К проблеме создания списка высокочастотных слов и выражений немецкого языка для специальных целей

М.С. Коган[1], А.М. Ярошевич[1], А.Ю. Колотаева[1], В.П. Захаров[2], З. Шрот-Вихерт[3], А. Тильманс[3]

[1] Санкт-Петербургский политехнический университет Петра Великого
[2] Санкт-Петербургский государственный университет
[3] Ганноверский университет имени Лейбница

m_kogan@inbox.ru, amjarr@mail.ru, anna.kolotaeva@mail.ru, v.zakharov@spbu.ru, schroth-wiechert@fsz.uni-hannover.de, anna.tilmans@fsz.uni-hannover.de

Аннотация

Статья посвящена статистическому анализу специальных немецких текстов с целью выявления устойчивых сочетаний. В отличие от английского языка, на материале которого подобные исследования проводятся очень широко, списков высокочастотных слов и выражений в немецком для специальных / научных целей не существует. Исследование было проведено на материале подкорпуса «Электротехника» немецкой части корпуса текстов PhD диссертаций по техническим специальностям, являющихся основой *Kod.ING* корпус, который разрабатывается в рамках совместного проекта Ганноверского университета имени Лейбница и Санкт-петербургского политехнического университета Петра Великого.

В работе представлены списки самых частотных существительных, найденных с помощью функции WordList, и высокочастотных коллокаций, найденных с помощью функции N-Grams корпусного менеджера третьего поколения AntConc. Обсуждаются лексические и грамматические особенности полученных выражений, их сходство и отличия от подобных англоязычных списков и возможности использования полученных результатов в дидактических целях в курсе немецкого языка для специальных целей. В частности, обращается внимание на преобладание предложных сочетаний в полученном списке. Также указывается на ограничения программы AntConc, проявляющиеся при анализе относительно больших корпусов немецких текстов.

Ключевые слова: корпусная лингвистика, устойчивые сочетания, лексические пучки, списки частотных слов и выражений, английский научный дискурс, Kod.ING корпус, немецкий для специальных целей

1. Введение

Работы, посвященные количественной оценке лингвистических данных, велись давно. Еще в 1897-1898 гг. немецким лингвистом Ф. Кедингом (F. Keding) был составлен первый корпус текстов в бумажном виде объемом 11 млн. слов для сравнения частоты распределения букв в словах и выявления их сочетаемости. Особенно активно количественные методы стали развиваться с появлением компьютеров. На их основе стали создаваться частотные словари и проводиться различные исследования как теоретического, так и прикладного характера [1–5].

Следующий шаг в использовании количественных методов в лингвистике был сделан с появлением корпусов текстов. Результаты обработки запросов к корпусу всегда сопровождаются выдачей соответствующих статистических данных. В настоящее время репрезентативные корпусы служат источником для создания словарей того или иного языка.

Один из популярных предметов в корпусной лингвистике — это устойчивые сочетания. Среди них можно выделить 3 основных типа: грамматические сочетания (например, составные предлоги), семантизированные устойчивые сочетания, как свободные, так и идиоматические, и просто n-граммы.

Исследователи считают, что одним из важнейших результатов исследований в области корпусной лингвистики за последние десятилетия является вывод о том, что язык состоит не из отдельных слов, а сочетаний, регулярно встречающихся вместе и использующихся в устном и письменном дискурсе [6]. Носители языка как в устной, так и в письменной речи, тяготеют к использованию одних и тех же линейно организованных сочетаний лексических единиц. Такого рода сочетания не являются фразеологизмами, а представляют собой часто повторяющиеся цепочки слов, не всегда обладающие конструктивной законченностью, но хранящиеся в языковой памяти говорящего как некие «строительные блоки». Например, «per cent of the», «for the first time», «at the end of the», «it has been (shown / observed / argued) that». Наиболее разработана тема таких словосочетаний англоязычными исследователями. Тем не менее, в англоязычной литературе не существует общепринятого термина для определения таких словосочетаний. В рамках нашего исследования мы будем использовать предложенный Д. Байбером (D. Biber) с соавторами термин «lexical bundles», который авторы определяют как «последовательность слов, часто встречающуюся в определенном стиле речи» [7]. В отечественных публикациях линейно организованным сочетаниям лексических единиц, не образующих семантического целого, почти не уделяется внимания и, следовательно, отсутствует общепринятое название для данного явления. Поэтому вслед за Н.В. Денисовой и Е.С. Петровой мы будем использовать термин-кальку «лексические пучки» [8].

2. Списки частотной лексики в английском научном дискурсе

Наиболее широко описано использование корпусов в учебных целях в англоязычной литературе. В английском языке эти исследования в 21 веке проводятся с очень большим размахом и широко используются на практике, например, при создании учебников по английскому языку для специальных целей. Речь идет о созданном в 2000г. частотном списке слов общенаучной лексики — AWL (academic wordlist), который включает в себя 570 гнезд слов (word families) [9] на базе корпуса, составленного из статей из научных журналов и учебников по 28 предметным областям, относящимся к четырем крупным областям знания (естествознание, юриспруденция, бизнес и гуманитарные науки) и содержащего 3,5 млн. слов. Список составлялся на основе трех критериев: *specialised occurrence, range and frequency* [9, p.221] (на данный момент работу цитировали 2670 раз). В 2014 г. Д. Гарднер и М. Дэвис (D. Gardner и M. Davies) опубликовали статью, в которой показали необходимость создания нового списка частотных слов общенаучной лексики английского языка на базе подкорпуса научной периодики (academic subcorpus), содержащего 120 млн. слов — части Корпуса современного американского английского (COCA) [10]. Целью исследователей было выделить зону общенаучной «ядерной» лексики, отсекая «слева» высокочастотные слова, встречающиеся во всех типах речи, и «справа» — узкопредметную лексику. Для этого они использовали 4 критерия (ratio, range, dispersion, discipline measure). В результате был получен новый список частотных общенаучных слов, содержащий 3000 слов, полностью доступный на сайте COCA[1]. В данной работе приведены

[1] Corpus of Contemporary American English <https://www.academicwords.info/>

500 наиболее частотных слов из этого списка [10, p. 317–320]. В 2013 г. Л. Валипори и Х. Нассаи (L. Valipouri и H. Nassaji) опубликовали список частотных слов, встречающихся в научном дискурсе по химии (Chemistry Academic WordList (CAWL)), основанный на анализе корпуса, состоящего из 1185 научных статей по химии [11]. Они обнаружили, что 27,85% слов из их списка (CAWL) отсутствовали в широко известном списке слов общенаучной лексики — AWL, составленном А. Кохед (A. Coxhead) [9].

Масштабный проект по созданию и обработке корпусов на 9 языках (английский, польский, итальянский, шведский, норвежский, русский, китайский, греческий, арабский), описан в [12]. Результатом проекта стали списки частотных слов для 9 языков и 72 языковых пар для изучающих иностранные языки и переводчиков. База данных этого проекта, получившего название KELLY, находится в открытом доступе[2], и авторы призывают преподавателей указанных языков и лингвистов активно использовать этот ресурс в своей работе [12, p.155].

Что касается языка для специальных целей, то по мнению М. Маккарти (M. McCarthy) с соавторами составление обычного частотного списка слов уже может предоставить достаточно информации для выделения характерных особенностей дискурса корпуса [13]. Типичные повторяющиеся сочетания слов («лексические пучки»), выделенные из специального корпуса, дают дополнительное представление о языке определенной области (языке для специальных целей).

В англоязычной литературе, особенно в Великобритании, в последнее время широкое распространение получило понятие formulaic language — лингвистический термин для вербальных выражений, которые устойчивы по форме, но не являются единицами плана содержания и которые тесно связаны с коммуникативно-прагматическим контекстом. Проблемы, связанные с изучением formulaic language, особенно активно исследуются в работах, посвященных проблемам изучения иностранного языка. Исследователи обращают внимание на то, что «лексические пучки» занимают, с одной стороны, очень важное место в научном дискурсе, а с другой — в разных дисциплинах они ведут себя по-разному и используются с разной частотностью [7, 14, 15]. Несмотря на наличие работ по исследованию роли лексических пучков в изучении таких языков, как корейский и испанский, большинство исследований проведено на материале английского языка [16]. В статье, опубликованной в 2008 г., К. Хайлэнд (K. Hyland) приводит списки 50 наиболее частотных выражений состоящих из 4-х слов в 4-х предметных областях: биологии, электротехнике, прикладной лингвистике и деловому администрированию, полученных на основе анализа специальных корпусов, собранных из научных статей, магистерских и докторских (PhD) диссертаций общим объемом 3,5млн. слов. При этом он отмечает, что в корпусе по электротехнике таких четырехкомпонентных лексических пучков значительно больше, чем в других корпусах. Больше половины выражений из каждого предметного списка не встречаются в других корпусах, собранных для исследования, и только 5 самых частотных лексических пучков (*on the other hand, as well as, at the same time, the results of the, in the case of*) встречаются во всех корпусах, а 14 лексических пучков — в 3-х разных областях [15, p.12–13].

В 2010 г. Р. Симпсон-Влах и Н. Эллис (R. Simpson-Vlach и N. Ellis) опубликовали список научных «шаблонов» (Academic Formulas List), включив в него выражения, которые встречаются в научных текстах гораздо чаще, чем в других видах дискурса, и типичны для разных подвидов научного дискурса [17]. В 2015 г. Дж. Фокс и М. Тигчелаар (J. Fox и M. Tigchelaar) предложили список из 99 лексических пучков, которые, по их мнению, целесообразно использовать при обучении будущих инженеров письменному дискурсу. Этот список был составлен на основе анализа корпуса научных статей по инженерным специальностям и разделен на 3 функциональные категории: референтные выражения (Referential expressions), выражения, организующие дискурс, и выражения, вводящие новое

[2] KELLY lists<https://www.npmjs.com/package/kelly-lists>, KELLY DB http://kelly.sketchengine.co.uk>

утверждение (stance expressions) [18]. Широкий обзор исследований, в основном, зарубежных авторов и на материале английского языка, по теме *formulaic language* дан в монографии Д. Вуда (D. Wood) [19].

3. Доступность немецких корпусов для специальных целей

Насколько нам известно, специальных корпусов немецкого языка, которые могли бы стать ресурсом для студентов инженерных специальностей при освоении письменной профессиональной коммуникации на немецком языке, не существует. Основанием для такого вывода является сравнительный анализ работ, посвященных немецкому и английскому языку для специальных целей за 1998–2012 гг., проведенный С. Яворска [20]. Автор ссылается на международный проект *Gesprochene Wissenschaftssprache Kontrastiv (GeWiss)* «Разговорный научный язык в сравнении», содержащий 1,2 млн. слов, свободно доступный онлайн для проведения сравнительных исследований устного научного дискурса на 3-х языках: немецком, английском и польском (https://gewiss.uni-leipzig.de). GeWiss-корпус дополняет такие известные корпусы, как Мичиганский корпус разговорного научного английского (MICASE) и корпус Британского научного разговорного английского (BASE) [20, p. 187]. Что касается обучения письменному научному дискурсу, то С. Яворска ссылается на исследования Г. Грэфен (G. Graefen), которая в обучении студентов письменному немецкому для специальных целей подчеркивала важность сопоставления того, как используется общенаучная лексика в научном дискурсе и обыденной речи, ее метафорический характер в научном дискурсе [21]. Также обращается внимание на важность списков частотных слов и словосочетаний, Примеры подобного рода содержатся в пособии [22], предназначенном для студентов и аспирантов инженерных направлений, которые пишут свои диссертации на немецком языке, не являющимся для них родным. Яворска признает, что использование подходов корпусной лингвистики в изучении общенаучного словаря немецкого языка находятся в начальной стадии [20].

Зная это, доцент Ганноверского университета им Лейбница З. Шрот-Вихерт (S. Schroth Wiechert) решила создать специальный корпус немецкого языка как ресурс для изучающих общенаучный и технический немецкий. Проведенный ею анализ показал, что подобных ресурсов очень мало, они не являются легкодоступными, а находящиеся в них примеры не соответствуют направлению подготовки магистрантов и аспирантов по таким специальностям как, например, *турбостроение, механика жидкостей и газов* или *гражданское строительство.* Ее инициатива получила развитие в рамках программы «Стратегическое партнерство» между Санкт-Петербургским Политехническим университетом Петра Великого (СПбПУ) и Ганноверским университетом им. Лейбница (ГУЛ). С 2014 г. Лингвистический центр ГУЛ и кафедры гуманитарного института СПбПУ работают над созданием трилингвальной платформы, ***D**eutsch, **E**nglish and **R**usskii* (DEaR)-корпус, в названии которого отражены языки (немецкий, английский и русский), на которых написаны диссертации и научные статьи, входящие в корпус [23]. Предполагается, что DEaR -корпус будет доступным онлайн аннотированным корпусом с встроенной поисковой системой (корпусным менеджером). Однако, кроме технических проблем, необходимо решить правовые, которые на данный момент запрещают пользователям за пределами ГУЛ обращаться к материалам немецкой части корпуса — наиболее разработанной и названной the Korpus der Ingenieurwissenschaften *(Kod.ING).*

С учетом этого обстоятельства было решено проанализировать подкорпус немецких диссертаций по электротехнике с использованием собственного корпусного менеджера.

4. Сервис корпусного менеджера

Корпус текстов становится мощным инструментом в руках лингвиста лишь посредством специализированных программных средств. Неотъемлемой частью понятия «корпус

текстов» является система управления текстовыми и лингвистическими данными, которую чаще всего называют корпус-менеджером (или корпусным менеджером). Это специализированная система, включающая программные средства для поиска данных в корпусе, получения статистической информации и предоставления результатов пользователю в удобной форме. Также сюда с некоторой долей условности можно отнести сюда средства подготовки и загрузки текстов в корпус.

Современные корпусные системы позволяют не только формировать конкордансы для заданных слов и частотные списки, но и решать достаточно сложные задачи, такие как выявление коллокаций (устойчивых сочетаний), ключевых слов и словосочетаний, построение лексико-семантических групп и др. Функциональность корпусного менеджера определяется типом данных и конкретной задачей, для которой корпус создается. Это же правомерно и для корпусов, обрабатывающих специальные тексты.

Большинство современных инструментов, используемых корпусной лингвистикой, предлагают множество функций, включая статистические методы, обладают некоторой масштабируемостью для работы с большими корпусами, предлагают многоязычную поддержку и включают в себя дружественный интерфейс. Наибольшим их ограничением является то, что они плохо работают с большими корпусами. Более мощные системы 4-го поколения, такие как corpus.byu.edu, CQPweb, Sketch Engine, Wmatrix предлагают лучшую масштабируемость за счет хранения корпуса в базе данных веб-сервера и предварительной индексации данных для обеспечения быстрого поиска.

Однако несмотря на перечисленные выше преимущества новых корпусных систем, они также имеют ряд ограничений. Они требуют покупки программного обеспечения, наличие собственного сервера и поддержание его работоспособности, установки программного обеспечения на сервер, для чего необходима высокая программистская квалификация. Еще одна проблема заключается в том, что в инструментах 4-го поколения размыты границы между данными и инструментом. Из-за способа хранения данных в индексированной форме на внешнем сервере, пользователи не имеют возможности обратиться к исходным данным непосредственно, по крайней быстро соотнести их с результатами поиска в корпусе.

Анализ показывает, что набор свободно доступных корпусных программ совсем не велик. Наиболее известны такие свободно распространяемые универсальные корпусные менеджеры как XAIRA (BNC), Manatee/Bonito, CQP, DDC, Wordsmith, MonoConc и AntConc.

Мы остановили свой выбор на одном из наиболее эффективных корпусных менеджеров, AntConc, программе, разработанной профессором Университета Васэда (Япония) Л. Антони (L. Anthony) и доступной на его сайте[3]. Выбор данного компьютерного инструмента обусловлен тем, что это свободно распространяемое мультиплатформное программное обеспечение, оснащенное удобным интерфейсом, имеющее множество функций по автоматической обработке текстов в разной кодировке и в разных форматах. AntConc позволяет обрабатывать тексты в кодировках Unicode, iso, cp, koi8, ascii для текстов на европейских языках и в специальных кодировках для текстов на восточных языках. Форматы входных файлов, поддерживаемые в AntConc — это .txt, .html, .htm, .xml. В состав AntConc входят статистический модуль и модуль машинного обучения. AntConc допускает обработку как «сырых», так и размеченных текстов. Возможно подключение списков стоп-слов, списков лемм, списков отрицательных ключевых слов и пр.

С помощью AntConc можно производить следующие операции:

- просмотр файла с текстом;
- построение конкорданса для целевого слова в пределах контекстного окна;
- построение графиков к конкордансу;

[3] Laurence Antony’s AntConc, a freeware corpus analysis toolkit for concordancing and text analysis <http://www.laurenceanthony.net/software/antconc/>.

- выделение ключевых слов по двум критериям log-ikelihood и χ2 в анализируемом корпусе с выдачей результатов о ранге и частоте;
- выделение коллокатов целевого слова на основе двух коэффициентов ассоциации (MI и log-likelihood) в пределах контекстного окна;
- построение частотного списка словоформ и/или лемм для обрабатываемого корпуса с указанием ранга и абсолютной частоты;
- построение частотного списка кластеров для заданных слов;
- выделение N-грамм с целевым словом в пределах контекстного окна и построение частотного списка выделенных N-грамм.

Именно последние две функции были использованы нами для получения результатов для написания данной статьи.

5. Материал и инструмент исследования

Выбранный для анализа корпуса немецких диссертаций инструмент, AntConc, относится к корпусным менеджерам 3-го поколения [24]. Их главным недостатком является то, что они не могут справиться с очень большими корпусами, содержащими больше 100 млн. слов [24, p. 152]. Однако, как мы указали в предыдущем разделе, выбор корпусного менеджера 4-го поколения в данный момент не представлялся нам возможным,

В системе AntConc имеется два инструмента, позволяющих выявить наиболее частые N-граммы: «Кластеры» (The Clusters Tool или просто Cluster) и N-граммы (The N-Grams Tool). Первый из них фактически суммирует результаты, полученные в инструменте Concordance Tool или Concordance Plot Tool, подсчитывая частоты левых и правых кластеров (окружений, контекстов) вокруг заданного слова. При этом имеется возможность задать длину кластера и другие параметры настройки. Второй инструмент позволяет сканировать весь корпус и вычислять частотность для кластеров длины N. Все параметры настройки, доступные в инструменте «Кластеры», также доступны в инструменте N-грамм.

Очень привлекательным выглядит поиск в режиме *N-Gram* с указанием минимального и максимального размера лексического пучка, минимальной частоты единицы в корпусе и количества текстов, в которых она встречается (range). К сожалению, для поиска по всем текстам Kod.ING-корпус оказался слишком велик. Для проведения поиска с использованием инструмента «Clusters/N-Grams» нами был использован стационарный компьютер с 64-битной операционной системой Windows 8 с частотой 4-ядерного процессора 3,40 МГц и объемом оперативной памяти 16 Гб. Неоптимальное использование программой AntConc оперативной памяти в процессе обработки массива данных приводит к переполнению последней и последующему аварийному прекращению работы программы. Решить проблему можно, если обрабатывать только небольшие группы файлов, или искать компьютер с большим объёмом оперативной памяти. Для обработки всех файлов корпуса нужно очень большой объем оперативной памяти. Мы установили, что для обработки каждого подкорпуса в отдельности используется 95–99% оперативной памяти указанного компьютера.

6. Получение лексических пучков для высокочастотных слов Kod.ING-корпуса

6.1.Получение лексических пучков с помощью функции *Cluster*

Получение лексических пучков с помощью функции *Cluster* предполагает генерирование частотного списка слов всего корпуса текстов диссертаций, выделение из него самых частотных существительных и дальнейший поиск частотных лексических пучков на их основе или на основе сложного слова, составной частью которого является

слово из списка. Мы проанализировали списки из 100 самых частотных слов в целом Kod.ING корпусе и в каждом подкорпусе.

Проанализировав разные подкорпусы, можно сделать вывод, что после вспомогательных частей речи самыми частотными значимыми словами являются существительные. Общеупотребительных существительных, которые входят в 100 самых употребляемых слов, больше чем общенаучных. Самыми частотными словами оказались такие существительные, которые можно будет встретить в диссертационных текстах по разным специальностям, например: *Daten* (данные), *Abbildung* (изображение), *Kapitel* (глава), *Vergleich* (сравнение), *Untesuchung* (исследование). К самым частотным общенаучным существительным стоит отнести: *Simulation* (симуляция), *Berechnung* (измерение), *Verfahren* (процесс), *Temperatur* (температура). Также присутствуют слова, которые представляют пласт общеупотребительной лексики: *Einfluss* (влияние), *Anzahl* (количество), *Bereich* (область).

В своих работах Г. Грэфен предлагает использовать для составления списков для изучения в курсе немецкого языка распространенные существительные, глаголы и прилагательные в сочетании с их наиболее частотными коллокациями [21]. Мы проверили, являются ли выделенные ею слова *Analyse* (анализ) and *Aspekt* (аспект) высокочастотными в нашем корпусе. В режиме поиска *Cluster/N-Grams* поисковый запрос формируем следующим образом: *Analyse, указав размер кластера Min.2 — Max.5 и минимальную частотность 10. Такой запрос позволяет найти сложные слова с корнем «Analyse» и коллокации с ними. Результаты представлены в таблице 1.

Таблица 1. Наиболее частотные лексические пучки Kod.ING-корпуса, отобранные с использованием сложных слов с корнем «Analyse»

№	Частотность (в корпусе)	Лексический пучок
1	46	Praxisanalyse / Gegenüberstellung
2	42	Praxisanalyse / Gegenüberstellung Zwischenergebnisse
3	20	Feinanalyse und
4	19	Feinanalyse und technische
5	19	Feinanalyse und technische Realisierung
6	11	Sensitivitätsanalyse der
7	10	Bildanalyse-Objektmodellen
8	10	Eigenwertanalyse zyklischer
9	10	Qualitätsanalyse von
10	10	Qualitätsanalyse von gdv
11	10	Qualitätsanalyse von gdv bei

Как и следовало ожидать, выделить сочетания существительных с глаголами автоматически не удалось. Причина состоит в том, что из-за синтаксических особенностей немецкого языка расстояние между существительным и относящимся к нему глаголом часто бывает больше 5–6 слов.

6.2. Получение лексических пучков с помощью функции N-Grams

Так как поиск по всему Kod.ING-корпусус помощью этой функции оказался невозможным, то мы ограничились поиском по подкорпусу из 35 PhD диссертаций на немецком языке в области электротехники. Поиск проводился по следующим параметрам: длина лексического пучка 2–4 слова, минимальная частота встречаемости — 10 раз во всех текстах; распределение (range) — 5, что означает, что искомые пучки должны встречаться не менее чем в 5 разных текстах подкорпуса. Главным критерием отбора лексических пучков являлась их частотность в корпусе. Оборотная сторона применения этой простой метрики состоит в том, что для нахождения «нужных выражений» приходится вручную

обрабатывать длинные списки лексических пучков, найденных программой как удовлетворяющих поисковому запросу.

Поясним, какие лексические пучки мы искали. М. Маккарти (McCarthy) с соавторами указывают, что найденные в корпусе лексические пучки могут состоять:

1) из случайного набора слов, часто встречающихся вместе (*so dass die, die durch die*);

2) из синтаксически неполных, но «осмысленных» выражений: (*zwischen den beiden, auf Basisder, in bezug auf*) и

3) из семантически и прагматически «законченных» выражений: (*im Rahmen dieser Arbeit, ist in Abbildung dargestellt)* (примеры взяты из Kod.ING-корпуса*).* (В своей монографии авторы приводят примеры лексических пучков каждой категории на материале английского языка: as *are to my, this one for* (1), *to be able to, a lot of the* (2) и *on the other hand* and *as a result* (3) [13, p. 61]). Нас интересовали выражения, относящиеся ко второй и к третьей группе. Всего при указанных параметрах нами было идентифицировано 50 выражений. Они представлены в таблице 2.

7. Обсуждение и выводы

Подавляющее большинство выражений представляют собой предложные конструкции: *предлог + существительное*. В немецком более распространены предложные конструкции, которые по структуре представляют собой связку глагол + предлог, которая называется «управление глаголов» и является грамматической доминантой в сочетаемости предлогов с другими частями речи.

Обычно конструкции с предлогами являются составными частями более крупных конструкций. В традиционной лингвистике сочетаемость предлогов, как правило, описывается с точки зрения их грамматических (падежных) функций, однако семантике предлогов и их функциональной роли до настоящего момента должного внимания не уделялось. И совсем не исследованным остается вопрос о семантике и сочетаемости предлогов в специальных текстах. Корпусных исследований, посвященных предлогам, крайне мало. Исчерпывающие формальные описания функционирования предлогов в составе конструкций отсутствуют.

Необходимость корпусно-статистического описания немецких предлогов ставит перед лингвистами сложную задачу. В связи с тем, что глагол и предлог в немецком предложении могут находиться не рядом, что обусловлено рамочной конструкцией, при автоматическом поиске таких сочетаний находится лишь незначительная их часть. Среди трудностей следует назвать также синонимию предлогов и вариативность конструкций. Для исследований необходимо использовать несколько корпусных источников, т.к. типы конструкций и их частотные характеристики для одного и того же предлога в тематически разнородных текстах могут не совпадать.

В предложных конструкциях с существительными предлог может находиться в предпозиции (***in*** *der Regel, bei der Modellierung*) или постпозиции к существительному (*abhängig von, Beispiel für*). Как известно, в немецком языке предлоги управляют следующим после себя словом, т.е. после предлогов zu/von последующее слово будет находиться в дательном падеже, а после предлогов auf/für- в винительном. Маркером рода и падежа является артикль, поэтому он и является такой частотной частью речи в немецких предложениях и широко представлен в корпусе.

В большинстве частотных лексических пучков присутствует артикль, который выполняет несколько функций:

- определять род главного существительного предложного сочетания, например: ***Der*** *Zugriff auf, in* ***der*** *Literatur;*
- являться частью следующего слова, которое не входит в данный лексический пучок, например: *Einfluss auf* ***die*** *,Auf Basis* ***der****, Parameter für* ***die.*** Хайленд также включает в свои списки высокочастотных выражений в разных предметных областях английского

языка выражения, заканчивающиеся определенным артиклем: *due to **the**, in terms of **the**, as a result of **the*** [15, p.7].

– Отдельно можно выделить такие предложные конструкции, как предлог+сущ.+предлог*: i**m** Bereich von; i**n** Form von.* В целом, преобладающими в постпозиции являются сочетания существительных с предлогами*von* и *zu*: сущ. + ***von,*** сущ + ***zu***, а в препозиции – с предлогами *bei* и *in (im*): ***bei***+ сущ. ***in/im*** + сущ.

Таблица 2. Наиболее частотные лексические пучки Kod.ING- корпуса, отобранные с использованием функции *N-Grams*

№	Частотность (в корпусе)	Трехкомпонентные лексические пучки	Двух и четырех компонентные лексические пучки	Частотность (в корпусе)
1	417	in dieser Arbeit	im Rahmen dieser Arbeit	219
2	336	in diesem Fall	zum Zeitpunkt	71
3	219	in der Regel	Beispiel für	69
4	225	mit Hilfe der	Abhängig von	58
5	221	die Anzahl der	Unter Berücksichtigung	58
6	179	auf diese Weise	Vor allem	57
7	179	aus diesem Grund	Mit einem Durchmesser von	21
8	168	mit Hilfe von	Differenz zwischen	10
9	164	im Vergleich zu	Verfahren für	10
10	143	in Bezug auf	für die Herstellung von	10
11	127	In der Literatur		
12	120	in Abhängigkeit von		
13	137	In Abbildung dargesellt		
14	130	Einfluss auf die		
15	113	Stand der Technik		
16	104	im Gegensatz zu		
17	103	auf Basis der		
18	75	in Richtung der		
19	74	in Form von		
20	73	die Verwendung von		
21	72	unter der Annahme		
22	70	es ergibt sich		
23	66	für die Berechnung der		
24	64	der Zugriff auf		
25	62	als Funktion der		
26	58	im Bereich von		
27	56	im folgenden werden		
28	53	im folgenden wird		
29	46	bei der Entwicklung		
30	43	im folgenden Abschnitt		
31	38	auf der Oberfläche		
32	30	bei der Simulation		
33	20	bei einer Frequenz		
34	19	bei der Analyse		
35	18	Parameter für die		
36	18	in der Datenbank		
37	18	bei der Untersuchung		
38	17	bei der Implementierung		
39	17	bei der Modellierung		
40	13	in der Größenordnung		

Еще одним признаком научного стиля является большое количество пассивных конструкций с вспомогательным глаголом *werden,* например*:im Folgenden **werden***. Но

выявить конструкцию целиком автоматически не получается из-за большого расстояния между вспомогательным и смысловым глаголом.

Проведенный анализ позволяет сделать следующие предварительные выводы.

1) В немецком языке, как и в любом флективном языке с изменяющимися формами существительных, правильное употребление требуемой формы слова является непростой задачей для студентов инженерных специальностей. Имея информацию о частоте употребления определенных форм в письменном научном дискурсе, можно целенаправленно уделять больше внимания именно им при обучении студентов письменной практике для научных целей.

2) В состав большинства предложных сочетаний с существительными входит артикль. С употреблением артиклей связано большое количество ошибок. Поэтому возможность проанализировать их употребление в частотных лексических пучках является очень ценной. Наблюдая, в каких случаях на практике используется определенный артикль, а в каких неопределенный, как в "*der* direkte Vergleich" и "*ein* direkter Vergleich" можно с помощью расширенного контекста эмпирически понять, в чем разница в их употреблении.

3) Автоматический поиск по текстам корпуса с использованием программы *AntConc* не позволяет учесть некоторые особенности немецкого синтаксиса при формировании запросов. Например, личные формы глагола в немецком языке могут состоять из нескольких частей (один или два вспомогательных глагола и смысловой глагол), которые не всегда занимают в предложении место рядом с существительным, к которому непосредственно относятся, будь оно подлежащим или дополнением, что не позволяет легко находить в корпусе сочетания существительных с глаголами. В связи с этим, необходима ручная обработка полученного конкорданса. Аналогичной обработки требуют и глаголы с отделяемыми приставками. Техническим решением может стать использование корпусного менеджера с более гибким языком запросов и учет особенностей немецкого синтаксиса и нужных функций для анализа корпуса при разработке корпусного менеджера *HanConc*, встроенного в разрабатываемый *Kod.ING*-корпус.

Полученные результаты позволяют надеяться, что при дальнейшем исследовании Kod.ING-корпуса могут быть получены новые интересные данные по частотности и сочетаемости слов разных предметных областях, которые могут быть использованы в дидактических целях.

Исследование поддержано программой сотрудничества «Стратегическое партнерство», проекты DAAD (Deutscher Akademischer Austauschdienst) № 56268450 (2013-2016) и № 57271274 (2017-2018).

Литература

[1] Алексеев П.М. Статистическая лексикография. Л.: Изд-во ЛГПИ, 1975.
[2] Арапов М.В. Квантитативная лингвистика. М.: Наука, 1988.
[3] Головин Б.Н. Язык и статистика М.: Просвещение, 1971.
[4] Пиотровский Р.Г. Информационные измерения языка. Л.: Наука. Ленинградское отделение, 1968.
[5] Фрумкина Р.М. Статистические методы изучения лексики. М.: Наука, 1964.
[6] Römer U. The inseparability of lexis and grammar: Corpus linguistic perspectives // Annual Review of Cognitive Linguistics. 2009.Vol. 7. P. 141–163.
[7] Biber D., Conrad S., Cortes V. If you look at ...: Lexical bundles in university teaching and textbooks // Appl. Linguist. 2004. Vol. 25, № 3. P. 371–405.
[8] Денисова Н.В., Петрова Е.С. Маркеры ряда: кластерность и вариантность (на материале английского языка) // Вестник Челябинского государственного университета. 2008. Т. 30. С 44–49.
[9] Coxhead A. A new academic word list // TESOL Quarterly. 2000. Vol. 34, № 2. P.213–238.

[10]Gardner D., Davies M. A New Academic Vocabulary List // Applied Linguistics. 2014. Vol. 35, № 3. P. 305–327.
[11]Valipouri L., Nassaji H. A corpus-based study of academic vocabulary in chemistry research articles // J. of English for Academic Purposes. 2013. Vol. 12, № 4. P. 248–263.
[12]Kilgarriff A, Charalabopoulou F., Gavrilidou M., Johannessen J. B., Khalil S., Johannessen K., Sofie J., et al. Corpus-based vocabulary lists for language learners for nine languages // Language Resources and Evaluation. 2014. Vol. 48, № 1. P. 121–163.
[13]McCarthy M., O'Keeffe A., Cartner R. From Corpus to Classroom. UK, Cambridge: CUP, 2007. 315 p.
[14]Cortes V. Teaching lexical bundles in the disciplines: An example from a writing intensive history class // Linguistics and Education. 2016. Vol.17. P. 391–406.
[15]Hyland K. As can be seen: Lexical bundles and disciplinary variation // ESP.2008. Vol. 27, № 1. P. 4–21.
[16]Hyland K. Bundles in Academic Discourse // Annual Review of Applied Linguistics. 2012. № 32. P. 150–169.
[17]Simpson-Vlach R., Ellis N.C. An Academic Formulas List: New Methods in Phraseology Research // Applied Linguistics. 2010. Vol. 31, № 4. P. 487–512.
[18]Fox J., Tigchelaar M. Creating an engineering academic formulas list // The Journal of Teaching English for Specific and Academic Purposes. 2015. Vol. 3, № 2. P. 295 –304.
[19]Wood D. Fundamentals of Formulaic Language: An introduction. London: Bloomsbury. 2015. 205 p.
[20]Jaworska S. Review of recent research (1998–2012) in German for Academic Purposes (GAP) in comparison with English for Academic Purposes (EAP): cross-influences, synergies and implications for further research // Lang. Teach. 2015. Vol. 48, № 2. P. 163–197.
[21]Graefen G. Die Didaktik des wissenschaftlichen Schreibens: Möglichkeiten der Umsetzung // GFL-journal. 2009. № 2-3. S. 106–127.
[22]Schroth-Wiechert S. Deutsch als Fremdsprache in den Ingenieurwissenschaften: Formulierungshilfen für schriftliche Arbeiten in Studium und Beruf. Berlin: Cornelsen Verlag. 2011. 160 p.
[23]Gärtner T., Schroth-Wiechert S., Kogan M. A trilingual platform for academic technical writing // Коммуникация в поликодовом пространстве: лингво-культорологические, дидактические, ценностные аспекты: Материалы междунар. Научн. конф. СПб: Изд-во Политехн. ун-та, 2015. С. 67–68.
[24]Anthony L. A critical look at software tools in corpus linguistics // Linguistic Research. 2013. Vol. 30, № 2. P. 141-161.

On Problem of Creating a German Engineering Academic Words and Formulas List

M.S. Kogan [1], A.M. Yaroshevich [1], A.Y. Kolotaeva [1,] V.P. Zakharov [2,]
S. Schroth Wiechert [3], A. Tilmans [3]

[1] Peter the Great Saint Petersburg Polytechnic University,
[2] Saint Petersburg State University,
[3] Leibniz Universität Hannover

The article deals with statistical analysis of German texts in order to find reoccurring lexical bundles. Unlike the English language there are no academic formulas and words frequency lists in German academic discourse. The authors analyzed the *Elektrotechnik* subcorpus of the Kod.ING corpus, which contains PhD dissertations in German in different engineering fields. The Kod.ING corpus is being developed at Leibniz University of Hannover (LUH) as a part of a joint project with St. Petersburg Polytechnic University.

The article presents lists of the most frequent nouns and lexical bundles found with *WordList* and *N-gram* functions correspondingly of the AntConc freeware corpus analysis toolkit for concordancing and text analysis. The paper discusses lexical and grammatical peculiarities of the selected lexical bundles, highlighting similarities with English academic formulas and words frequency lists. In particular, the dominance of prepositional phrases in the list has been noticed. Limitations of AntConc program for analysis of relatively large German corpora are described and perspectives of applying the findings in teaching German for specific purposes are considered.

Keywords: corpus linguistics, lexical bundles, academic formulas and word list, English academic discourse, Kod.ING corpus, German for specific purposes