

Ольга Юрьевна САВИНА¹

УДК 81'33

МЕТОДИКА ФОРМИРОВАНИЯ ЛЕКСИЧЕСКОГО МИНИМУМА С ПОМОЩЬЮ КОНКОРДАНСЕРА

¹ старший преподаватель кафедры немецкой филологии
Института филологии и журналистики
Тюменского государственного университета
osawina@gmail.com

Аннотация

В статье рассматривается дидактический потенциал современной корпусной лингвистики, в частности, использование корпуса как источника данных для лингводидактических материалов. Автором сформулированы этапы создания мини-корпуса немецкоязычных публицистических текстов с помощью специализированной поисковой машины Yomunda. Представлена методика работы с конкордансером, включающая предварительную обработку текстов, их конвертацию, составление частотного списка лексики и лемматизацию словоформ, вошедших в частотный список. Описан опыт формирования тематического лексического минимума на базе корпуса аутентичных текстов и его верификации с помощью базы данных Profile Deutsch.

Проведенный анализ показал, что предлагаемый алгоритм создания лексического минимума позволяет эффективно отбирать наиболее употребительные лексемы из актуальных аутентичных текстов. Достоинством создаваемого таким образом лексического минимума является его словообразовательное богатство. Представленный алгоритм позволяет формировать лексический минимум для любой целевой группы.

Ключевые слова

Лексический минимум, лингводидактический мини-корпус, лингвистическая поисковая машина Yomunda, конкордансер AntConc, база данных Profile Deutsch.

Цитирование: Савина О. Ю. Методика формирования лексического минимума с помощью конкордансера / О. Ю. Савина // Вестник Тюменского государственного университета. Гуманитарные исследования. Humanitates. 2016. Том 2, № 1. С. 92-99.
DOI: 10.21684/2411-197X-2016-2-1-92-99

DOI: 10.21684/2411-197X-2016-2-1-92-99

Данная статья представляет опыт создания тематически ориентированного лексического минимума на основе корпуса аутентичных текстов. Проведенное исследование показало, что оптимальным инструментом для формирования подобного минимума может стать конкордансер, а материалом — мини-корпус актуальных публицистических текстов.

Очевидно, что корпус может функционировать как источник статистических данных не только для лингвистических, но и для лингводидактических исследований. Количественные данные о языке позволяют также формировать содержание обучения — список наиболее частотной лексики с данными о вероятности ее употребления. При этом для отбора актуальной лексики необходимо использовать актуальные тексты.

Новые дидактические возможности лингвистических корпусов эффективно внедряются в систему обучения иностранным языкам. Различные стороны такого подхода неоднократно описаны (К. В. Александров [1], М. С. Мальцева [4], Л. К. Раицкая [5], Е. П. Соснина [6], П. В. Сысоев [7] и др.). Возможность самостоятельной, но эффективной разработки дидактических материалов отмечает Л. И. Раицкая, предлагая составлять упражнения на основе т. н. мини-конкордансов, т. е. «на основании одного или нескольких учебных текстов или текстов, используемых в учебных пособиях» [5, с. 126].

Мини-корпусом может стать любой представительный массив текстов. Однако для получения качественных лингвистических данных, отражающих реальную ситуацию словоупотребления, необходимо иметь собрание аутентичных иноязычных текстов, отвечающих определенному уровню владения иностранным языком. Именно аутентичные тексты иллюстрируют функционирование языка в форме, принятой носителями языка [3]. При этом под аутентичным текстом следует понимать «текст, являющийся реальным продуктом речевой деятельности носителей языка и не адаптированный для нужд учащихся с учетом их уровня владения языком» [9, с. 41].

Базовая технология создания корпуса была предложена В. П. Захаровым [2, с. 7-8]. Основываясь на ней, авторы предлагают собственную методику формирования лексического минимума (списка слов и выражений в соответствии с требованиями к уровню владения языком, по определенной теме, «которые должны быть усвоены учащимися за определенный промежуток учебного времени» [9, с. 134]), включающую шесть этапов:

- 1) выборка текстов с учетом тематики и уровня сложности;
- 2) предварительная обработка текстов — коррективная, графематический анализ, при необходимости — экстралингвистическое описание;
- 3) конвертация текстов (согласно требованиям корпусного менеджера) и ввод корпуса в программу;
- 4) составление частотного списка лексики;

- 5) исключение стоп-слов (не релевантных для целевой темы);
- 6) лемматизация словоформ, вошедших в частотный список.

Представим детальное описание каждого этапа.

- 1) По мнению Л. К. Раицкой, лучшим корпусом являются газетные тексты. Автор подчеркивает, что при создании собственного корпуса необходимо учитывать тематику и уровень сложности текстов для целевой группы [5, с. 125]. Для немецкого языка в качестве инструмента выборки газетных текстов с учетом темы и уровня сложности можно использовать специализированную лингвистическую поисковую машину Yomunda [10]. Она представляет собой онлайн-сервис, ведущий поиск по 15 электронным СМИ Германии и Швейцарии. Положенные в основу сервиса авторские алгоритмы, разработанные сотрудниками semtracks gmbh (Цюрих, Фрайбург, Токио), позволяют вести отбор по дате, теме и уровню сложности (текста в целом, лексического наполнения, грамматики).
- 2) Предварительная обработка текстов подразумевает в т. ч. удаление нетекстовых элементов.
- 3) Для работы с корпусом (поиска словоформ/лемм, составления конкорданса, получения различных статистических данных и т. д.) могут использоваться различные корпусные менеджеры (специализированные поисковые системы для конкретного корпуса) или конкордансеры [8, с. 62]. С помощью конкордансеров, в частности, легко решаются узкоспециальные лингводидактические задачи, например, отбор примеров грамматических форм слова или лексической сочетаемости [6]. П. В. Сысоев также отмечает потенциал использования корпусов и конкордансеров для формирования именно лексико-грамматических навыков обучающихся [7, с. 104].
- 4) Один из хорошо зарекомендовавших себя конкордансеров — программа AntConc [11] — может быть использован для вычисления соотношения количества словоформ и словоупотреблений (types и tokens) и относительной частоты различных слов, а также для выявления типичных контекстов (коллокатов) для конкретных словоформ в данном корпусе и вычисления их частотности и др. Программа позволяет исключить определенные слова во время операции вычисления (Use a stoplist listed below).
- 5) С помощью функции «Список слов» (Word List) составляется частотный список. Полученный в результате частотный словарь имеет алфавитную (Word), частотную (Frequency) и порядковую (Rank) системы.
- 6) В связи с тем, что конкордансер AntConc не имеет функции морфологического анализа (т. е. не может различать служебные и знаменательные слова), необходимо использовать стоп-лист, включающий лексемы, не релевантные для темы (артикли, союзы, предлоги, вспомогательные глаголы и др. слова). Файл стоп-листа может использоваться при работе с конкордансером многократно.

- 7) Словоформы, вошедшие в частотный список, лемматизируются либо вручную, либо с помощью программы-лемматизатора, например, Morphu (свободно распространяемый программный пакет для морфологического анализа и синтеза, а также для контекстного определения частей речи в немецком языке [14]).

Представленная выше методика составления лексического минимума была апробирована на материале «Охрана окружающей среды» (Umweltschutz) для порогового уровня B1 [12]. К данной целевой группе относятся школьники, обучающиеся в образовательных учреждениях с углубленным изучением языков или готовящиеся сдать ЕГЭ, учащиеся лингвистических гимназий, студенты, готовящиеся к стипендиальным программам DAAD.

С помощью лингвистической поисковой машины Yomunda были отобраны тексты 2009-2012 гг. следующих немецкоязычных электронных СМИ: “Magazin Deutschland”, “Spiegel Online”, “Süddeutsche Zeitung”, “20 Minuten Online”, “Neue Zürcher Zeitung” и “Tages-Anzeiger”. На момент исследования (март 2012) поисковая машина Yomunda предлагала 50 текстов, соответствующих заданной теме (Umweltschutz) и пороговому уровню B1. В ходе анализа были отсеяны 32 текста, лишь косвенно затрагивающие вопросы охраны окружающей среды. В результате в исследовательский корпус были включены 18 текстов, содержащих полезный лингвистический материал для последующего формирования тематического лексического минимума.

Все отобранные тексты были преобразованы в распознаваемый программой AntConc формат .txt и прошли перекодировку в UTF8.

Сформированный нами частотный список включил 4909 словоформ и 15 949 словоупотреблений. Общее число словоформ, вошедших в стоп-лист, составило 4373. В результате был получен список из 536 словоформ и 1332 словоупотреблений.

Словоформы, вошедшие в частотный список, были лемматизированы вручную. Количество лемм составило 467. В число лемм вошли, например, следующие слова (в скобках указана частотность): Umweltschutz (52), Umwelt (45), Tonne (32), Müll (23), Energie (21), CO (20), Wasser (19), Klima (17), Greenpeace (16), Strom (15).

Для проведения верификации составленного лексического минимума в качестве эталона была использована база данных Profile Deutsch — авторитетный источник тематической лексики, отвечающий требованиям Общеввропейских компетенций обучения иностранным языкам [13]. Эталонный список лексики включил 269 слов.

Сравнительный анализ двух списков лексики (составленной по предлагаемой методике — 467 лемм и эталонной — 269 лемм) показал, что первый лексический минимум содержит практически все элементы второго, обладая при этом явным словообразовательным богатством. Продемонстрируем это на примере слов Land и Umwelt (табл. 1).

Таблица 1

Примеры лемм в составе тематического лексического минимума

Эталонная база данных Profile Deutsch	Исследовательский мини-корпус	
Land- Landschaft Landschaftsarten Landwirtschaft	Land Landesumweltministerium Landschaft Landschaftsbild Landwirtschaft	
Umwelt Umweltschutz Umweltverschmutzung	Umwelt Umweltauszeichnung Umweltbelastung Umweltbewegung umweltbewusst Umweltbewusstsein	Umweltbezug Umweltbilanz Umweltbundesamt Umweltbürokratie Umweltexperte Umweltfolge

Очевидно, что предлагаемый алгоритм формирования тематического лексического минимума позволяет вести эффективный отбор наиболее употребительных слов на основе актуальных аутентичных текстов и формировать лексический минимум для отдельных целевых групп.

СПИСОК ЛИТЕРАТУРЫ

1. Александров К. В. Компьютерная программа для формирования лексической компетенции при изучении иностранного языка / К. В. Александров // Иностранные языки в школе. 2010. № 6. С. 67-73.
2. Захаров В. П. Корпусная лингвистика: учеб.-метод. пособ. / В. П. Захаров. СПб., 2005. 48 с.
3. Киян О. Н. Аутентичность как методическая категория в обучении иностранному языку. URL: <http://main.isuct.ru/files/konf/antropos/SECTION/4/KIYAN.html>
4. Мальцева М. С. Корпусные технологии в методике преподавания иностранных языков / М. С. Мальцева // Журнал Социально-экономические явления и процессы. 2011. № 8. С. 214-218. URL: <http://cyberleninka.ru/article/n/korpusnye-tehnologii-v-metodike-gerodavaniya-inostrannyh-yazykov#ixzz3nDU9ATZE>
5. Раицкая Л. К. Дидактические возможности корпусных Интернет-технологий в преподавании иностранного языка в высшей школе / Л. К. Раицкая // Вестник Московского государственного областного университета. 2009. № 4. С. 123-127.
6. Соснина Е. П. Корпусная лингвистика и корпусный подход в обучении иностранному языку. URL: http://ling.ulstu.ru/linguistics/resourses/literature/articles/corpus_linguistics_language_teaching/

7. Сысоев П. В. Лингвистический корпус в методике обучения иностранным языкам / П. В. Сысоев // Язык и культура. 2010. № 1(9). С. 99-111.
8. Щипицина Л. Ю. Информационные технологии в лингвистике: учеб. пособ. / Л. Ю. Щипицина. М.: Флинта, 2013. 128 с.
9. Шукин А.Н. Лингводидактический энциклопедический словарь. М.: Астрель: АСТ: Хранитель, 2008. 746 с.
10. Yomunda. Поисковая машина для немецкоязычных текстов. URL: <http://www.yomunda.com/>
11. Anthony L. A freeware corpus analysis toolkit for concordancing and text analysis. URL: <http://www.laurenceanthony.net/software/antconc/>
12. Gemeinsamer europäischer Referenzrahmen für Sprachen: Lernen, lehren, beurteilen. URL: <http://www.goethe.de/z/50/commeuro>
13. Glaboniat M., Müller M., Schmitz H., Rusch P., Wertenschlag L. Profile Deutsch. Version 2.0. CD-Rom. 2005.
14. Lezius W. MORPHY für Windows. Ein Tool zur Deutschen Morphologie. URL: <http://www.wolfganglezius.de/doku.php?id=cl:morphy>

Olga Yu. SAVINA¹

**THE METHOD OF FORMING
A VOCABULARY MINIMUM
BY APPLYING CORPUS ANALYSIS TOOLKIT
FOR CONCORDANCING**

¹ Senior Lecturer, German Language Department,
Institute for Philology and Journalism,
Tyumen State University
osawina@gmail.com

Abstract

The article deals with the didactic potential of the modern corpus linguistics, in particular, the use of the corpus of texts as a source of data for didactic materials. The author formulates the steps for creating of a mini-corpus with German publicist texts applying a specialized search engine Yomunda. The method employs the work with corpus analysis toolkit for concordancing, including preprocessing of texts, their conversion, compiling a list of vocabulary frequency, and the lemmatization of types that are included in the frequency list. The article describes the experience of the compilation of thematic vocabulary minimum on the basis of the corpus of authentic texts and its verification using database Profile Deutsch. The analysis shows that the proposed algorithm for creating vocabulary minimum can effectively select the most common tokens of topical authentic texts. The advantage of this created vocabulary minimum is its derivational wealth. The algorithm allows forming vocabulary minimum for any target group.

Keywords

Vocabulary minimum, mini-corpus of texts, linguistic search engine Yomunda, corpus analysis toolkit for concordancing AntConc, database Profile Deutsch.

DOI: 10.21684/2411-197X-2016-2-1-92-99

Citation: Savina, O. Yu. 2016. "The Method of Forming a Vocabulary Minimum by Applying Corpus Analysis Toolkit for Concordancing". Tyumen State University Herald. Humanities Research, vol. 2, no. 1, pp. 92-99.

DOI: 10.21684/2411-197X-2016-2-1-92-99

REFERENCES

1. Aleksandrov, K. V. 2010. "Kompyuternaya programma dlya formirovaniya leksicheskoy kompetencii pri izuchenii inostrannogo yazyka" [A Computer Program for the Formation of Lexical Competence when Learning a Foreign Language]. *Inostrannye yazyki v shkole* [Foreign Languages at School], no. 6, pp. 67-73.
2. Maltseva, M. S. 2011. "Korpusnye tehnologii v metodike prepodavaniya inostrannyh yazykov" [Corpus Technology in Methodology of Teaching Foreign Languages]. *Zhurnal Socialno-ekonomicheskie yavleniya i processy* [Social-Economic Phenomena and Processes], no. 8, pp. 214-218. <http://cyberleninka.ru/article/n/korpusnye-tehnologii-v-metodike-prepodavaniya-inostrannyh-yazykov#ixzz3nDU9ATZE>
3. Raitskaya, L. K. 2009. "Didakticheskie vozmozhnosti korpusnyh Internet-tehnologiy v prepodavanii inostrannogo yazyka v vysshey shkole" [Didactic Possibilities of Corpus Internet Technologies in Teaching a Foreign Language in High School]. *Vestnik Moskovskogo gosudarstvennogo oblastnogo universiteta* [Moscow Regional University Herald], no. 4, pp. 123-127.
4. Sosnina, Ye. P. Korpusnaya lingvistika i korpusnyy podhod v obuchenii inostrannomu yazyku [Corpus Linguistics and Corpus Approach in Foreign Language Teaching]. http://ling.ulstu.ru/linguistics/resourses/literature/articles/corpus_linguistics_language_teaching/
5. Sysoev, P. V. 2010. "Lingvisticheskiy korpus v metodike obucheniya inostrannym yazykam" [Linguistic Corpus in the Methodology of Teaching Foreign Languages]. *Yazyk i kultura* [Language and Culture], no. 1(9), pp. 99-111.
6. Kiyani, O. N. Autentichnost kak metodicheskaya kategoriya v obuchenii inostrannomu yazyku [Authenticity as a Methodical Category in Foreign Language Teaching]. <http://main.isuct.ru/files/konf/antropos/SECTION/4/KIYAN.html>
7. Shchukin, A. N. 2008. *Lingvodidakticheskiy enciklopedicheskiy slovar* [Lingvodidactic Encyclopedic Dictionary]. Moscow: Astrel; AST; Khranitel.
8. Zaharov, V. P. 2005. *Korpusnaya lingvistika: uchebno-metod. Posobie* [Linguistic Corpus: a Teaching Method: a Manual]. St. Petersburg.
9. Yomunda. "Poiskovaya mashina dlya nemeckoyazychnyh tekstov" [Search Engine for German-Speaking Texts]. <http://www.yomunda.com/>
10. Shhipicina, L. Yu. 2013. *Informacionnye tehnologii v lingvistike. Uchebnoe posobie* [Information Technology in Linguistics. A Manual]. Moscow: Flinta.
11. Anthony, L. A Freeware Corpus Analysis Toolkit for Concordancing and Text Analysis. <http://www.laurenceanthony.net/software/antconc/>
12. Lezius, W. MORPHY für Windows. Ein Tool zur Deutschen Morphologie. <http://www.wolfganglezius.de/doku.php?id=cl:morphy>
13. Gemeinsamer Europäischer Referenzrahmen für Sprachen: Lernen, Lehren, Beurteilen. <http://www.goethe.de/z/50/commeuro>
14. Glaboniat, M., M. Müller, H. Schmitz, P. Rusch and L. Wertenschlag. 2005. *Profile Deutsch. Version 2.0.*