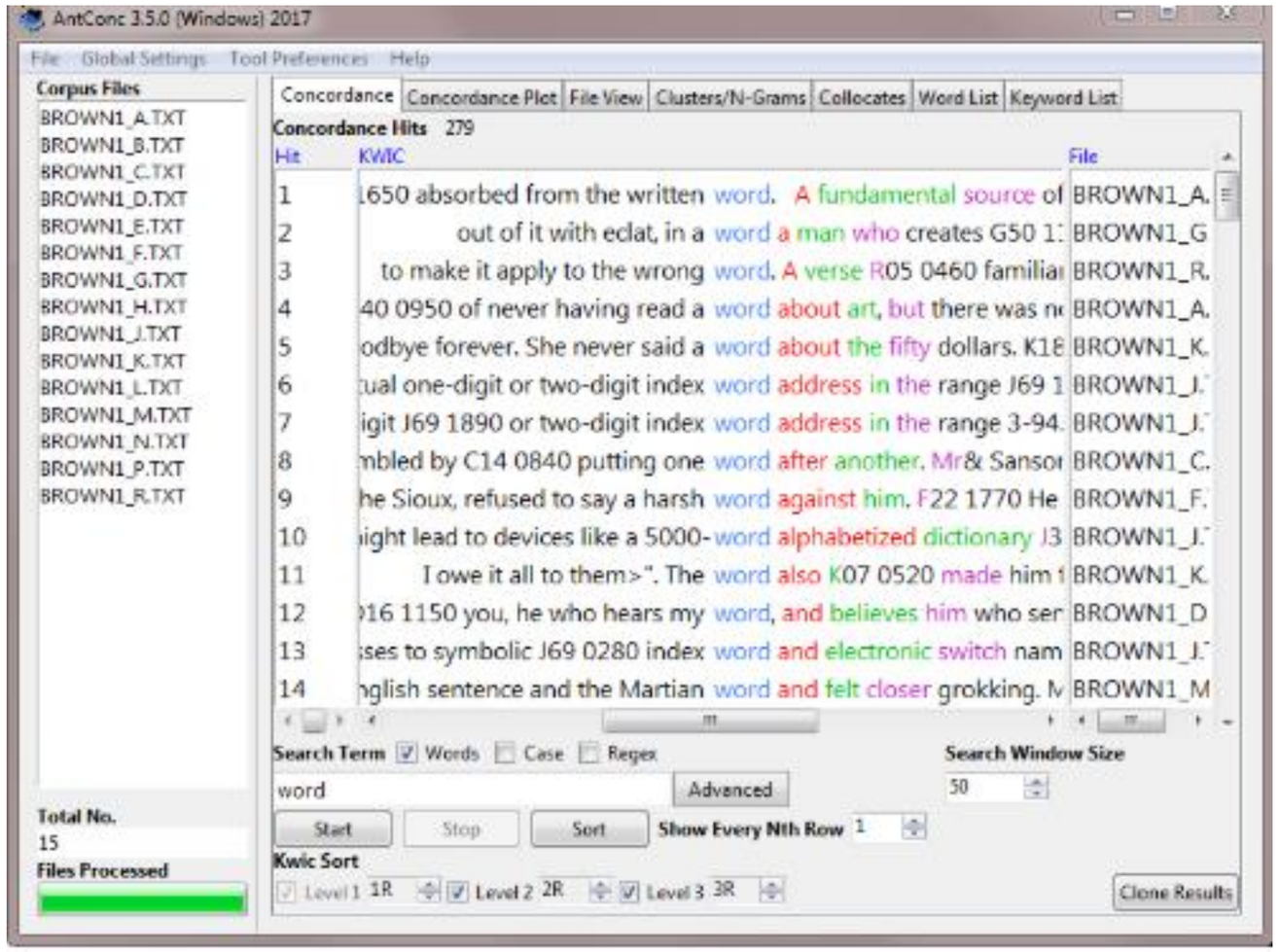


Elizaveta Smirnova
esmirnova2@hse.ru
cmelizaveta@yandex.ru

ANTCONC

- разработчик – Laurence Anthony, Waseda University (Japan)
- Сайт https://laurenceanthony.net/software/antconc/

- интерфейс



Elizaveta Smirnova
esmirnova2@hse.ru
cmelizaveta@yandex.ru

СРАВНЕНИЕ ФУНКЦИОНАЛА

Name	Functions							Pricing
	Tagging (pos/semantic)	concordancer	wordlists	keywords	statistics	Corpus Creation	Semantic Analysis	
AntConc	With Ant Tag	+	+	+	+			Free
WMatrix	Pos, semantic	+	+	+	+			50 pounds per username per year
WordSmith	Pos, semantic	+	+	+	+			60 euros per licence
Sketch Engine	pos	+	+	+	+	+	+	30-day trial, then 4.85 per month

Elizaveta Smirnova
esmirnova2@hse.ru
cmelizaveta@yandex.ru

ANTCONC: FUNCTIONS

- Wordlist
- Keyword list
- Concordance
- Clusters/Ngrams
- Concordance plot

Elizaveta Smirnova
esmirnova2@hse.ru
cmelizaveta@yandex.ru

ANTCONC: WORDLIST

- запустите **AntConc**, кликните вкладку **File** в левом верхнем углу, затем – **Open files** и загрузите файлы из папки **Data** на рабочем столе
- Список файлов и их общее количество должны появиться слева
- Кликните вкладку **Wordlist** (предпоследняя в верхней строчке) и нажмите кнопку **Start** внизу
- Найдите данные о:

- общем количестве слов (токенов) в корпусе

- количестве типов слов

-какое слово является самым частотным? сколько слов имеет частотность больше 100?

Elizaveta Smirnova
esmirnova2@hse.ru
cmelizaveta@yandex.ru

ANTCONC: KEYWORD LIST

- Ключевое слово в корпусной лингвистике – слово, которое встречается чаще в исследуемом корпусе/тексте, чем в контрольном корпусе, который обычно больше по размеру.
- Для определения ключевых слов используются статистические тесты (log likelihood и effect size в AntConc)

Elizaveta Smirnova
esmirnova2@hse.ru
cmelizaveta@yandex.ru

ANTCONC: KEYWORD LIST (2)

- Кликните вкладку Tool **Preferences**, затем – **Keyword list**
- В разделе **Reference corpus** нажмите кнопку **Add Files**, откройте папку **Reference corpus** на рабочем столе, выделите все файлы, нажмите на кнопку открыть, затем **Load**
- В правом нижнем углу окна кликните **Apply**
- Теперь кликните вкладку **Keyword list** и кнопку **Start**
- Слова можно отсортировать по keyness (статистической значимости) и по частотности.

Elizaveta Smirnova
esmirnova2@hse.ru
cmelizaveta@yandex.ru

ANTCONC: KEYWORD LIST (3)

- Какое самое частотное значимое слово в нашем корпусе?
- Какой ранг у слова Perm при сортировке по keyness? Меняется ли он, если отсортировать ключевые слова по частотности?

Elizaveta Smirnova
esmirnova2@hse.ru
cmelizaveta@yandex.ru

ANTCONC: CONCORDANCE

Слова в контексте

- Кликните вкладку **Concordance**
- Введите в строку поиска слово *Perm* и нажмите **Start**
- Для того, чтобы отсортировать результаты по слову, которое находится слева, нужно поставить галочку в первом квадратике в нижней строке, а рядом выбрать **1L** и нажать **Sort**.
- Отсортируйте кокордансы по первому слову справа. Сколько раз встречается сочетание *Perm company/companies*?

Elizaveta Smirnova
esmirnova2@hse.ru
cmelizaveta@yandex.ru

ANTCONC: CLUSTERS/NGRAMS

Эта функция позволяет искать частотные словосочетания в корпусе

Кликните вкладку **Clusters/Ngrams**. В строке поиска у нас стоит слово Perm. Кликните **Start**.

Самое частотное правое сочетание, состоящее из двух слов, – *Perm krai*

Найдите самое частотное сочетание слов слева

Elizaveta Smirnova
esmirnova2@hse.ru
cmelizaveta@yandex.ru

ANTCONC: CONCORDANCE PLOT

Данная функция позволяет посмотреть, как распределены слова в текстах.

Откройте вкладку **Concordance plot**

Кликните **Start**

В каком количестве текстов нашего корпуса используется слово *Perm*?

Каково максимальное количество употреблений данного слова в одном тексте?

Elizaveta Smirnova
esmirnova2@hse.ru
cmelizaveta@yandex.ru

ANTCONC: ПОЛЕЗНЫЕ ССЫЛКИ

Официальный сайт
https://laurenceanthony.net/software/antconc/

Ссылка на канал разработчика на Youtube (на английском)
https://www.youtube.com/user/AntlabJPN

Ссылка на инструкцию на русском
https://lektsii.org/5-72958.html

Elizaveta Smirnova
esmirnova2@hse.ru
cmelizaveta@yandex.ru

DDL- DATA DRIVEN LEARNING

В широком смысле – использование учащимися корпусов на занятии под руководством преподавателя

Учащиеся выступают как исследователи, изучающие грамматические и лексические явления

Прямой (1) и косвенный (hard) (2) подходы (soft)

1 учащиеся сами работают с корпусом

2 преподаватель готовит материалы, используя корпус, и работает с ними на занятии

Elizaveta Smirnova
esmirnova2@hse.ru
cmelizaveta@yandex.ru

DDL- DATA DRIVEN LEARNING

- Готовые корпусы
- http://ota.ox.ac.uk/catalogue/index.html
- Собственный корпус (E.g. Antconc, SketchEngine)

Elizaveta Smirnova
esmirnova2@hse.ru
cmelizaveta@yandex.ru

REFERENCES

- Anthony, L., Flowerdew, J., & Costley, T. 2016. Introducing corpora and corpus tools into the technical writing classroom through Data-Driven Learning (DDL). In *Discipline specific writing* (pp. 162-180). Routledge
- Anthony, L. 2013. A critical look at software tools in corpus linguistics. *Linguistic Research 30(2*): 141–161.
- Biber, D, Conrad, S, and Reppen, R. 1998. *Corpus linguistics*. Cambridge: Cambridge University Press.

Elizaveta Smirnova
esmirnova2@hse.ru
cmelizaveta@yandex.ru

THANK YOU FOR YOUR ATTENTION!

NATIONAL RESEARCH
UNIVERSITY

www.perm.hse.ru/en/bi/sfcr/ Phone: +7(912)4961553 Address: 38, Studencheskaya St., Perm, Russia