УДК 001.891, 81.139, 81-112
DOI: 10.17223/19996195/52/3

КОРПУСНЫЕ ИССЛЕДОВАНИЯ С ПОМОЩЬЮ СЕРВИСА ANTCONC В УСЛОВИЯХ РАБОТЫ В ВУЗЕ

И.А. Котюрова

Аннотация. Описывается один из инструментов корпусных исследований – программа AntConc, которая успешно может быть использована как в учебной, так и в научно-исследовательской работе студентов. Перечислены основные принципы работы с программой, ее функциональные возможности, при этом для лучшего понимания описание подкрепляется скриншотами окна программы. Затем приводится ряд конкретных заданий, которые могут быть предложены студентам в курсе истории немецкого языка как на очном практическом занятии, так и в режиме дистанционной работы. Пример способов решений этих заданий со скриншотами демонстрирует возможности работы с программой и призван также мотивировать к составлению любых других подобных заданий. Приводятся задания, наглядно демонстрирующие лексические, грамматические, фонетические и синтаксические особенности древневерхненемецкого языка. Поскольку программа позволяет работать с документами на любом языке, данные задания могут быть адаптированы и к другим языкам. Делается анализ преимуществ AntConc для использования в вузе. Во-первых, он бесплатный и не требует регистрации, что делает сервис доступным как для работы непосредственно на практическом занятии, так и в режиме дистанционной / самостоятельной работы. Во-вторых, программа AntConc имеет интуитивно понятный интерфейс и очень проста в использовании. Это позволяет в ходе даже одного занятия провести небольшое исследование по изучаемой теме и научить студентов проводить собственные исследования с помощью AntConc. В-третьих, данная программа дает возможность работать с любым текстовым файлом txt любого объема на любом языке, что позволяет выполнить статистический анализ практически любого материала. В-четвертых, сервис дает возможность анализировать несколько файлов или целиком, как единый корпус, или параллельно, как отдельные файлы, в результате чего можно провести сравнительный анализ количественных показателей в разных текстах. Все это делает сервис AntConc почти идеальным инструментом для исследовательской работы студентов.
Ключевые слова: корпусные исследования; AntConc; история немецкого языка.

Введение

Корпусная лингвистика – очень динамично развивающееся направление, за которым стоит будущее. Значимость корпусных исследований для современной лингвистики сегодня не ставится под сомнение (см., в частности, [1]). В работах современных отечественных и зарубежных лингвистов анализ как отдельных языковых единиц, так и

целых дискурсов, подтверждаемый корпусными данными, встречается в последнее время все чаще [1–7]. Очевидно, что в ближайшие годы лекции и практические занятия по корпусной лингвистике станут обязательными при обучении бакалавров различных направлений и профилей подготовки. Все больше появляется статей по теме включения корпусных технологий в обучение [8–10].

Первые массивы репрезентативных аннотированных текстов появились в США, поэтому самые обширные и наиболее функциональные корпусы представлены англоязычными материалами. Очень хорошие по качеству, т.е. по объему и репрезентативности языка, корпусы есть и в других языках. Например, для русского языка это НКРЯ, а для немецкого – Cosmas II. На сегодня разработанный и развиваемый в Мангеймском университете корпус Cosmas II заслуженно считается наиболее полным и по объему, и по функциям.

Однако работа с Cosmas II со студентами на занятиях в российский вузах затрудняется тем, что программа эта насколько богата в своих возможностях, настолько и трудна в пользовании. Требуется долгое погружение в то, как устроен проект и каким образом нужно строить запросы поиска. Это безусловно важно и нужно для специалистов, занимающихся лингвистическими исследованиями немецкого языка. Но в условиях подготовки бакалавров и очень ограниченного времени на знакомство обучающихся с современными технологиями исследований, в частности с возможностями корпусной лингвистики, в некоторых случаях более подходящим, на наш взгляд, может оказаться сервис AntConc.

Antconc – это бесплатная, мультиплатформенная программа, представляющая собой инструмент для статистических исследований текстов. Она была разработана профессором Лоуренсом Антони (Laurence Anthony), директором Центра обучения английскому языку в науке и технике Школы науки и техники университета Васеда (Япония). Для масштабных глубоких исследований немецкого языка эта программа вряд ли составит конкуренцию Cosmas II. Но для лингвистов, только начинающих свое знакомство с корпусными технологиями, Antconc – один из самых популярных на сегодняшний день сервисов.

Методология

Кратко опишем основные принципы работы с AntConc, приведем конкретные примеры использования ее функций и сделаем выводы относительно целесообразности применения сервиса в условиях высшей школы.

Итак, AntConc запускается открытием файла .exe, загружаемого с официального сайта разработчика программы Лоуренса Антони: http://www.antlab.sci.waseda.ac.jp/software.html [11]. Открывшееся окно программы содержит семь вкладок, соответствующих семи инструмен-

там анализа, которые могут быть актуализированы как кликом на ту или иную вкладку, так и функциональными клавишами от F1 до F7.

1. Concordance – конкорданс, инструмент, позволяющий найти все контексты слова или словосочетания в указанном тексте (KWIC – Key Words in Context) (рис. 1).

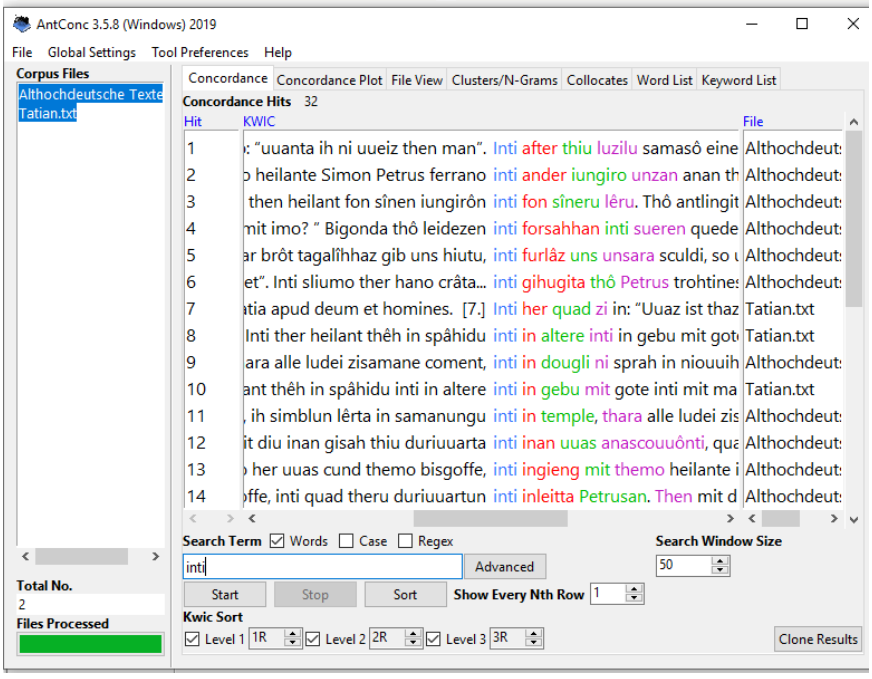

Рис. 1. Конкорданс для слова inti в текстах древневерхненемецкого языка

2. Concordance plot отображает наличие исследуемых слов или словосочетаний в тексте в виде штрих-кода, что делает возможным визуально оценить, как часто в какой части текста встречается искомый объект. Например, на рис. 2 показано, где в загруженном файле встречается слово inti. Отчетливо видно, что inti используется только в текстах одного источника (файл Althochdeutsche Texte содержит несколько разных текстов VIII–X вв.). В других источниках это слово будет выглядеть как enti, endi.

3. File View отображает текст выбранного файла в начальном виде. При этом цветом маркируются элементы, указанные в поле поиска (рис. 3).

4. Функция Words Clusters представляет собой инструмент отбора группы слов с заданным количеством элементов слева и справа от заданного слова.

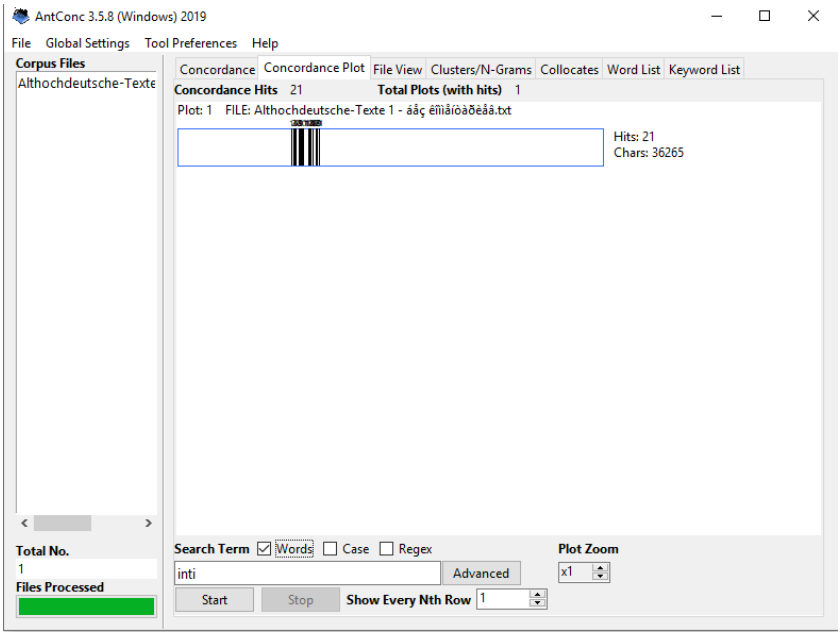

Рис. 2. Конкорданс в виде штрих-кода для слова inti
в текстах древневерхненемецкого языка

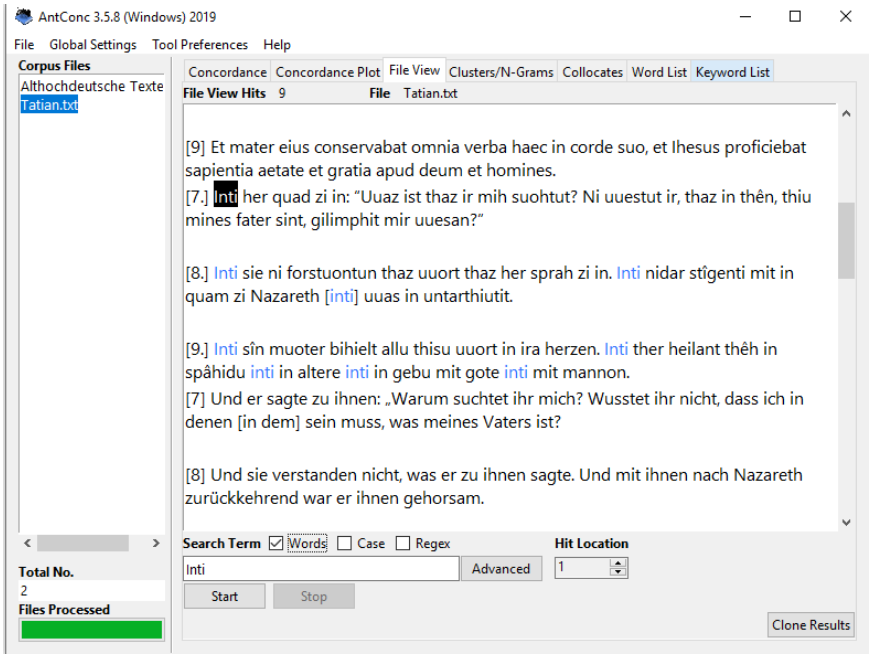

Рис. 3. Отображение искомого элемента inti в исходном файле

Это бывает нужно, например, чтобы проверить, с артиклем какого падежа используется тот или иной предлог (например, предлог trotz) в текстах разных стилей или разных периодов развития языка (рис. 4). Сортировка при этом может быть как по количеству, так и по первой или последней букве в кластере, а также по степени вероятности, что первое слово в кластере предшествует остальным. Для этого инструмента есть дополнительные опции, подробно о которых можно прочитать в инструкции разработчика на сайте [11].

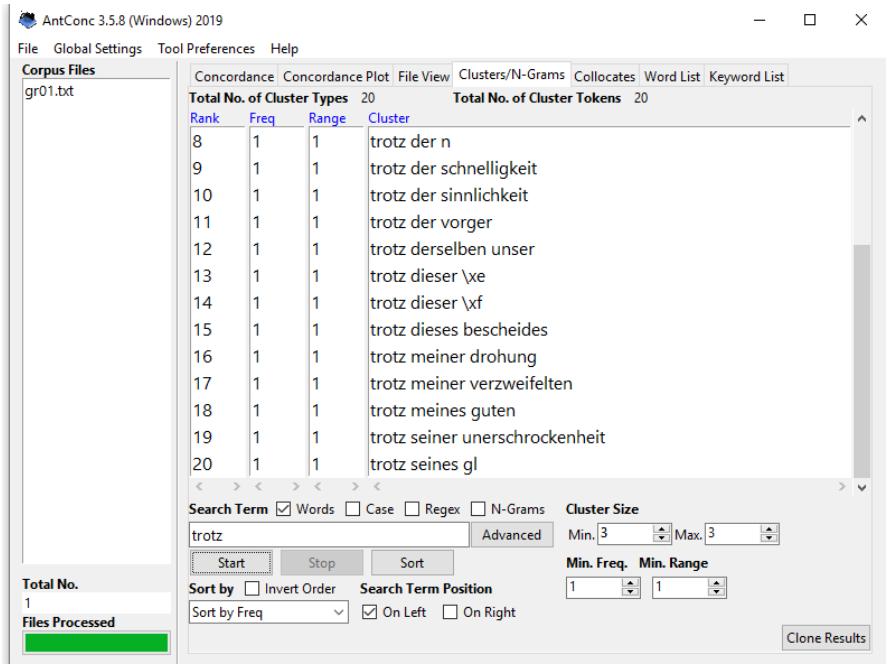

Рис. 4. Отбор групп слов с тремя элементами, начинающихся с предлога trotz, в романе К. Майя «Через пустыню»

Функция N-Grams также связана с поиском в загруженных файлах групп слов (кластеров) заданной длины, но в отличие от Clusters речь идет не о поиске групп слов вокруг ключевого элемента, а о любых сочетаниях стоящих рядом слов. Это позволяет найти наиболее распространенные в тексте словосочетания. Например, в предложении Das ist meine Pflicht программа выдаст три кластера с двумя элементами: Das ist, ist meine, meine Pflicht. Так, на рис. 5 видно, что наиболее употребимым сочетанием из трех слов в романе В. Херндорфа «Чик» является выражение die ganze Zeit, встретившееся в тексте 29 раз.

Для того чтобы переключиться с функции Words Clusters на N-Grams, нужно, открыв вкладку Clusters / N-Grams, выбрать ниже поля результатов в условиях поиска Words или N-Grams соответственно.

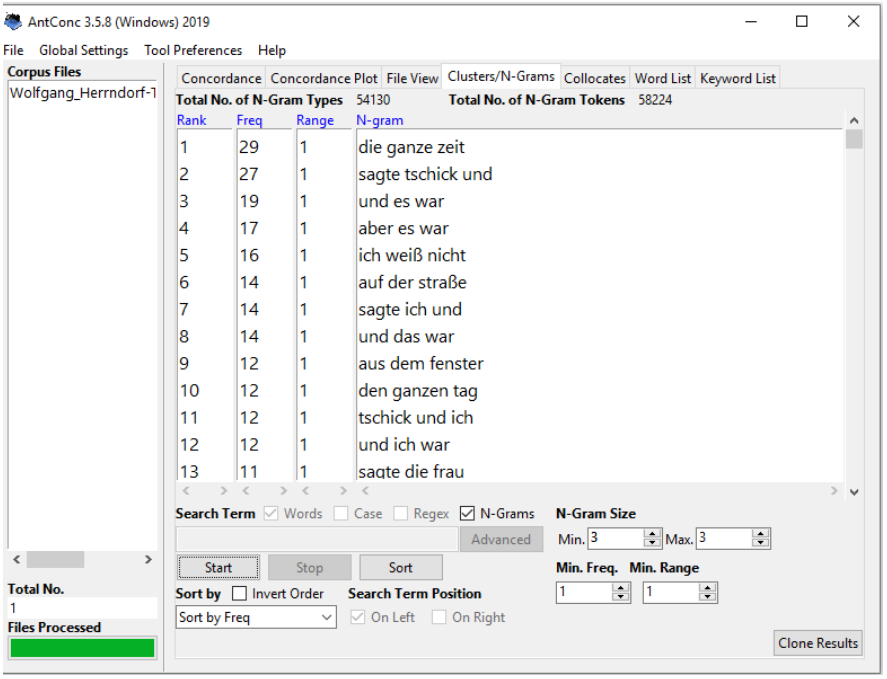

Рис. 5. Результат поиска кластеров из трех элементов в романе В. Херндорфа «Чик»

5. Функция Collocates дает возможность сделать мгновенный статистический анализ по словам, стоящим слева и справа от искомого элемента. Так, чтобы посмотреть, с какими словами используются например, словоформы существительного Gott в романе В. Херндорфа «Чик», достаточно вписать это слово в строку поиска, выставить параметр количества элементов слева и справа, а также способ представления результата: по общей частотности употребления, по частотности употребления слов слева или справа, по алфавиту или по конечной букве в слове. На рис. 6 показано, что статистически после слова zweifeln вероятнее других слов будет стоять gott, в отличие, например, от слова mein, которое хоть и встречается со словом Gott чаще, чем zweifeln, но сила его коллокации со словом Gott значительно меньше, поскольу mein в исследуемом тексте встречается и с другими словами, а не только с Gott. Также видно, что наиболее частотным коллокациями слова gott в диапазоне четырех слов слева и справа являются элементы mein и o.

6. Word List – инструмент для подсчета и представления всех словоупотреблений, встречающихся в корпусе, в виде упорядоченного списка. Другими словами, это инструмент для составления частотного словаря конкретного корпуса (рис. 7).

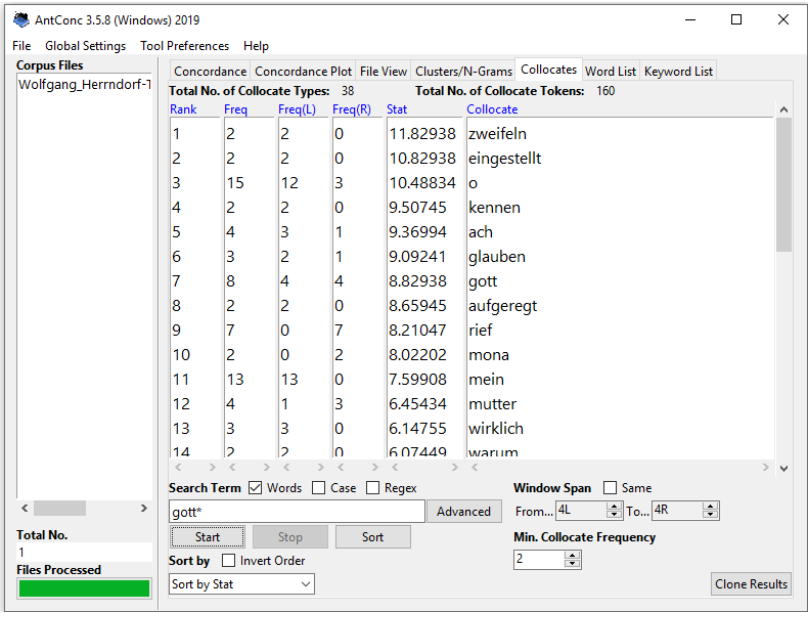

Рис. 6. Результат поиска коллокаций словоформ gott* в романе В. Херндорфа «Чик», где сортировка результатов представлена по силе коллокации

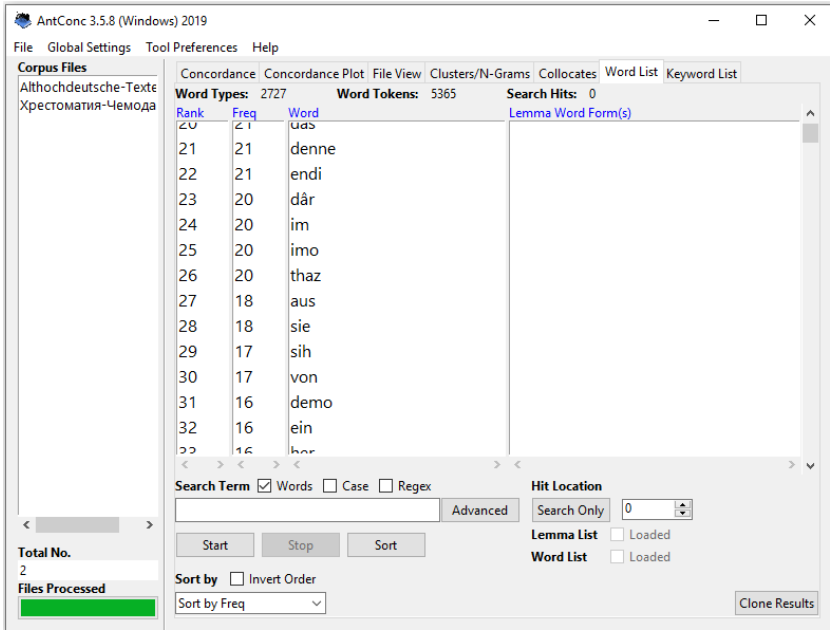

Рис. 7. Перечень всех словоформ, встретившихся в загруженных для анализа древненемецких текстах

7. Key Word List – инструмент, рассчитывающий, какие слова в корпусе являются необычно высокочастотными или необычно низкочастотными по сравнению с эталонным корпусом. Например, если в качестве эталона использовать текст классического художественного произведения, то в исследуемом тексте газеты или политической речи необычно высокочастотные единицы могут быть охарактеризованы как маркеры публичного стиля соответствующего жанра (разумеется, для определения таких характерных особенностей потребуется целый ряд подобных сравнительных исследований).

Приведем примеры заданий, которые можно предложить обучающимся в качестве пробной работы с сервисом AntConc. Эти примеры могут быть использованы на занятии по истории немецкого языка или по современным технологиям научных исследований. С помощью приложения AntConc проследим некоторые характеристики древневерхненемецкого языка.

Древневерхненемецкие тексты VIII–X в. можно взять в открытом доступе на различных отечественных и зарубежных порталах, например на портале http://www.mediaevum.de/texte/ahd.htm. Необходимо заранее подготовить файл для работы с приложением (или файлы, если планируется сопоставление статистических данных по разным текстам этого периода).

Большинство общедоступных материалов в интернете представлены в формате pdf или html. С вышеуказанного сайта можно скопировать материалы непосредственно в Word, а затем сохранить файл в формате txt, используя кодировку Unicode 8. Файлы в формате pdf требуют предварительной конвертации в txt.

Исследование и результаты

Студенты загружают подготовленный заранее файл / файлы, используя функцию Open File, и выполняют на его основе задания, предложенные ниже.

Задание 1. Определите по два наиболее часто встречающихся в текстах древневерхнего периода имени существительных, местоимения, глагола и служебных слова. Прокомментируйте полученные данные.

Решение: задание выполняется с помощью функции Word List. Полученные данные служат основанием для дискуссии (существительные: man, gottes; местоимения: er (her), ih; глаголы: ist (is), scul; служебные слова: in, so) (рис. 7).

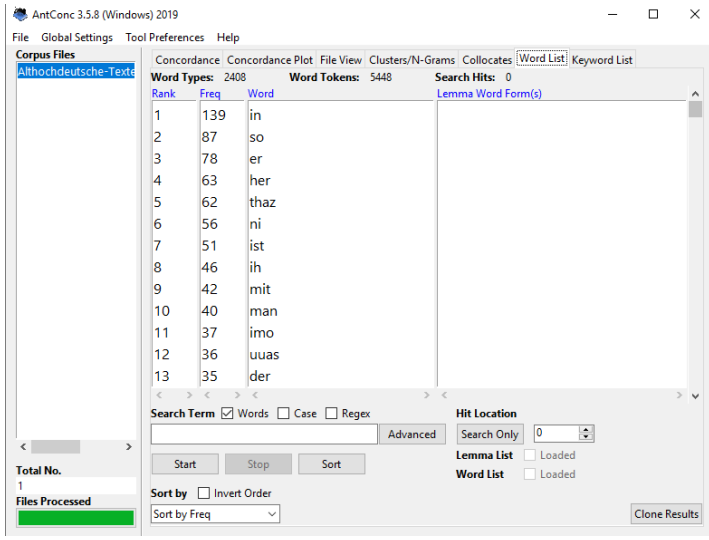

Рис. 7. Пример полученных данных по поиску наиболее частотных слов в древневерхненемецких текстах

Задание 2. Определите, является ли ein в древневерхненемецких текстах числительным или артиклем.

Решение: задание выполняется с помощью функции Concordance. Вводим в качестве искомого элемента ein* (звездочка (*) делает возможным поиск других форм этого слова) и определяем по контексту однозначность отнесения слова к той или иной части речи (рис. 8).

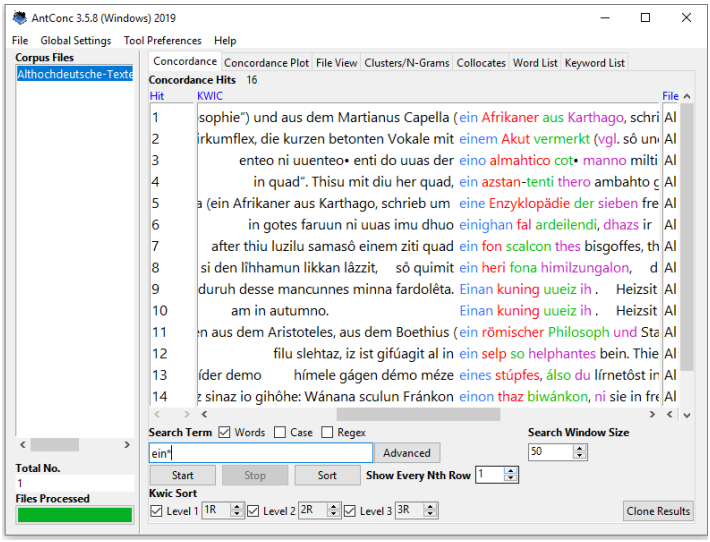

Рис. 8. Пример полученных данных по поиску контекстов употребления ein* в древневерхненемецких текстах

Задание 3. Найдите в текстах древневерхненемецкого периода формы генитива и дайте количественную и качественную оценку этих форм.

Решение: задание выполняется с помощью функций Concordance и Concordance Plot (рис. 9). Для количественной оценки необходимо подготовить еще два файла с современными текстами такого же объема – один религиозной тематики, а другой – содержащий разговорную речь. В поиск по древним текстам вводится форма thes, с помощью Concordance определяется общее количество использований, а с помощью Concordance Plot – равномерность употребления форм генитива по разным текстам этого периода. То же самое проводится с современными текстами и формой des. Затем делается сравнительный анализ количественных показателей и равномерности распределения форм по тексту. Для представления результатов полезной является возможность «клонировать результат» с помощью кнопки в правом нижнем углу Clone Results (в нашем примере в древних текстах объемом 35 500 знаков встретилось 24 формы генитива thes, в современных религиозных проповедях того же объема – 38 форм, а в отрывке из романа «Чик» – всего 3 формы).

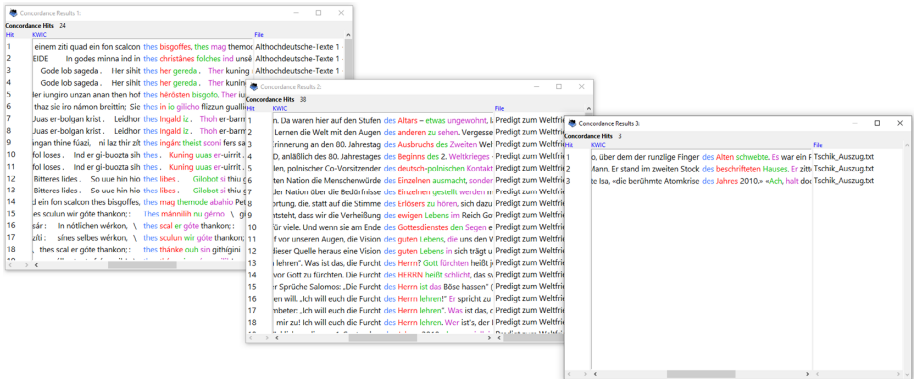

Рис. 9. Пример полученных данных по поиску генитивных форм
в древневерхненемецких и современных текстах

Для качественной оценки необходимо определить функции генитива в каждом конкретном употреблении, посмотрев в Concordance контекст и осуществив перевод релевантных форм в контексте.

Задание 4. Определите наиболее употребимые выражения в древневерхненемецких текстах и дайте свой комментарий полученному результату.

Решение: задание выполняется с помощью функции Cluster / N-Grams. Возможно варьировать количество элементов в кластере. Наиболее целесообразным представляется выставление параметров количества элементов от 2 до 4 (рис. 10).

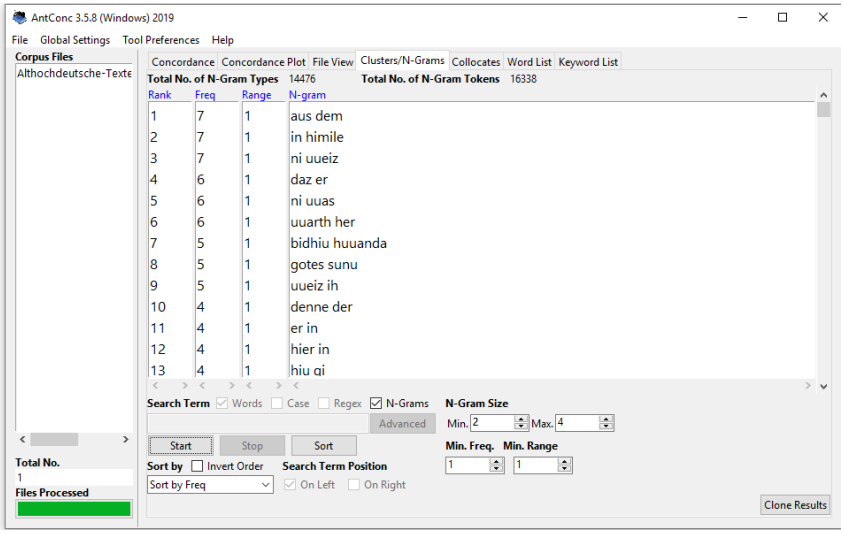

Рис. 10. Пример полученных данных по поиску наиболее частотных коллокаций слов в древневерхненемецких текстах

Задание 5. Найдите все формы глагола stantan и определите их функцию в каждом конкретном случае.

Решение: чтобы найти разные формы одного слова, в котором могут меняться как конечные звуки, так и корневые гласные и согласные, необходимо варьируемую часть слова заменить на звездочку (*). Таким образом, используем Concordance и вводим в поиск st*n* (рис. 11).

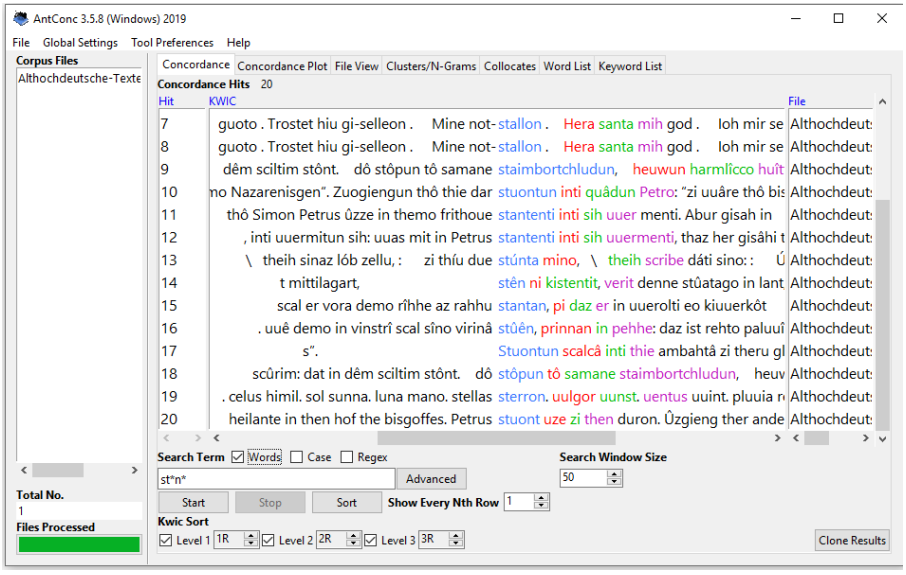

Рис. 11. Пример полученных данных по поиску форм глагола stantan в древневерхненемецких текстах

В перечне отобранных программой словоформ, очевидно, окажутся и формы других слов, которые следует отсортировать самостоятельно. В комментарии результата необходимо обратить внимание на то, какие формы используются и почему. То, что не вся парадигма спряжения представлена в текстах, является нормой (см. современные корпусные исследования на эту тему, в частности [6]).

Задание 6. Продемонстрируйте отсутствие качественной редукции конечных гласных в древневерхненемецком языке.

Решение: задав в поиск формулы «*i» «*u» «*o», можно вывести списки слов, оканчивающиеся на эти гласные, и сопоставить их с современными соответствующими словоформами (рис. 12).

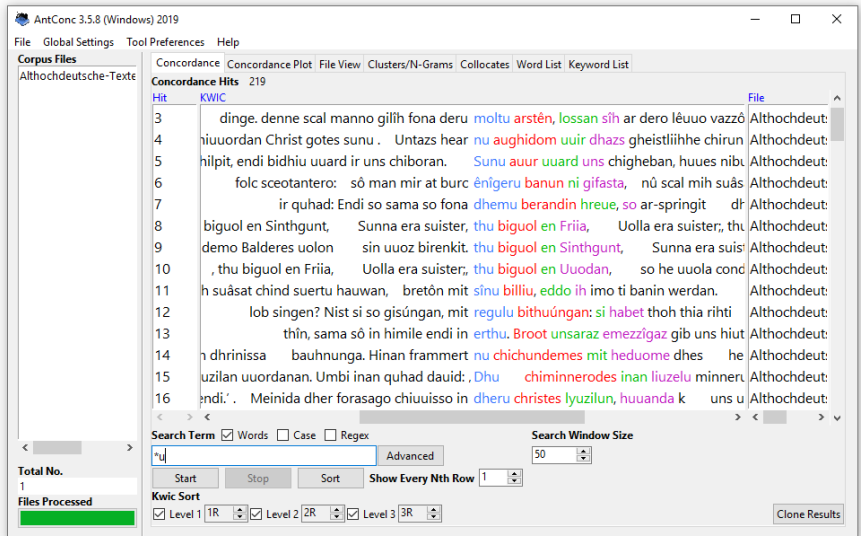

Рис. 12. Пример полученных данных по поиску слов, оканчивающихся на -u в древневерхненемецких текстах

Заключение

Таким образом, приложение AntConc позволяет легко и быстро проводить корпусные исследования в рамках занятий в вузе – как очных, так и дистанционных. Для работы в условиях высшей школы у этого ресурса есть ряд преимуществ.

Во-первых, он бесплатный и не требует регистрации, что крайне важно для вузов, работающих только с официально приобретенным лицензионным продуктом.

Во-вторых, эта программа очень проста и понятна в использовании, чего нельзя сказать про Cosmas II – программы более функциональной, но требующей длительного обучения по работе с ней. Для то-

го, чтобы начать работать с Antconc, не требуется долгого изучения интерфейса программы, формул запросов и т.п. Это позволяет в ходе даже одного занятия в вузе провести небольшое исследование, которое, с одной стороны, продемонстрирует принципы работы с корпусами, в частности создание конкорданса, с другой – создаст у обучающихся ситуацию успеха, мотивирующую на дальнейшие собственные исследования.

В-третьих, данная программа позволяет работать с любым текстовым файлом txt, а не только с аннотированными текстами, включенными в традиционные корпусы, такие как Cosmas II, DWDS, DDD, Zeitungstextkorpora и др. Корпус для анализа с помощью AntConc составляется самим пользователем и может быть любого объема, на любом языке и включать любое количество файлов, предварительно собранных в формате txt в одну общую папку. Это оказывается очень удобным, так как дает возможность выполнить статистический анализ практически любого материала, как цифрового (т.е. файлов с текстом), так и не цифрового (т.е. тексты на бумаге). Конечно, во втором случае текст нужно будет сначала оцифровать (например, отсканировать и распознать).

В-четвертых, сервис позволяет анализировать несколько файлов или целиком, как единый копус, или параллельно, как отдельные файлы. Это делает возможным сравнительный анализ количественных показателей в разных текстах.

Резюмируя, следует сказать, что сервис AntConc может быть для начинающего исследователя-лингвиста очень полезным и несложным в использовании инструментом, легко и быстро осуществляющим статистический анализ текстового материала. Однако интерпретация полученных с его помощью результатов все равно остается за человеком.

Литература

1. ***Плунгян В.А.*** Почему современная лингвистика должна быть лингвистикой корпусов. URL: https://polit.ru/article/2009/10/23/corpus/ (дата обращения: 09.08.2019).
2. ***Баранов М.М., Вознесенская А.Н., Добровольский О.Г., Киселева К.Л., Козеренко А.Д.*** Корпусное обеспечение исследований в области фразеологии и фразеографии // Русская лексикография XXI века: проблемы и способы их решения. М. ; СПб. : Нестор-История, 2016. С. 14–15.
3. ***Баркович А.А.*** Корпусная лингвистика: специфика современных метаописаний языка // Вестник Томского государственного университета. 2016. № 406. С. 5–13.
4. ***Laura A. Janda.*** Aspectual clusters of Russian verbs // Studies in Language. 2007. Vol. 31 (3). P. 607–648.
5. ***Lüdeling A., Walter M.*** Korpuslinguistik // Handbuch Deutsch als Fremd- und Zweitsprache. HSK 35. De Gruyter, 2010. S. 315–322.
6. ***Wallner F.*** Korpora im DaF-Unterricht – Potentiale und Perspektiven am Beispiel des DWDS. Revista Nebrija de Lingüística Aplicada 13, Nr. número especial – Actas de Congreso (2013).
7. ***Dobrovolsky O.*** Constructions in Parallel Corpora: a Quantitative Approach // Computational and Corpus-Based Phraseology (Second International Conference,

Europhras 2017. London, UK, November 13–14, 2017. Proceedings) / ed. by Ruslan Mitkov. Berlin : Springer, 2017. P. 41–53.
8. ***Горина О.А.*** Использование технологий корпусной лингвистики для развития лексических навыков студентов-регионоведов в профессионально-ориентированном общении на английском языке : дис. ... канд. пед. наук. М., 2014. 321 с
9. ***Станкевич А.Ю.*** Поиск контекстов и оценка их типичности средствами AntConc (Laurence Anthony) // Теория и практика преподавания русского языка как иностранного: достижения, проблемы и перспективы развития : материалы V Междунар. науч.-метод. конф. / редкол.: С.И. Лебединский (гл. ред.) и др. Минск : Изд. Центр БГУ, 2011. С. 210–213.
10. ***Ahrenholz B., Wallner F.*** Digitale Korpora und Deutsch als Fremdsprache // Deutsch als Fremdsprache (Deutschunterricht in Theorie und Praxis / hrsg. Bernt Ahrenholz, Ingelore Oomen-Welke. Schneider Verlag Hohengehren, 2013. Bd. 10. S. 261–272.
11. ***Страница*** разработчика приложения AntConc Атонио Лоуренса. URL: http://www.antlab.sci.waseda.ac.jp/index.html (дата обращения: 09.08.2019).

Сведения об авторе:
Котюрова Ирина Аврамовна – кандидат филологических наук, доцент, Петрозаводский государственный университет (Петрозаводск, Россия). E-mail: koturova@petrsu.ru

Поступила в редакцию 28 октября 2020 г.

CORPUS-BASED STUDIES WITH ANTCONC SERVICE AT THE UNIVERSITY

Kotiurova I.A., Ph.D. (Philology), Associate Professor, Petrozavodsk State University (Petrozavodsk, Russia). E-mail: koturova@petrsu.ru

DOI: 10.17223/19996195/52/3

Abstract. The article describes one of the tools of corpus-based studies – the AntConc program, which can be successfully used in both educational and research work of students. The article describes the basic principles of working with the program, its functionality. For a better understanding the description is supported by screenshots of the program window. Then, a number of specific tasks are given that can be offered to students in the course of the history of the German language both in full-time practical training and in remote work mode. An example of ways to solve these tasks with screenshots demonstrates the possibilities of working with the program and is also intended to motivate to draw up any other similar tasks. The article provides tasks that clearly demonstrate the lexical, grammatical, phonetic and syntactic features of the Old High German language. Since the program allows you to work with documents in any language, these tasks can be adapted to other languages. In conclusion, an analysis of the attractions of AntConc for use in the university is made. Firstly, it is free and does not require registration, which makes the service available both for working directly in a practical lesson and in remote / independent work mode. Secondly, AntConc has an intuitive interface and is very easy to use. This allows even a single lesson at the university to conduct a small study on the topic being studied and teach students to conduct their own research using AntConc. Thirdly, this program allows you to work with any txt file of any size, in any language, which allows you to perform statistical analysis of almost any material. Fourth, the service allows you to analyze several files either as a whole corpus, or as separate files, which allows a comparative analysis of quantitative indicators in different texts. All this makes the AntConc service an almost perfect tool for student research.
Keywords: corpus studies; AntConc; history of the German language.

References

1. Plungyan V.A. Why modern linguistics should be linguistics of corps. Access Mode: https://polit.ru/article/2009/10/23/corpus/ (Date of access 09.08.2019).
2. Baranov MM, Voznesenskaya AN, Dobrovolsky OG, Kiseleva KL, Kozerenko AD. Case support for research in the field of phraseology and phraseography // Russian lexicography of the XXI century: problems and methods their decisions. M., St. Petersburg: Nestor-History 2016, p. 14–15.
3. Barkovich A.A. Corpus linguistics: the specifics of modern meta-descriptions of language // Bulletin of Tomsk State University. 2016. No 406. P. 5-13.
4. Laura a. Janda Aspectual clusters of Russian verbs / Studies in Language, Vol. 31: 3. 2007 Pp. 607–648.
5. Lüdeling, Anke; Walter, Maik: Korpuslinguistik. In: Hans-Jürgen Krumm et al. (Hrsg.): Handbuch Deutsch als Fremd- und Zweitsprache, HSK 35, S. 315-322, De Gruyter, 2010.
6. Wallner, Franziska: Korpora im DaF-Unterricht – Potentiale und Perspektiven am Beispiel des DWDS. Revista Nebrija de Lingüística Aplicada 13, Nr. número especial – Actas de Congreso (2013).
7. Dobrovolsky O. Constructions in Parallel Corpora: a Quantitative Approach // Computational and Corpus-Based Phraseology (Second International Conference, Europhras 2017. London, UK, November 13-14, 2017. Proceedings) / Ruslan Mitkov (Ed.). Berlin etc.: Springer, 2017. Pp. 41–53.
8. Gorina O.A. The use of corpus linguistics technologies for the development of lexical skills of regional students in professionally oriented communication in English: dis. ... cand. teacher. sciences. M., 2014. 332 p.
9. Stankevich, A.Yu. Search for contexts and assessment of their typicality by means of AntConc (Laurence Anthony) / A.Yu. Stankevich // Theory and practice of teaching Russian as a foreign language: achievements, problems and development prospects: materials of the V Intern. scientific method. conf. Minsk / Editorial: S.I. Lebedinsky, (Ch. Ed.) [et al.]. Minsk: Publishing House. BSU Center, 2011. Pp. 210–213.
10. Ahrenholz, Bernt; Wallner, Franziska: Digitale Korpora und Deutsch als Fremdsprache. In: Bernt Ahrenholz, Ingelore Oomen-Welke (Hrsg.): Deutsch als Fremdsprache (Deutschunterricht in Theorie und Praxis, Bd. 10), Pp. 261–272. Schneider Verlag Hohengehren, 2013.
11. URL: http://www.antlab.sci.waseda.ac.jp/index.html (access date: 09.08.2019).

Received 28 October 2020