DOI: 10.24833/2410-2423-2020-4-24-14-24

FREQUENCY OF CO-OCCURRENCE OF CHINESE CHARACTERS AS AN INDICATOR OF LEXICALITY (WHEN SELECTING THE VOCABULARY OF CHINESE MILITARY DISCOURSE)

Dmitry S. Korshunov

Military University of Radio Electronics,
126, Sovetsky Prospect, Cherepovets, Vologda region, 162600, Russia.

Abstract. *Teaching a foreign language in a non-linguistic college or university should be professionally oriented, which brings up the question of selecting the relevant vocabulary of a professional discourse under study. Modern text corpora are too general in subject matter and the time span. Therefore, a specially compiled collection of texts can serve the purpose of selecting the vocabulary. In the case of the Chinese language, the task is complicated by the lack of word segmentation in such texts. Taking into account the fact that most words in Chinese are written in two characters, it is assumed that one of the methods applicable in this situation is a comprehensive frequency analysis of text sequences of two characters – character bigrams. The analysis of frequent bigrams has showed that 70% of the most frequent lexical units are representative of the discourse, including 11% of out-of-vocabulary ones. The remaining part of bigrams pertain to syntactic constructions, including structurally incomplete ones, and fragments of longer lexical units. Thus, the high frequency of character co-occurrence can with a rather high probability ($p > 0.7$) be considered as an indicator of lexicality in identifying representative vocabulary in an unsegmented thematic collection of texts in Chinese.*

Key Words: *Chinese military discourse, Chinese language, military news reports, vocabulary selection, lexical units, out-of-vocabulary words, character bigrams, frequency analysis, frequency of co-occurrence*

For citation: Korshunov D.S. 2020. Frequency of Co-Occurrence of Chinese Characters as an Indicator of Lexicality (When Selecting the Vocabulary of Chinese Military Discourse). *Philological Sciences at MGIMO.* Vol. 6. No 4(24). P. 14–24. https://doi.org/10.24833/2410-2423-2020-4-24-14-24

ЧАСТОТА СОВМЕСТНОЙ ВСТРЕЧАЕМОСТИ ИЕРОГЛИФОВ КАК ПОКАЗАТЕЛЬ ЛЕКСИЧНОСТИ (ПРИ ОТБОРЕ ЛЕКСИКИ КИТАЙСКОГО ВОЕННОГО ДИСКУРСА)

Д.С. Коршунов

Военный университет радиоэлектроники,
162600, Россия, Вологодская область, г. Череповец, пр. Советский, 126

Аннотация. *Преподавание иностранного языка в неязыковом вузе должно быть профессионально ориентированным, что ставит перед преподавателем задачу отбора лексики, репрезентативной для профессионального дискурса изучаемой специальности и актуальной для его текущего состояния. Современные корпусы текстов являются слишком общими по тематике и охватываемому периоду для такой узкой задачи. Поэтому материалом для отбора лексики должна выступать специально составленная коллекция текстов. В случае с китайским языком задача осложняется отсутствием сегментации таких текстов на слова. С учётом того, что большинство слов китайского языка записываются двумя иероглифами, предполагается, что одним из применимых в этой ситуации методов может быть сплошной частотный анализ текстовых последовательностей из двух иероглифов – иероглифических биграмм. В результате такого анализа среди частотных биграмм получено более 70 % репрезентативных для данного дискурса лексических единиц, в том числе 11 % несловарных. Оставшаяся доля биграмм приходится на синтаксические конструкции, в том числе структурно незавершённые, и фрагменты более длинных лексических единиц. Таким образом, высокая частота совместной встречаемости иероглифов может с достаточно большой вероятностью ($p > 0,7$) рассматриваться как показатель лексичности при выявлении репрезентативной лексики в несегментированной тематической коллекции текстов на китайском языке.*

Ключевые слова: *китайский военный дискурс, китайский язык, новостные сообщения военной тематики, отбор лексики, лексические единицы, несловарные слова, иероглифические биграммы, частота совместной встречаемости*

Для цитирования: Коршунов Д.С. 2020. Частота совместной встречаемости иероглифов как показатель лексичности (при отборе лексики китайского военного дискурса). *Филологические науки в МГИМО.* Том 6. № 4(24). С. 14–24. https://doi.org/10.24833/2410-2423-2020-4-24-14-24

1. Введение

1.1 Постановка задачи

Актуальные вопросы лингвистической теории нередко возникают из практических, прикладных задач, требующих осмысления, проверки и переосмысления эмпирических результатов этой проверки.

Одной из традиционных прикладных задач языкознания (в широком смысле этого слова) является преподавание иностранного языка. В подавляющем большинстве вузов иностранный язык не является профильным предметом, однако предполагается, что вся терминологическая система изучаемой в вузе специальности во всём многообразии внутрисистемных отношений, а также сопутствующая ей лексика, которые изучаются в течение 4–5 лет на родном языке, должны быть изучены и на иностранном языке. Правда, бюджет времени на иностранный язык в неязыковом вузе выделяется обычно скромный. Это ставит перед преподавателем задачу оптимизации содержания обучения, подбора наиболее репрезентативного лексико-грамматического материала и ряд других дидактических задач [11; 3].

Настоящая работа исходит из частной прикладной задачи отбора лексики, репрезентативной для профессионального дискурса изучаемого языка, а именно – китайского военного дискурса, в том его сегменте, который представлен в новостных сообщениях военной тематики электронных СМИ КНР на китайском языке. При решении этой задачи возникает вопрос не только и не столько критерия репрезентативности лексических единиц, сколько критерия – или хотя бы ин-

дикатора, показателя – лексичности, то есть на каком основании ту или иную последовательность иероглифов можно считать лексической единицей – словом или словосочетанием. Поскольку репрезентативность лексики обеспечивается прежде всего большим объёмом обрабатываемого материала, предпочтение должно отдаваться таким показателям, которые могут применяться при автоматической обработке письменных текстов. Это, к сожалению, исключает из рассмотрения (на начальном этапе) такие классические характеристики слова как «семантическое единство» и «фонетическая целостность» [1, с. 157].

Уже традиционно одним из ключевых факторов в лингвистических исследованиях является частота употребления тех или иных единиц в представительном массиве текстов. Очевидно, что высокая частота употребления может являться одним из критериев репрезентативности лексики, но может ли высокая частота совместной встречаемости иероглифов в текстах быть показателем лексичности конкретного сочетания иероглифов – это остаётся вопросом. Если да, может, то стопроцентный ли это показатель или лишь указание на вероятность? Если нет, то какие ещё основания, кроме принадлежности к одному слову, могут так часто соединять иероглифы в текстах?

В настоящей работе предпринимается попытка на эмпирическом материале ответить на поставленные вопросы. Поскольку статистической нормой в современном китайском языке является двуморфемное слово, которое записывается двумя иероглифами [2, с. 244], на данном этапе представляется возможным ограничиться изучением сочетаний именно двух иероглифов – иероглифических биграмм.

1.2 Исходные рассуждения

1.2.1. Необходимо оговориться, что используя слово «дискурс», мы имеем в виду довольно узкий лингвистический аспект этого разнообразно трактуемого термина и понимаем под ним совокупность текстов, коммуникативно значимых для конкретной сферы социального (профессионального) взаимодействия людей и актуальных для её текущего состояния. В значительной степени такое понимание профессионального дискурса синонимично понятию подъязыка специальности, но с условием его регулярного «обновления» до актуального состояния[1]. Например, номинация 北京军区 *бэйцзин цзюньцюй* ‘Пекинский военный округ’ по-прежнему принадлежит китайскому военному подъязыку, но уже выпала из китайского военного дискурса, так как обозначаемая ею структура с 2016 года утратила актуальность (упразднена) в результате реформы вооружённых сил КНР.

Из такого понимания объекта исследования вытекают ограничения в отношении изучаемого материала. Существующие корпусы китайских текстов (к примеру, восемь перечислено в [16, p. 60–63], ещё один представлен в [13]) оказываются слишком общими с точки зрения тематики и, как правило, ориентированными на слишком широкий временной период с точки зрения актуальности терминологии. Поэтому удовлетворяющим требованиям тематичности и актуальности языковым материалом представляется только специально отобранная коллекция текстов.

1.2.2. Обзор различных взглядов на отбор репрезентативной лексики, необходимой для составления лексических минимумов, представленный в работе [6], предлагает нам три возможных подхода: объективный – с опорой на частотность слов в тексте, субъективный – руководствующийся представлениями составителя о репрезентативности и достаточности, и смешанный. Субъективный подход по определению неидеален в силу своей субъективности, опора на абсолютную частотность может приводить к нарушению целостности семантических полей (к примеру, в частотный список могут попасть не все наименования дней недели или месяцев), поэтому смешанный подход должен предложить разумный баланс между объективными и субъективными основаниями. Именно такой баланс закладывается нами в понятие репрезентативности

[1] Нам близки взгляды Е.И. Шейгал: «Институциональный дискурс оказывается предельно широким понятием, охватывающим как языковую систему (ту её часть, которая специфически ориентирована на обслуживание данного участка коммуникации), так и речевую деятельность (в совокупности лингвистических и экстралингвистических факторов) и текст. Сказанное можно представить в виде формулы: ДИСКУРС = ПОДЪЯЗЫК + ТЕКСТ + КОНТЕКСТ» [10, с. 27].

лексики – объективно вычисляемая частотность и на основе опыта субъективно определяемое соответствие понятийной базе предметной области. В любом случае, исследователем в первую очередь должны быть получены статистические данные на достаточном массиве тематически адекватных текстов.

1.2.3. Как отмечает Е.И. Риехакайнен, «чаще всего объектом статистического изучения оказываются отдельные слова. При работе с такими единицами исследователи сталкиваются с рядом проблем, большинство из которых связано с определением того, что считать словом» [7, с. 9]. Эта вечно актуальная проблема языкознания усугубляется, если предметом нашего интереса является китайский язык. Вот что об этом писали выдающиеся отечественные китаисты В.М. Солнцев и Н.В. Солнцева:

Понятие слова вообще относится к числу трудно определяемых понятий. <…> И многие учёные (кстати, к ним относится крупнейший советский лингвист Л.В. Щерба) вообще сомневаются в возможности дать определение слова, одинаково пригодное для разных языков. <…> Трудность определения слова объясняется тем, что, во-первых, слово – это величина, сильно варьирующая от языка к языку и, во-вторых, это величина, или единица, отнюдь не самым лучшим образом отграниченная от других единиц. В частности, слово не всегда ясно отграничивается от морфемы, единицы меньшей, чем слово, и далеко не всегда существует возможность достаточно чётко провести границу между тем, что называется сложным словом, и тем, что называется сочетанием слов. Этот второй момент – недостаточно чёткая отграниченность слова от единицы меньшей, чем слово, и единицы большей, чем слово, – особенно характерен для изолирующих языков, к которым принадлежит китайский язык» [9, с. 29–30].

Эта проблема подробно описана в отечественном китаеведении. В частности, в классическом труде В.М. Солнцева «Язык как системно-структурное образование» есть целый параграф под названием «Проблема неразличимости сложного слова и словосочетания», в котором автор с опорой на работы китайских лингвистов и исследования своего учителя Н.Н. Короткова приходит к выводу, что применительно к китайскому языку «факт неразличимости при некоторых условиях слова и словосочетания можно считать установленным» [8, с. 167].

Кроме того, во всех языках слова могут образовываться непосредственно в речи, в том числе письменной. Однако «в некоторых языках, например в китайском и других языках китайско-тибетской семьи, а также в типологически близких им языках Юго-Восточной Азии, производимость слов в речи – массовое и широко распространённое явление» [там же, с. 155]. И само такое спонтанное, окказиональное словообразование, и особенно его масштаб в китайском языке создают проблему для статистической обработки текстов.

Ещё одну специфическую трудность в выделении китайских слов в тексте отмечает профессор А.Н. Алексахин: «В китайском языке центральная роль слова затеняется иероглифической письменностью текста, в котором иероглифы располагаются на равном расстоянии друг от друга и слова не отделяются пробелами» [1, с. 139].

Перечисленные проблемы дают возможность современным лингвистам прибегать к достаточно радикальным формулировкам, например: «Лексический уровень в китайском языке – слабый и представлен размытыми, легко соединяющимися в целое и легко разъединяющимися единицами» [5, с. 288]. Нам представляется не вполне оправданной характеристика всего лексического уровня китайского языка как слабого – вряд ли китайская лексика беднее или в чём-либо слабее лексики других языков, но со второй частью цитаты трудно не согласиться. Иероглифы (морфемы) легко соединяются в слова или словосочетания, и зачастую так же легко разъединяются обратно. Поэтому в целях нашего исследования представляется обоснованным выбор для поиска лексических единиц метода сплошного анализа всех иероглифических последовательностей в тексте.

1.2.4. Получение статистических данных невозможно без автоматизации обработки текстов. В этом процессе указанные выше лингвистические проблемы описываются в литературе в виде двух ключевых технических трудностей: 1) разделение текста на слова (word segmentation), проблематичное из-за отсутствия пробелов, и 2) распознавание новых слов и выражений, не зафик-

сированных в словарях, в том числе именованных сущностей (named entities), или неоднословных номинаций. На практике эти две трудности обычно усиливают друг друга [14, p. 6154].

Для человека, владеющего китайским языком, «группирование и членение иероглифического текста на слова происходит на основе знания слов» [1, с. 139]. При автоматической обработке текста его сегментация на слова начинается с сопоставления последовательности иероглифов со словарём. Однако по называвшимся выше причинам трудно ожидать исчерпывающей полноты от любого словаря китайского языка: слова в тексте могут иметь краткую и полную формы (записываться одним или двумя иероглифами), словосочетания могут «смешиваться» со словами. Кроме того, в речи, в том числе письменной, постоянно легко образуются и широко используются новые сочетания лексического уровня, которые словари фиксировать не успевают. При этом китайские лексикографы, похоже, и не ставят перед собой такую задачу – как применительно к русскому языку не ставится задача зафиксировать в словарях все возможные словосочетания. Но если в русском языке речь идёт о комбинациях целых слов с потенциально легко выводимым суммарным значением, то в китайском языке часто приходится иметь дело с морфемной контракцией – словообразованием путём сложения усечённых слов, разновидностью аббревиации (подробнее см. [4, с. 127–165]), что нередко затрудняет интерпретацию получившейся лексической единицы.

При этом, как отмечают работающие в сфере автоматической обработки китайских текстов исследователи, «даже наличие полного словаря не гарантирует правильную интерпретацию последовательностей символов», «практически любое сочетание иероглифов может быть интерпретировано тем или иным образом», и «конкретный смысл иероглиф приобретает только в контексте» [12, с. 138].

Обзор последних работ в этой сфере подтверждает, что проблема полноты словаря актуальна для IT-специалистов любого уровня. Ма Цзи с коллегами из лингвистического подразделения Google констатирует, что «несловарные слова остаются вызовом для нейросетевых моделей» и «являются основным препятствием для достижения высокой точности сегментации»[2] [18, p. 4902–4906]. Другие исследователи заявляют о прогрессе в этом направлении, однако и у них точность выделения несловарных слов в используемых для оценки разных наборах данных колеблется от 62,7 % до 92,9 % (среднее значение – 83,2 %) [15], что подтверждает наличие проблемы.

1.2.5. Биграмма как последовательность двух иероглифов, строго говоря, не является типичным языковым объектом, она может совпадать со словом, но может и не совпадать. Биграмма (как и любая n-грамма) представляет собой фрагмент реализации в тексте синтагматических отношений, к ней применимо понятие «морфемная синтагма», частным случаем которой может оказаться слово или словосочетание, в том числе терминологическое. Однако очевидно, что в иных частных случаях соседство двух иероглифов может оказаться совершенно ситуативным, случайным. Тем не менее, сам факт неоднократного появления биграммы в тексте придаёт ей определённую языковую значимость. И чем больше количество таких фактов, тем больше значимость даже не совпадающей со словом биграммы и тем очевиднее необходимость адекватной интерпретации её высокой частоты появления в тексте.

В своё время группа китайских учёных специально исследовала роль биграммы и слова в задачах автоматической обработки текстов на китайском языке. Общий вывод был таким – биграммы лучше слов подходят для выполнения этих задач [17, p. 545]. Впрочем, нас интересуют отмеченные исследователями свойства биграмм.

Во-первых, биграммы отличаются от слов тем, что в тексте они накладываются друг на друга, а слова нет. Биграмма может быть «внутрисловной» (intraword bigram), когда, допустим, в слове из трёх иероглифов получаются две накладывающиеся биграммы, каждая из которых с технической точки зрения может с хорошей долей вероятности репрезентировать полное слово. Во-вторых, биграмма может быть «межсловной» (interword bigram), захватывая по иероглифу от смежных

[2] В оригинале: “Out-of-vocabulary words remain challenging for neural-network models” (p. 4902) и “OOV is a major obstacle to achieving high segmentation accuracy” (p. 4906).

слов, что в некоторых случаях позволяет репрезентировать целое словосочетание. В-третьих, биграммы могут выявлять новые (несловарные) слова, что очень полезно в нашем случае.

Впрочем, возможны и казусы. «Внутрисловная» биграмма иногда может совпасть с совершенно другим словом. Например, в слове 天文学 *тяньвэньсюэ* ‘астрономия’ вторая биграмма совпадает со словом 文学 *вэньсюэ* ‘литература’. Однако чаще – в силу специфики китайского словообразования – вторая биграмма может репрезентировать родовое понятие: например, биграмма 织物 *чжиу* совпадает со словом ‘ткань’ и является частью слов 棉织物 *мяньчжиу* ‘хлопчатобумажные изделия’ и 针织物 *чжэньчжиу* ‘трикотаж’. Эта интересная особенность объясняет ещё одно наблюдение авторов: каждое слово, содержащее биграмму, имеет меньшую частоту, чем сама биграмма [17, p. 549–550].

Таким образом, в настоящем исследовании объектом выступает совокупность (коллекция) текстов, относящихся к китайскому военному дискурсу, предметом – иероглифические биграммы, их лексические и синтаксические свойства в изучаемых текстах, а основным методом – сплошной частотный анализ таких биграмм.

2. Материалы и методы

Языковым материалом исследования послужили информационные сообщения военного раздела сайта китайской государственной службы новостей «Чжунсинь» (*www.chinanews.com/mil*), публикующего новости ведущих информационных агентств и изданий Китая («Синьхуа», «Жэньминь жибао», «Цзефанцзюнь бао» и др.) и имеющего доступный онлайн архив новостей. Для текущего исследования методом сплошной выборки были взяты все публикации военного раздела сайта за третий квартал 2018 года, что составило 560 текстовых сообщений общим объёмом 720708 знаков (иероглифов и знаков препинания).

Далее с помощью специальной компьютерной программы подсчитывалась частота совместного употребления последовательностей иероглифов в тексте. Как уже упоминалось, с учётом преимущественно двуморфемной нормы современного китайского слова в рамках текущей работы мы ограничились данными о частотах пар иероглифов – иероглифических биграмм.

Полученные биграммы проверялись на наличие в словаре. Использовалась электронная версия «Нового китайско-русского словаря» под редакцией А.В. Котова («Русский язык-Медиа», 2004), входящая в комплект электронного словаря ABBYY Lingvo x3. Это наиболее новый официальный словарь китайского языка, доступный в электронном виде. Биграммы, соответствие которым в данном словаре не было найдено, определены в настоящей работе как несловарные.

Для всех биграмм помимо абсолютной частоты (фактического количества появлений в коллекции текстов) подсчитывалась частота в пересчёте на миллион употреблений (ipm, instances per million).

3. Результаты

3.1 Частота совместной встречаемости

В общей сложности получилось 35370 биграмм с частотой употребления не менее двух раз. Из них словарных 6334, соответственно, несловарных – 29036. Словарные биграммы присутствуют во всех частях полученного частотного рейтинга, последняя из них находится на 35360 месте (это слово 齐全 *цицюань* ‘полностью’).

Практический интерес представляют в первую очередь частотные биграммы. Чётких критериев разделения частот на высокие, средние или низкие не существует. Е.И. Риехакайнен, изучая этот вопрос, приводит примеры различных работ, в которых исследователи на своё усмотрение определяли границу высокой частоты на уровне 60, 80, 84 или 90 употреблений на миллион [7, с. 22–23]. Возьмём нижний из указанных порогов (ipm ≥ 60).

Таких биграмм получилось 1161, из них 467 не имеет соответствий в словаре (40 %). Соответственно, 60 % биграмм, выявленных по критерию фактической частоты совместного употребле-

ния иероглифов в коллекции текстов, оказались словарными словами, причём принадлежащими словарю общего назначения.

Первые три десятка самых частых иероглифических сочетаний приведены ниже [таблица 1]. Получившийся список представляется вполне репрезентативным, отражающим лексическое наполнение источника – китайских новостных сообщений военной тематики.

Как видно из таблицы 1, даже среди самых частых сочетаний имеются три слова, не зафиксированных в словаре:

1) 官兵 *гуаньбин* – пример копулятивного способа китайского словообразования, буквально эти иероглифы означают ‘офицеры [и] солдаты’, что вместе составляет собирательное понятие ‘военнослужащие’;

2) 一个 *игэ* – числительное ‘один’ и универсальное счётное слово (≈ ‘штука’), вместе эти иероглифы лексически составляют полноценное слово с количественным значением ‘один’, но функционально часто используются в грамматическом значении, аналогичном неопределённому артиклю;

Таблица 1.

Наиболее частые иероглифические биграммы в новостных сообщениях военной тематики[3]

№ п/п	Биграмма	Чтение	Перевод	Наличие в словаре[4]	Частота в коллекции	Частота в ipm
1.	中国	*чжунго*	Китай	да	2009	2787,5
2.	训练	*сюньлянь*	тренировка	да	1110	1540,2
3.	部队	*будуй*	войска, части	да	1061	1472,2
4.	军事	*цзюньши*	военный	да	970	1345,9
5.	官兵	*гуаньбин*	военнослужащие	нет	848	1176,6
6.	工作	*гунцзо*	работа	да	784	1087,8
7.	一个	*игэ*	один	нет	768	1065,6
8.	他们	*тамэнь*	они	да	763	1058,7
9.	我们	*вомэнь*	мы	да	741	1028,2
10.	记者	*цзичжэ*	корреспондент	да	724	1004,6
11.	美国	*мэйго*	США	да	711	986,5
12.	任务	*жэньу*	задача	да	708	982,4
13.	进行	*цзиньсин*	проводить	да	691	958,8
14.	比赛	*бисай*	соревнования, игры	да	675	936,6
15.	作战	*цзочжань*	(вести) боевые действия	да	662	918,5
16.	战斗	*чжаньдоу*	бой	да	650	901,9
17.	海军	*хайцзюнь*	военно-морские силы	да	634	879,7
18.	军人	*цзюньжэнь*	военнослужащий	да	623	864,4
19.	国防	*гофан*	оборона страны	да	610	846,4
20.	飞行	*фэйсин*	полёт	да	570	790,9
21.	来源	*лайюань*	источник	да	566	785,3
22.	军队	*цзюньдуй*	войска	да	566	785,3
23.	导弹	*даодань*	управляемая ракета	да	564	782,6
24.	装备	*чжуанбэй*	вооружение	да	539	747,9
25.	空军	*кунцзюнь*	военно-воздушные силы	да	532	738,2
26.	参赛	*цаньсай*	участвующие в играх	нет	523	725,7
27.	国家	*гоцзя*	государство	да	515	714,6
28.	报道	*баодао*	доклад; сообщать	да	511	709,0
29.	第一	*ди и*	первый	да	492	682,7
30.	组织	*цзучжи*	организация	да	488	677,1

[3] Источник – коллекция текстов всех публикаций сайта *www.chinanews.com/mil* в период с 01.07.2018 по 30.09.2018.

[4] Новый китайско-русский словарь: около 4100 иероглифов и свыше 26000 слов и лекс. словосочетаний. Котов А. В. (ред.). М.: Русский язык Медиа, 2004. Доступ из электронного словаря ABBYY Lingvo.

3) 参赛 *цаньсай* – пример морфемной контракции, легко образуемого сокращения со значением ‘участвующие в [военных] играх’ от полных форм 参加 *цаньцзя* ‘участвовать’ и 比赛 *бисай* ‘соревнования, игры’.

На наш взгляд, это вполне показательные примеры специфики китайского лексического уровня. Первый и третий наглядно демонстрируют упоминавшуюся выше лёгкость соединения в целое и разделения на части; второй пример показывает, как грамматические значения в условиях бедной китайской морфологии распределяются по другим языковым уровням. Притом что эти три лексические единицы можно с достаточным основанием считать высокочастотными словами, они не зафиксированы в словаре, и это характеризует не столько качество конкретного лексикографического источника, сколько неохватную вариативность лексического уровня китайского языка.

В целом оказалось, что критерий частоты достаточно эффективно выявляет словарные слова: уровень «словарности» в первой десятке составил 80 %, в первой сотне – 76 %, среди биграмм с частотностью свыше 100 употреблений на миллион – 66 % (и 60 % среди биграмм с ipm ≥ 60, о чём уже говорилось выше).

Оставшиеся 40 % несловарных биграмм можно разделить на три основные группы: 1) *лексические единицы* (133 биграммы: 11,5 % среди всех частотных, 28,5 % среди несловарных); 2) *синтаксические конструкции* (198 биграмм: 17,1 %; 42,4 %); 3) *синтагматические фрагменты* (136 биграмм: 11,7 %; 29,1 %).

1. Несловарные *лексические единицы*, в свою очередь, подразделяются на следующие категории:

– знаменательные слова или словосочетания, такие как приводившееся выше 官兵 *гуаньбин* ‘военнослужащие’ (16,3 % от несловарных);

– сокращения, полученные в результате морфемной контракции, такие как 参赛 *цаньсай* ‘участвующие в играх’ (6,2 %);

– служебные слова (числительные и различные неличные местоимения), такие как 一个 *игэ* ‘один’ или 这一 *чжэи* ‘этот’ (5,6 %);

– имена собственные: таких (из двух иероглифов) встретилось всего два (0,4 %).

2. Под *синтаксическими конструкциями* здесь подразумеваются различные не имеющие лексической цельности, но синтаксически значимые комбинации грамматических (служебных, дискурсивных) элементов с частотными словообразовательными элементами либо друг с другом:

– собственно синтаксические конструкции 这是 *чжэ ши* ‘это есть’ (аналог английского *this is*), 也是 *е ши* ‘также является’ и т.п. (9 % от несловарных);

– незаконченные сочетания с грамматическими элементами, такими как определительная частица 的 *дэ* (типа 的军 *дэцзюнь* ‘[атрибутив/генитив] + войска’ либо 军的 *цзюньдэ* ‘войска + [атрибутив/генитив]’), показатель места 在 *цзай* (在这 *цзай чжэ* ‘в этом…’), связка 是 *ши* (是一 *ши и* – аналог английского *is a*)[5] (33,4 % от несловарных).

3. К *синтагматическим фрагментам* мы отнесли различные последовательности иероглифов, не имеющие лексического или грамматического значения, которые принадлежат более длинным лексическим единицам и примерно показывают их долю в частотном списке, не ограниченном сочетаниями только двух иероглифов:

– внутрисловные биграммы от «обычных» слов: 放军 *фанцзюнь* от 解放军 *цзефанцзюнь* ‘освободительная армия’, 斗机 *доуцзи* от 战斗机 *чжаньдоуцзи* ‘истребитель’ и т.п. (6,9 % от несловарных), а также внутрисловные биграммы, представляющие фрагменты имён собственных: 俄罗 *эло* от 俄罗斯 *элосы* ‘Россия’, 习近 *си цзинь* от习近平 ‘Си Цзиньпин’ (ещё 6 %);

– межсловные биграммы: 国海 *го хай* от 中国海军 *чжунго хайцзюнь* ‘ВМС Китая’, 际军 *цзи цзюнь* от 国际军事 *гоцзи цзюньши* ‘международные военные [игры]’ и т.п. (16,3 %).

5 Подобные сочетания, выделяемые на основе высокой частотности, называют в различных работах «лексическими пучками», «кластерами» и отмечают, что они «могут быть структурно незаконченными» [3, с. 75]. Ли Шоуцзи и Го Шулунь относят их к издержкам подхода с опорой на данные о частотности (data-driven approach), названного у нас выше «объективным» [16, p. 57].

В итоге опора на частоту совместной встречаемости пар иероглифов позволяет выявить суммарно 71,2 % частотных лексических единиц (словарных и несловарных), а также обнаружить наиболее распространённые в исследуемом дискурсе синтаксические конструкции и их элементы. Кроме того, внутрисловные и межсловные биграммы тоже указывают нам на конкретные лексические единицы данного дискурса, с их учётом доля частотной лексики в общих результатах повысится до 82,9 %.

Содержательно полученный частотный список представляется вполне репрезентативным. Приведённая в таблице 1 его первая часть, а также остальные не показанные в данной работе пункты этого списка соответствуют основанным на личном опыте субъективным ожиданиям.

Таким образом, выделение наиболее репрезентативной лексики тематического дискурса на основе критерия фактической частоты встречаемости в коллекции текстов можно считать хорошим методом, дающим ожидаемый результат.

Заключение

В настоящей работе рассматривался вопрос возможности выделения китайских лексических единиц, записываемых двумя иероглифами, на основе частоты совместной встречаемости этих иероглифов. Сложность этого вопроса применительно к китайскому языку определяется отсутствием явного разделения текста на слова и трудностью однозначной автоматической идентификации китайского слова. Ограниченность объекта исследования достаточно специфическим профессиональным дискурсом с требованием обязательной актуальности текстов не оставляет возможности обращаться к существующим большим сегментированным корпусам. Поэтому исследование проводилось на материале коллекции текстов.

Сплошной частотный анализ иероглифических биграмм показал себя хорошим методом, позволяющим среди частых результатов выявить 71,2 % лексических единиц, в том числе 60 % словарных единиц и более 11 % новых слов, выражений и сокращений, отличающих лексику предметной области от словаря общего назначения. Помимо них такой метод обнаруживает синтаксические конструкции (17,1 %), в основном структурно незавершённые, и различные синтагматические фрагменты более крупных единиц и конструкций (ещё 11,7 %). Последняя категория приобретёт характер осмысленных лексических единиц при рассмотрении сочетаний более чем двух иероглифов.

Таким образом, высокая частота совместной встречаемости иероглифов может с достаточно большой вероятностью ($p > 0,7$) рассматриваться как показатель лексичности при выявлении репрезентативной лексики в несегментированной тематической коллекции текстов на китайском языке. Однако дальнейший отбор лексики в прикладных целях, отделение её от синтаксических или просто синтагматических сочетаний исследователь должен выполнять «вручную».

Список литературы

1. Алексахин А.Н. Алфавит китайского языка путунхуа. Буква – фонема – звук речи – слог – слово. 4-е изд., испр. и доп. М.: Восточная книга, 2018. 212 с.
2. Алексахин А.Н. Современная политика КНР в отношении иероглифической и буквенной письменности // Вестник МГИМО. 2011. № 3. С. 243–252.
3. Горина О.Г. Использование технологий корпусной лингвистики для развития лексических навыков студентов-регионоведов в профессионально-ориентированном общении на английском языке: дисс. … канд. пед. наук. М.: МГУ, 2014. 321 с.
4. Кленин И.Д. Лексикология китайского языка / И.Д. Кленин, В.Ф. Щичко. М.: Восточная книга, 2013. 272 с.
5. Курдюмов В.А. Динамический подход к научному изучению китайского языка // III Готлибовские чтения: Востоковедение и регионоведение Азиатско-Тихоокеанского региона в фокусе современности : материалы Междунар. науч. конф. Иркутск, 10–16 сент. 2019 г. / ФГБОУ ВО «ИГУ»; [отв. ред. Е.Ф. Серебренникова]. Иркутск: Изд-во ИГУ, 2019. С. 285–291.

6. Муравьёв Н.А. Подходы к составлению лексических минимумов в России и за рубежом: проблемы и перспективы / Н.А. Муравьёв, М.Ю. Ольшевская // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2019. Т. 17, № 1. С. 78–89.
7. Риехакайнен Е.И. Восприятие русской устной речи: контекст + частотность: монография. СПб.: С.-Петерб. гос. ун-т, 2016. 270 с.
8. Солнцев В.М. Язык как системно-структурное образование. Изд. 2-е, доп. М.: Наука, 1977.
9. Солнцев В.М. Теоретическая грамматика современного китайского языка (проблемы морфологии): курс лекций / В.М. Солнцев, Н.В. Солнцева. М.: Военный институт, 1978.
10. Шейгал Е.И. Семиотика политического дискурса: дисс. ... д-ра филол. наук. Волгоград, 2000. 431 с.
11. Шемет Г.И. Совершенствование обучения иностранному языку курсантов военных вузов на основе оптимизации лексической компоненты: дисс. … канд. пед. наук. М.: Военный университет, 2011. 249 с.
12. Юй Чуцяо. Автоматический синтаксический анализ китайских предложений при ограниченном словаре / Юй Чуцяо, И.А. Бессмертный // Программные продукты и системы. 2017. Т. 30. № 1. С. 138–142.
13. Da Jun. 2004. Chinese text computing. [Electronic resource] – URL: http://lingua.mtsu.edu/chinese-computing (accessed: 23.03.2020)
14. Deng K. On the unsupervised analysis of domain-specific Chinese texts / K. Deng, P.K. Bol, K.J. Li et al. // Proceedings of the National Academy of Sciences of the United States of America, 2016. Vol. 113, No. 22. Pp. 6154–6159.
15. Huang W. Toward Fast and Accurate Neural Chinese Word Segmentation with Multi-Criteria Learning / W. Huang, X. Cheng, K. Chen et al. [Electronic resource] Cornell University > Computer Science > Computation and Language. – URL: https://arxiv.org/abs/1903.04190 (accessed: 22.03.2020)
16. Li Sh. Collocation Analysis Tools for Chinese Collocation Studies / Sh. Li, Sh. Guo // Journal of Technology and Chinese Language Teaching. Vol. 7, No. 1, 2016. Pp. 56–77.
17. Li J. A Comparison and Semi-Quantitative Analysis of Words and Character-Bigrams as Features in Chinese Text Categorization / J. Li, M. Sun, X. Zhang // Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, July 2006. Pp. 545–552.
18. Ma J. State-of-the-art Chinese Word Segmentation with Bi-LSTMs / J. Ma, K. Ganchev, D. Weiss // In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, October 31 – November 4, 2018. Pp. 4902–4908.

References

1. Aleksakhin, A.N. *Alfavit kitayskogo iazyka putunkhua. Bukva – fonema – zvuk rechi – slog – slovo* [The Chinese language alphabet of Putonghua. Letter – phoneme – speech sound – syllable – word]. 4th ed., rev. and add. M.: Vostochnaia kniga, 2018. 212 p. (In Rus.)
2. Aleksakhin, A.N. Sovremennaia politika KNR v otnoshenii ieroglificheskoi i bukvennoi pis'mennosti [Modern policy of the PRC in relation to hieroglyphic and alphabetic writing] // *MGIMO Bulletin,* 2011. No. 3. Pp. 243–252. (In Rus.)
3. Gorina, O.G. *Ispol'zovanie tekhnologii korpusnoi lingvistiki dlia razvitiia leksicheskikh navykov studentov-regionovedov v professional'no-orientirovannom obshchenii na angliiskom iazyke* [The use of corpus linguistics technologies for the development of the lexical skills of regional students in professionally oriented communication in English]. PhD thesis. Moscow: Moscow State University, 2014. 332 p. (In Rus.)
4. Klenin, I.D. *Leksikologiia kitaiskogo iazyka* [Lexicology of the Chinese language] / I.D. Klenin, V.F. Shchichko. M.: Vostochnaia kniga, 2013. 272 p. (In Rus.)
5. Kurdyumov, V.A. *Dinamicheskii podkhod k nauchnomu izucheniiu kitaiskogo iazyka* [Dynamic approach to exploring Chinese]. III Gotlibovskie chteniia: Vostokovedenie i regionovedenie Aziatsko-Tikhookeanskogo regiona v fokuse sovremennosti: materialy Mezhdunar. nauch. konf. Irkutsk, 10–16 Sent. 2019. FGBOU VO «IGU»; [otv. red. Ie. F. Serebrennikova]. Irkutsk: Izd-vo IGU, 2019. Pp. 285–291. (In Rus.)
6. Murav'ëv, N.A. Podkhody k sostavleniiu leksicheskikh minimumov v Rossii i za rubezhom: problemy i perspektivy [Approaches to the composition of lexical minima in Russia and abroad: problems and prospects] / N.A. Murav'ev, M. Iu. Olshevskaia. *Vestnik NSU. Series: Linguistics and Intercultural Communication,* 2019, 17 (1). Pp. 78–89. (In Rus.)
7. Riehakainen, E.I. *Vospriiatie russkoi ustnoi rechi: kontekst + chastotnost'* [Perception of Russian spoken language: context + frequency]. Monograph. St. Petersburg: St. Petersburg. State Univ., 2016. 270 p. (In Rus.)
8. Solntsev, V.M. *Iazyk kak sistemno-strukturnoe obrazovanie* [Language as a systemic-structural formation]. Ed. 2nd, add. M .: Nauka, 1977. (In Rus.)
9. Solntsev, V.M. *Teoreticheskaia grammatika sovremennogo kitaiskogo iazyka (problemy morfologii)* [Theoretical grammar of modern Chinese (problems of morphology)]. Course of lectures, Moscow: Military Institute, 1978. (In Rus.)
10. Sheigal, E.I. *Semiotika politicheskogo diskursa* [Semiotics of Political Discourse]. Doctoral Thesis. Volgograd, 2000. 431 p. (In Rus.)
11. Shemet, G.I. *Sovershenstvovanie obucheniia inostrannomu iazyku kursantov voennykh vuzov na osnove optimizatsii leksicheskoi komponenty* [Improving the teaching of a foreign language to cadets of military colleges based on the optimization of the lexical component]. PhD thesis. Moscow: Military University, 2011. 249 p. (In Rus.)
12. Yu, Chuqiao. Avtomaticheskii sintaksicheskii analiz kitaiskikh predlozhenii pri ogranichennom slovare [Automatic syntactic analysis of Chinese sentences with a restricted dictionary] / Yu Chuqiao, I.A. Bessmertnyi // *Programmnye produkty i sistemy.* 2017. 30 (1). Pp. 138–142. (In Rus.)
13. Da, Jun. 2004. *Chinese text computing,* lingua.mtsu.edu/chinese-computing (accessed: 23.03.2020)

14. Deng, K. On the unsupervised analysis of domain-specific Chinese texts / K. Deng, P.K. Bol, K.J. Li et al. // *Proceedings of the National Academy of Sciences of the United States of America,* 2016. Vol. 113, No. 22. Pp. 6154–6159.
15. Huang, W. Toward Fast and Accurate Neural Chinese Word Segmentation with Multi-Criteria Learning / W. Huang, X. Cheng, K. Chen et al. [Electronic resource] Cornell University > Computer Science > Computation and Language, arxiv.org/abs/1903.04190 (accessed: 22.03.2020)
16. Li, Sh. Collocation Analysis Tools for Chinese Collocation Studies / Sh. Li, Sh. Guo // J*ournal of Technology and Chinese Language Teaching.* Vol. 7, No. 1, 2016. Pp. 56–77.
17. Li, J. A Comparison and Semi-Quantitative Analysis of Words and Character-Bigrams as Features in Chinese Text Categorization / J. Li, M. Sun, X. Zhang // *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics,* Sydney, July 2006. Pp. 545–552.
18. Ma, J. State-of-the-art Chinese Word Segmentation with Bi-LSTMs / J. Ma, K. Ganchev, D. Weiss // *In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.* Brussels, Belgium, October 31 – November 4, 2018. Pp. 4902–4908.

Сведения об авторе:

Коршунов Дмитрий Сергеевич – кандидат филологических наук, научно-педагогический работник ВУРЭ (Россия, Череповец). Сфера научных и профессиональных интересов: китайский военный дискурс, лексический уровень китайского языка, восприятие иероглифического текста, лингвострановедение, военный перевод.
E-mail: dmitry-korshunov@yandex.ru
Конфликт интересов: Автор заявляет об отсутствии конфликта интересов.

About the author:

Dmitry S. Korshunov – PhD (philology), scientific and pedagogical employee, MURE (Cherepovets, Russia). Spheres of research and professional interest: Chinese military discourse, Chinese lexical level, character text perception, country study and military translation.
E-mail: dmitry-korshunov@yandex.ru.
Conflicts of interest: The author declares absence of conflicts of interest.

* * *