



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



ПЛАН

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



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).

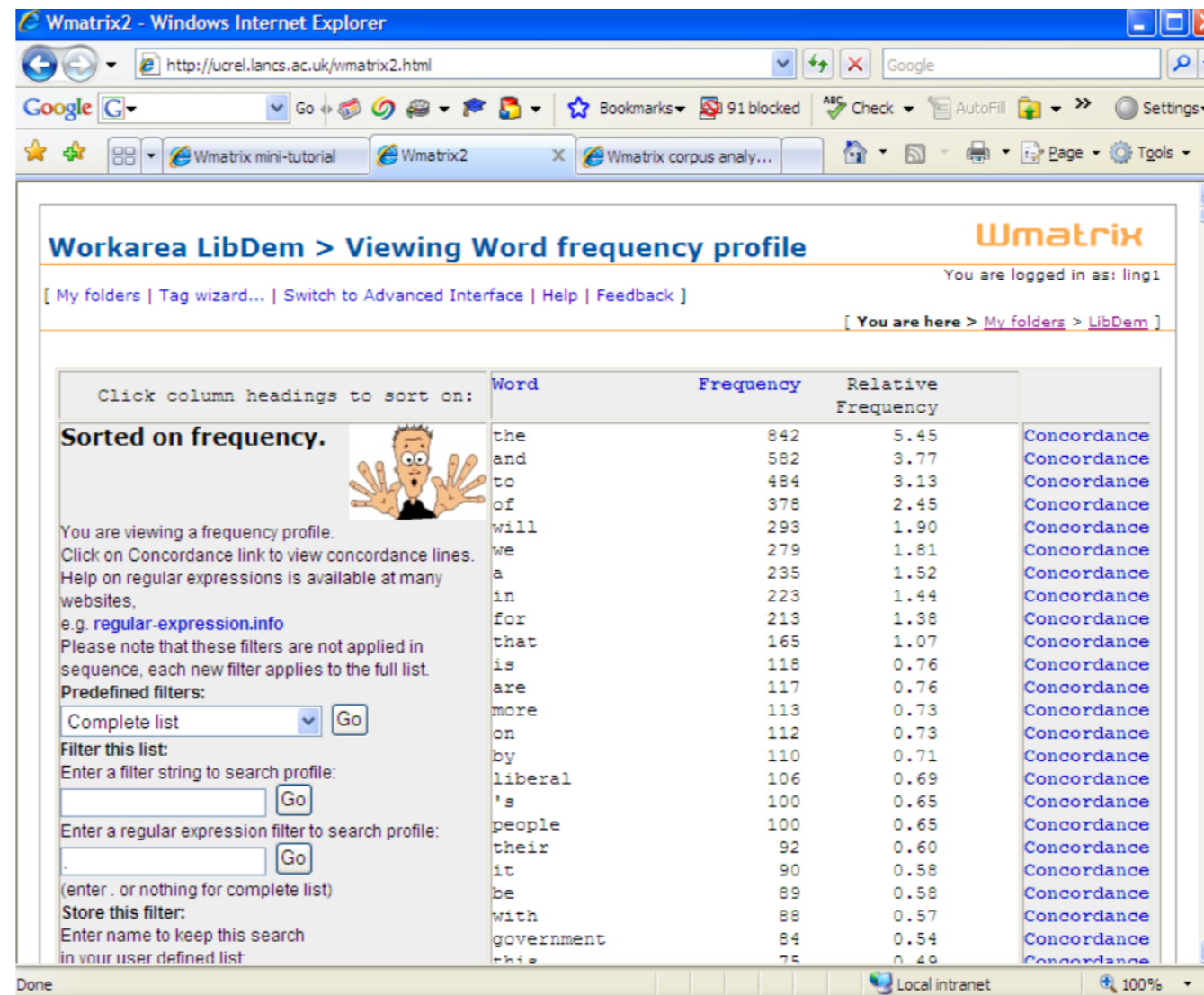


НАИБОЛЕЕ ПОПУЛЯРНЫЕ ИНСТРУМЕНТЫ

- AntConc
- WMatrix
- SketchEngine
- WordSmith

WMATRIX

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



Wmatrix2 - Windows Internet Explorer

http://ucrel.lancs.ac.uk/wmatrix2.html

Google

Wmatrix2

Workarea LibDem > Viewing Word frequency profile

You are logged in as: ling1

[My folders | Tag wizard... | Switch to Advanced Interface | Help | Feedback]

[You are here > My folders > LibDem]

Click column headings to sort on:

Sorted on frequency.

You are viewing a frequency profile.
Click on Concordance link to view concordance lines.
Help on regular expressions is available at many websites,
e.g. regular-expression.info
Please note that these filters are not applied in sequence, each new filter applies to the full list.

Predefined filters:
Complete list

Filter this list:
Enter a filter string to search profile:

Enter a regular expression filter to search profile:

(enter . or nothing for complete list)

Store this filter:
Enter name to keep this search
in your user defined list

Word	Frequency	Relative Frequency	
the	842	5.45	Concordance
and	582	3.77	Concordance
to	484	3.13	Concordance
of	378	2.45	Concordance
will	293	1.90	Concordance
we	279	1.81	Concordance
a	235	1.52	Concordance
in	223	1.44	Concordance
for	213	1.38	Concordance
that	165	1.07	Concordance
is	118	0.76	Concordance
are	117	0.76	Concordance
more	113	0.73	Concordance
on	112	0.73	Concordance
by	110	0.71	Concordance
liberal	106	0.69	Concordance
's	100	0.65	Concordance
people	100	0.65	Concordance
their	92	0.60	Concordance
it	90	0.58	Concordance
be	89	0.58	Concordance
with	88	0.57	Concordance
government	84	0.54	Concordance
this	75	0.49	Concordance

Done Local intranet 100%

SKETCH ENGINE

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

Sketch Engine

About Home Settings Change password Log out

Search in Help

user: Ms. Robyn Woodrow corpus: British National Corpus

Search save in British National Corpus

Corpus: British National Corpus
Hits: 12435 (110.8 per million)

Page 1 of 622 Go Next Last

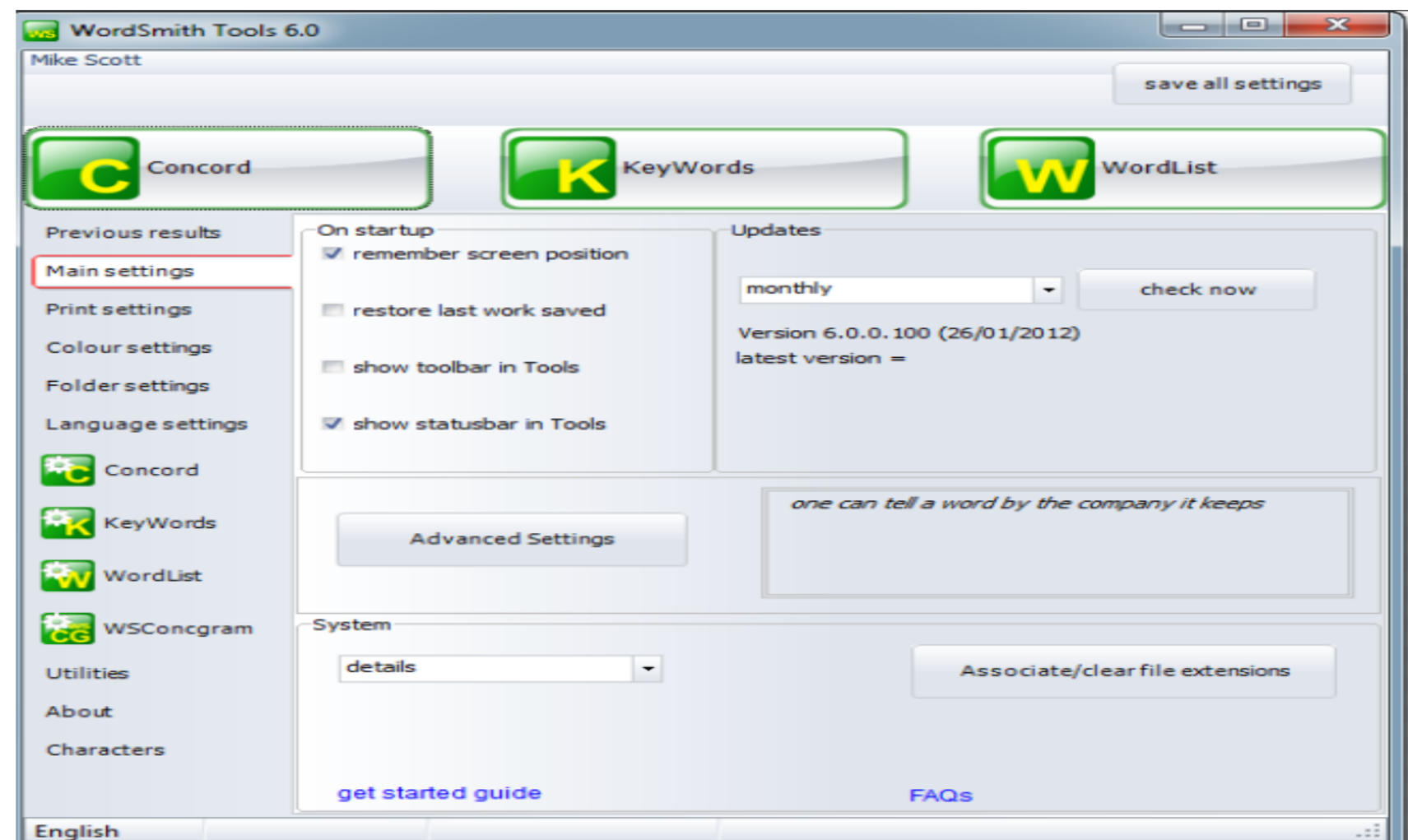
J2B the challenge is worth it. Diane Mynors saved us from oblivion by playing in the Oxford
J2B returned some royal manuscripts that he had saved from plunder and lectured the king about
J2B which my hardworking and frugal parents had saved enough money. </p><p> Then, in the summer
J2T state authorities. </p><p> The Alliance to Save Energy, a coalition of US environmental
J2T use 40 per cent less fuel by 2001 would save ten times the anticipated output of the
J2T power stations urgently need restriction to save areas that are acutely vulnerable to acid
J2G propping his feet on his canvas bag. " You saving your lamp oil? the ex-soldier said to
J29 collecting, but is certainly to be thanked for saving the many carriages, vans and carts which
J29 up or left to rot. Many of the vehicles saved in the 1940s and 50s, were being passed
J2W unless the Soviet government promises to save the vast Aral region. </p><p> Karimov said
J2W Rivers Authority has launched a programme to save some of England's disappearing rivers,
J2W introduced the scheme, which is estimated to have saved acres of grass, cabbages and turnips. </p>
J2D agents in the history of medicine and has saved the lives of millions of animals and human
J2D bacteria. Like penicillin, cephalosporin has saved the lives or relieved the suffering of
J2D estimated that this particular treatment has saved about 10 million lives. The treatment is
J2D person to another and many lives are now saved as a result. The vast majority of the biomedical
J2S recommended by scientists as the minimum needed to save the fish stocks of Europe. According to
J2S farmers to ban all imports of dairy cattle to save their stocks from BSE infection has been
J2S of Lorraine. It is launching appeals to save threatened wetlands in the valleys of the
J2S </p> Conservation: Species " Living fossil" saved from traders <p> The coelocanth fish, regarded

Page 1 of 622 Go Next Last

Lexical Computing Ltd.
Sketch Engine (ver: SKE-2.59.3-2.91.17)
Interface language: English | Český | 简体中文 | 繁體中文 | Gaeilge | slovenščina

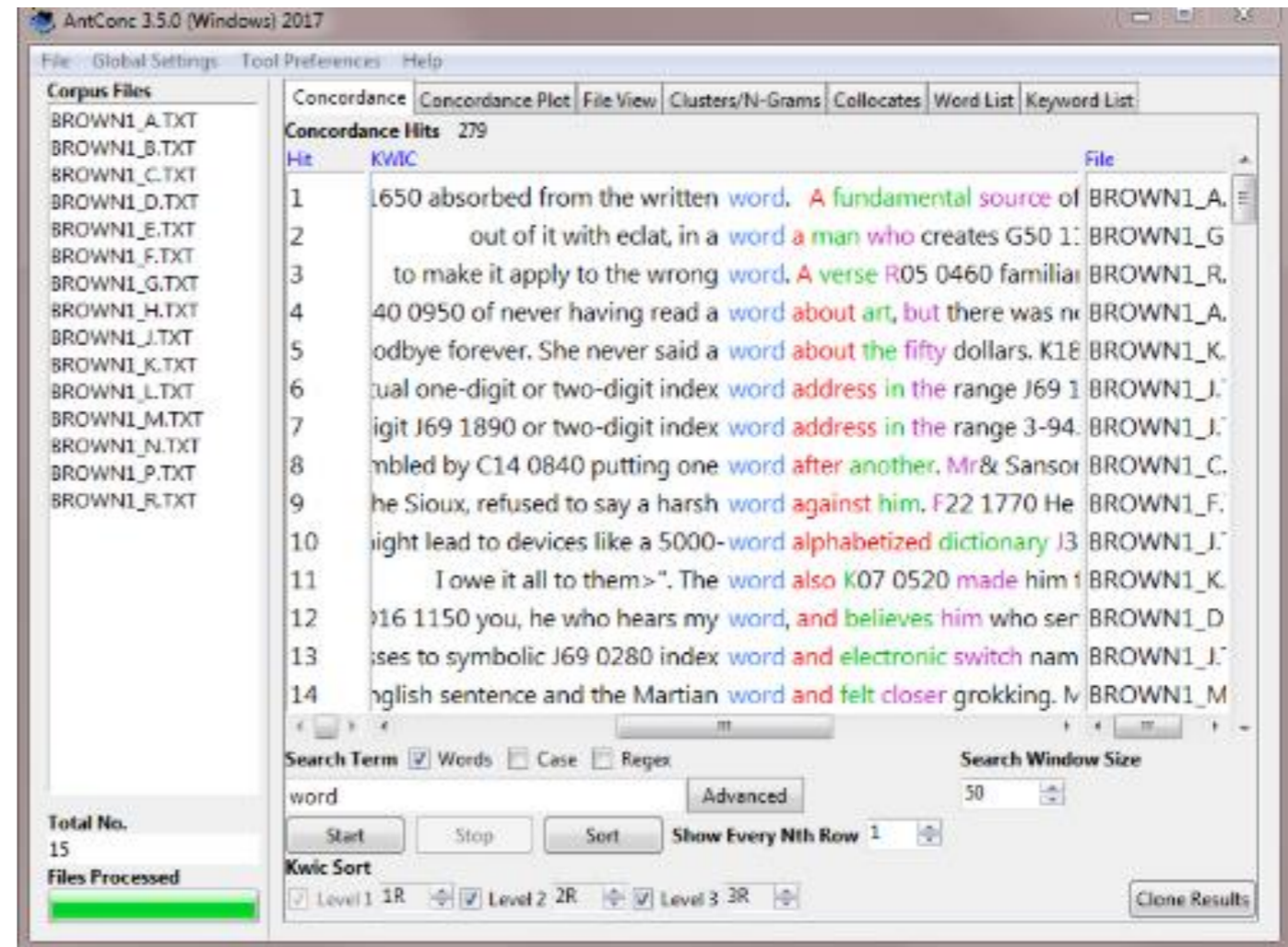
WORDSMITH

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



ANTCONC

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





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



ANTCONC: FUNCTIONS

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



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?

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)

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.



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?

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*?

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



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?



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>

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



DDL- DATA DRIVEN LEARNING

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



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