Научная статья

УДК 81.32 + 811.581.11 + 81'1
DOI 10.25205/1818-7935-2022-20-2-64-80

Особенности применения статистических мер в задачах выделения китайских иероглифических биграмм

Дмитрий Сергеевич Коршунов

Военный университет радиоэлектроники
Череповец, Россия
dmitry-korshunov@yandex.ru, https://orcid.org/0000-0002-1550-5904

Аннотация

Для изучения современной лексики определенной профессиональной тематики есть возможность создавать коллекции текстов и применять к ним программные средства лингвистического анализа. Однако существует проблема качества автоматической сегментации китайского текста на слова. Одним из способов выделения в китайском тексте лексических единиц является применение статистических мер выделения коллокаций к иероглифическим биграммам. Цель настоящей работы заключается в проведении сопоставительного анализа семи разных статистических мер оценки коллокаций как средства выделения двусложных лексических единиц (биномов) в несегментированном иероглифическом тексте на китайском языке. Предметом анализа являются лексико-грамматические и частотные характеристики биграмм, имеющих наибольшие значения рассматриваемых статистических мер. Их сопоставление позволяет сделать вывод об особенностях статистических мер, в частности о том, каким лингвистическим задачам какая мера лучше соответствует. Языковым материалом исследования послужила коллекция из 560 новостных текстов военной тематики на китайском языке объемом более 720 тысяч знаков. Результаты показывают, что рассмотренные статистические меры можно разделить на три группы по тому, какие характеристики биграмм получают наибольшие значения. К первой группе относятся меры MI, MS и logDice, которые дают приоритет редким биграммам с ограниченной сочетаемостью компонентов, таким как китайские двусложные одноморфемные слова «ляньмяньцзы». Эти меры плохо выделяют термины, но могут использоваться для поиска фразеологически связанных компонентов. Меры второй группы, t-score и log-likelihood, ориентированы на частотность, близки к анализу по частоте, но лучше него справляются с нелексическими биграммами, при этом log-likelihood несколько понижает ранг числительных и местоимений, лучше всех выделяя именно характерную для профессионального дискурса лексику. К третьей группе относятся меры MI^3 и MI.log-f, которые усредняют противоположные подходы первых двух групп. Мера MI^3 оценивается как наиболее универсальная, она могла бы использоваться для сравнения различных корпусов или коллекций текстов. Делается вывод, что использование статистических мер в отношении иероглифических биграмм возможно и целесообразно при учете соответствия их специфики исследовательской задаче.

Ключевые слова

китайский язык, бином, иероглифические биграммы, коллокации, статистические меры, коллекция текстов

Для цитирования

Коршунов Д. С. Особенности применения статистических мер в задачах выделения китайских иероглифических биграмм // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2022. Т. 20, № 2. С. 64–80. DOI 10.25205/1818-7935-2022-20-2-64-80

Distinctive Features of Association Measures Applied to Chinese Character Bigram Extraction Tasks

Dmitry S. Korshunov

Military University of Radio Electronics
Cherepovets, Russian Federation
dmitry-korshunov@yandex.ru, https://orcid.org/0000-0002-1550-5904

Abstract

Studying professional discourse, a researcher has now an opportunity to create collections of texts and apply linguistic analysis software tools to them. However, when it comes to Chinese discourse there is a problem with the reliability of automatic word segmentation of texts. One of the ways to extract lexical units in Chinese texts is to apply statistical association measures for collocations to Chinese character bigrams. The purpose of this work is to conduct a comparative analysis of seven different statistical measures for collocations as a means of extracting two-syllabic lexical units (binomes) in an unsegmented Chinese character text. The subject of the analysis is the lexical, grammatical and frequency characteristics of bigrams with higher values of the statistical measures. Their comparison makes it possible to draw a conclusion about the features of statistical measures, in particular, about the best correspondence of linguistic tasks to statistical measures. The linguistic material of the study was a collection of 560 military-related news texts in Chinese with more than 720 thousand characters. The results show that the statistical measures considered can be divided into three groups according to the characteristics of bigrams receiving the highest values. The first group includes measures MI, MS and logDice, which give priority to rare bigrams with limited compatibility of components, such as the Chinese two-syllable single morpheme words “lianmianzi”. These measures do not extract terms well, but can be used to search for phraseologically related components. The measures of the second group, t-score and log-likelihood, are frequency-oriented, similar to frequency analysis, but they cope with non-lexical bigrams better, while log-likelihood somewhat lowers the rank of numerals and pronouns, picking out best the typical vocabulary of professional discourse. The third group includes measures MI^3 and MI.log-f, which average the opposite approaches of the first two groups. The MI^3 measure is considered to be the most universal one; it could be used to compare different corpora or collections of texts. It is concluded that applying statistical association measures to Chinese character bigrams is possible and appropriate, when taking into account the correspondence of their specifics to a research task.

Keywords

Chinese language, binome, Chinese character bigrams, collocations, statistical association measures, text collection

For citation

Korshunov, D. S. Distinctive Features of Association Measures Applied to Chinese Character Bigram Extraction Tasks. *Vestnik NSU. Series: Linguistics and Intercultural Communication*, 2022, vol. 20, no. 2, pp. 64–80. (in Russ.) DOI 10.25205/1818-7935-2022-20-2-64-80

Введение

Описание проблемы

С развитием электронных средств массовой коммуникации у лингвистов появилась возможность не только обращения к большим корпусам текстов, но и формирования собственных коллекций текстов под частные задачи, к примеру, для изучения современной лексики определенной профессиональной тематики, создания лексических минимумов по языку специальности [Власова и др., 2019]. Для работы с такими коллекциями текстов доступны различные программные средства лингвистического анализа, такие как AntConc [1], Sketch Engine [2] и др. Они способны работать не только с английским, но и со многими другими языками, включая китайский.

Однако структурно-типологические особенности китайского языка и его системы письма представляют определенную проблему для автоматической обработки. Если в языках, например, с фонетическим письмом статистическим вычислениям в любом тексте предшествует его сравнительно простая в реализации сегментация на слова, то для китайского языка эта процедура является нетривиальной задачей – как известно, китайская письменность не

[1] https://laurenceanthony.net/software/antconc/
[2] https://www.sketchengine.eu/

предусматривает разделения слов пробелами. В таком случае сегментация на слова выполняется искусственно, для чего также существуют программные средства [3]. Они действуют на основе принятых разработчиками критериев, которые не всегда одинаковы, что, естественно, приводит к неоднозначным решениям: расхождениям в выделении слов в одном и том же предложении и, следовательно, в числе слов в одном и том же предложении.

Так, в работе [Chen et al., 2017] приводится пример различной сегментации в двух крупных китайских корпусах текстов предложения 姚明进入总决赛 *yáo míng jìnrù zǒng juésài* «Яо Мин выходит в финал», где неодинаково трактуются имя собственное Яо Мин и существительное ‘финал’ 总决赛 *zǒng juésài* (букв.: ‘главный решающий матч’). В одном корпусе оба комплекса идентифицируются как целые слова, в другом корпусе ИС раскладывается на фамилию и имя, а в составе лексического комплекса 总决赛 *zǒng juésài* (русс. ‘финал’) выделяется отдельная лексема 总 *zǒng* ‘главный’. Оба варианта могут быть рационально аргументированы в формально-семантическом плане.

Неоднозначность сегментации китайского текста на лексемы снижает качество и надежность получаемых из корпуса данных, поэтому исследователи ищут способы обойти эту структурно-семантическую проблему. Для ряда прикладных лингвистических задач разработаны компьютерные программы на основе нейронных сетей, которые, по заявлениям авторов, могут работать с опорой только на иероглифы, без сегментации текстов на слова, показывая при этом даже лучшие результаты (см., например, [Meng et al., 2019]).

Не прибегать к искусственной сегментации на лексемы в тех языках, где понятие слова элиминировано самим строем языка, решил и коллектив российских ученых, работающих над автоматизацией обработки тибетских письменных текстов. В них, как и в китайских, отсутствуют пробелы, а из-за грамматических особенностей тибетского языка «любое разбиение текста на словоформы оказывается <...> необоснованным» [Гроховский и др., 2019, с. 71]. Лексический уровень в тибетских языках приходится описывать через синтактику морфем, что возвращает нас к параллели с китайским языком, где слово, по выражению В. Б. Касевича, является всего лишь «частным случаем слогоморфемной синтагмы» [Касевич, 2011а, с. 392] [4]. А в корпусном исследовании Да Цзюня [Da, 2004] вместо сегментации текста на «слова» был выполнен частотный анализ иероглифических биграмм, т. е. текстовых последовательностей из двух иероглифов, независимо от семантической или грамматической связанности записываемых ими слогоморфем.

Да Цзюнь исходит из допущения, принятого в отношении коллокаций в европейских языках, что частое соупотребление двух компонентов не является случайным. Такие биграммы автор предлагает рассматривать «как близкое приближение к китайскому слову из двух иероглифов» (as a close approximation to a two-character word in Chinese) [Ibid., p. 505]. Это теоретическое решение, безусловно, имеет право на существование, поскольку известно, что роль так называемых биномов (сочетаний из двух слогоморфем) в организации китайского текста специфична и высока, а «категории слова в китайском языке принадлежит, скорее, периферийная позиция», при этом «грамматическая природа биномов может быть разной» [Касевич, 2011б, с. 616] [5]. Существенно для сопоставления и то, что «бином как целое в грамматическом и семантическом отношении часто аналогичен слову языков типа русского» [Там же, с. 619].

В китайском языке частотные биграммы могут совпадать со словарными и несловарными лексическими единицами, а также быть «внутрисловными» (intraword bigrams), когда, допустим, в слове из трех слогоморфем получаются две накладывающиеся биграммы (например, 锦标 и 标赛 из 锦标赛 *jǐnbiāosài* ‘чемпионат’), и «межсловными» (interword bigrams), «захва-

[3] Например, https://laurenceanthony.net/software/segmentant/

[4] Первая публикация в: Типология и грамматика. М., 1990. С. 67–72.

[5] Первая публикация в: Востоковедение. Филологические исследования / Отв. ред. проф. В. Г. Гузев, проф. О. Б. Фролова. СПб., 1993. Вып. 18. С. 66–77. В соавторстве с В. В. Рыбиным, Е. М. Шабельниковой.

тывая» по слогоморфеме от смежных слов (например, 气预 из 天气预报 *tiānqì yùbào* ‘прогноз погоды’) [Li Jingyang et al., 2006, p. 549]. Кроме того, биграммы могут совпадать с синтаксическими конструкциями (这是 *zhè shì* ‘это [есть]’), их смежными элементами (了一 *le yī*) и т. п.

Поскольку в ходе частотного анализа иероглифических биграмм выделяются не только биномы, но и довольно большое количество сочетаний, не совпадающих с лексическими единицами, для отделения значимых сочетаний от незначимых Да Цзюнь использовал статистические меры определения силы ассоциативной связанности элементов (measure of the strength of association). Для всех биграмм своего корпуса он рассчитал значение коэффициента взаимной информации MI (mutual information) [Church, Hanks, 1990], предположив, что биграммы с частотой более 50 вхождений в 14-миллионном корпусе и значением $MI \geq 3{,}5$ будут «хорошими кандидатами» на статус двусложных китайских слов [Da, 2004, p. 508]. Показательно, однако, что, указывая значение коэффициента для каждой биграммы, автор воздерживается от выводов об их лексичности.

Коэффициент взаимной информации и другие статистические меры были в свое время разработаны для выявления коллокаций – устойчивых словосочетаний, характер связи между компонентами которых в разных школах и традициях понимается по-разному. В отечественной лингвистике этот термин нередко используется как некое переходное понятие между фразеологизмами и свободными словосочетаниями, удобное для изучения, например, лексических функций [Иорданская, Мельчук, 2007; Грудева, Тиханович, 2014, с. 15–59]. С развитием статистических методов исследования коллокации стали чаще рассматриваться как любое «неслучайное сочетание двух и более лексических единиц», характерное как для языка в целом, так и для определенной выборки текстов [Ягунова, Пивоварова, 2010, с. 30]. М. В. Влавацкая, делая краткий обзор становления термина «коллокация» и интерпретаций его понимания, подчеркивает значимость этого понятия в лингвистической науке: «В глобальном смысле коллокация лежит в основе всего языкового использования» [2019, с. 439].

Идея применить концепцию коэффициента взаимной информации к иероглифическим биграммам для сегментации китайского текста впервые возникла еще в 1990 г. у американских исследователей [Sproat, Shih, 1990]. Преодолеть некоторые ограничения их подхода попытался коллектив китайских ученых, использовавших одновременно коэффициент MI и меру t-score [孙茂松 / Сунь Маосун и др., 1997]. В своей англоязычной работе часть этого авторского коллектива впервые в явном виде предложила под переменными в соответствующих статистических формулах понимать не слова, а китайские иероглифы [6]. Поскольку в китайском языке почти каждая записываемая иероглифом слогоморфема обладает свойствами единицы лексического уровня, любое неслучайное сочетание двух и более слогоморфем может считаться коллокацией.

Статистические меры

В настоящее время разработано порядка 80 статистических мер, позволяющих оценить силу связанности лексических единиц [Хохлова, 2017, с. 349]. Наиболее распространенными из них, судя по литературе, являются коэффициент взаимной информации MI и t-score.

Именно меры MI и t-score (по отдельности или вместе) в обязательном порядке присутствуют в программных средствах для поиска коллокаций в различных корпусах китайских текстов [Li Shouji, Guo Shulun, 2016, p. 65].

Каждая мера имеет свои достоинства и недостатки. На материале новостных текстов на русском языке Е. В. Ягунова и Л. М. Пивоварова делают вывод, что «коллокации, выделяемые с помощью MI, отражают предметную область», «MI наилучшим образом позволяет вы-

[6] В оригинале: “We adopt these measures almost completely here, with one major modification: the variables in two relevant formulae are no longer *words* but *Chinese characters*” [Sun et al., 1998, p. 1266] (курсив авторов. – *Д. К.*).

делять наименования объектов, термины, сложные номинации; t-score, напротив, лучше работает при выделении “общеязыковых устойчивых сочетаний” (производных служебных слов, дискурсивных слов) и “устойчивых конструкций”» [Ягунова, Пивоварова, 2010, с. 37]. С учетом этого китайский специалист Дэн Яочэнь предлагал в практических исследованиях коллокаций использовать эти меры вместе [邓耀臣 / Дэн Яочэнь, 2003, с. 77].

К особенностям меры MI относится ее способность находить в корпусе редкие словосочетания [Хохлова, 2017, с. 350], сильно завышая, однако, их значимость, что в большинстве случаев следует расценивать как недостаток. «Чем более редки слова, образующие коллокацию, тем выше будет для них значение MI, что делает данную меру совершенно “беззащитной” перед опечатками, иностранными словами и другим информационным шумом, который неизбежен в большой коллекции. Поэтому для данной меры используется порог отсечения по частоте» [Ягунова, Пивоварова, 2010, с. 33–34]. Ван Сугэ с коллегами ценят способность меры MI находить произвольную связь между компонентами коллокации [王素格 / Ван Сугэ и др., 2006, с. 35], но коллектив исследователей из провинции Хубэй отмечает, что простой анализ по частоте обнаруживает более осмысленные и эффективные коллокации, чем мера MI [Lan Huang et al., 2017, p. 31].

Что касается меры t-score, то она представляет собой лишь несколько модифицированное ранжирование коллокаций по частоте, что делает данную меру малопригодной, например, для поиска терминологических словосочетаний [Ягунова, Пивоварова, 2010, с. 33–34]. Вместе с тем разработчики известной онлайн-системы средств корпусного анализа Sketch Engine отмечают, что «в большинстве случаев мера t-score более надежна или более полезна, чем мера MI» [7].

Для преодоления недостатков названных мер был разработан ряд их модификаций. В исследованиях М. В. Хохловой, В. П. Захарова были отобраны и на материале ряда существительных русского языка сравнивались семь статистических мер: MI, log-likelihood (LL), t-score, MI^3, minimum sensitivity (MS), logDice и MI.log-f. В работе М. В. Хохловой были описаны некоторые особенности этих мер, но общий вывод указывал на их относительную взаимозаменяемость и необходимость дальнейшей экспертной оценки [Хохлова, 2017]. В. П. Захаров предлагал, в частности, методику интегрированной оценки коллокатов с помощью семи рассматриваемых статистических мер, а также оценивал степень соответствия результатов их применения оценкам экспертов [Zakharov, 2017a; 2017b].

Ниже приводятся формулы расчета указанных мер, взятые из работ [Захаров, Хохлова, 2010; Хохлова, 2017] [8]:

1) $$\mathrm{MI} = \log_2 \frac{f(n,c) \times N}{f(n) \times f(c)};$$

2) $$\text{t-score} = \frac{f(n,c) - \frac{f(n) \times f(c)}{N}}{\sqrt{f(n,c)}};$$

3) $$\text{log-likehood} = 2 \sum f(n,c) \times \log_2 \frac{f(n,c) \times N}{f(n) \times f(c)};$$

4) $$\mathrm{MI}^3 = \log_2 \frac{f^3(n,c) \times N}{f(n) \times f(c)};$$

[7] В оригинале: “In most cases, T-score is more reliable or more useful than MI Score” (https://www.sketchengine.eu/my_keywords/t-score).

[8] Формула меры log-likelihood для единообразия переменных приводится в форме, представленной не в самой работе [Захаров, Хохлова, 2010], а в одноименной презентации авторов, доступной онлайн: http://www.myshared.ru/slide/665373.

5) $\text{MI.log-}f = \text{MI} \times \ln\left(f\left(n,c\right)+1\right)$;

6) $\log \text{Dice} = 14 + \log_2 \frac{2f\left(n,c\right)}{f\left(n\right)\times f\left(c\right)}$;

7) $\text{MS} = \min\left\{\frac{f\left(n,c\right)}{f\left(n,c\right)\times f\left(n,\overline{c}\right)}, \frac{f\left(n,c\right)}{f\left(n,c\right)\times f\left(\overline{n},c\right)}\right\}$.

В этих формулах используются следующие переменные:

n – ключевое слово (в иероглифической биграмме – слогоморфема, записываемая первым иероглифом биграммы);

c – коллокат (слогоморфема, записываемая вторым иероглифом биграммы);

$f(n, c)$ – частота встречаемости ключевого слова *n* в паре с коллокатом *c* (частота встречаемости биграммы);

$f(n)$, $f(c)$ – абсолютные (независимые) частоты ключевого слова *n* и слова *c* в корпусе (тексте);

$f\left(n,\overline{c}\right)$ – частота встречаемости ключевого слова *n* без коллоката *c*;

$f\left(\overline{n},c\right)$ – частота встречаемости коллоката *c* без ключевого слова *n*;

N – общее число словоформ (иероглифов) в корпусе (тексте).

Цель настоящей работы – провести сопоставительный анализ применения разных статистических мер оценки коллокаций как средства выделения двусложных лексических единиц (биномов) в несегментированном иероглифическом тексте на китайском языке.

Предметом анализа являются лексико-грамматические и частотные характеристики биграмм, имеющих наибольшие значения рассматриваемых статистических мер. Их сопоставление позволит сделать вывод о тактике применения статистических мер, в частности о взаимосвязи между особенностью лингвистической задачи и соответствующей ее наилучшему достижению мерой.

1. Материалы и методы

В качестве языкового материала использовались информационные сообщения сайта Китайской государственной службы новостей (中国新闻), отражающего практику работы крупнейших информационных агентств и изданий Китая. В частности, из архива новостей военного раздела сайта [9] методом сплошной выборки были взяты все публикации за третий квартал 2018 г., что составило 560 текстовых сообщений общим объемом 720 708 знаков (иероглифов и знаков препинания).

Далее с помощью компьютерной программы, разработанной ранее нашими студентами, подсчитывалась частота отдельных иероглифов и совместного употребления последовательностей из двух иероглифов (иероглифических биграмм) в коллекции текстов. Частота биграмм вычислялась следующим образом: сочетание первого знака (иероглифа) со следующим справа знаком (иероглифом) проверялось по всей коллекции текстов на количество вхождений, далее операция повторялась для второго знака (иероглифа) в сочетании с третьим, третьего с четвертым и т. д. Затем биграммы, состоящие не из двух иероглифов (сочетания с цифрами, знаками препинания, латинскими буквами) удалялись. Помимо абсолютной частоты (т. е. фактического количества вхождений), для всех биграмм подсчитывалась частота в пересчете на миллион употреблений (ipm). Затем с помощью Microsoft Office Excel по приведенным выше формулам вычислялись и ранжировались значения отобранных семи ста-

[9] https://www.chinanews.com/mil/.

тистических мер. Далее анализировались частотные и лексико-грамматические характеристики биграмм, показавших наибольшие значения.

2. Результаты

Значения указанных выше статистических мер были рассчитаны для всех биграмм имеющегося языкового материала (35 370 биграмм с частотой появления в коллекции текстов не менее двух раз) и ранжированы в порядке убывания. Первые 20 результатов по каждой мере, а также по абсолютной частоте встречаемости биграмм в коллекции текстов приведены в табл. 1.

Таблица 1

Первые 20 биграмм, выделенные разными статистическими мерами

Table 1

The first 20 bigrams rated by different statistic measures

№	MI	MI^3	MI.log-f	logDice	t-score	MS	LL	Частота Frequency
1.	缥缈	训练	介绍	魔鬼	中国	魔鬼	中国	中国
2.	霹雳	记者	记者	牺牲	训练	牺牲	训练	训练
3.	叱咤	中国	牺牲	胳膊	部队	胳膊	记者	部队
4.	吭哧	介绍	什么	胞胎	军事	胞胎	部队	来源
5.	徜徉	任务	退役	絮叨	官兵	絮叨	官兵	军事
6.	愤恨	组织	组织	哔叽	工作	哔叽	任务	官兵
7.	憧憬	退役	研究	尴尬	他们	尴尬	工作	工作
8.	玻璃	官兵	指挥	橄榄	我们	橄榄	军事	一个
9.	珀玕	指挥	潜艇	盐碱	记者	盐碱	他们	他们
10.	羞涩	自己	训练	缥缈	一个	缥缈	我们	我们
11.	肆虐	技术	世界	霹雳	任务	霹雳	比赛	记者
12.	闵汝	什么	驾驶	叱咤	美国	叱咤	导弹	美国
13.	鞠躬	导弹	编辑	吭哧	进行	吭哧	进行	任务
14.	伉俪	工作	自己	徜徉	比赛	徜徉	组织	进行
15.	咳嗽	潜艇	包括	愤恨	战斗	愤恨	战斗	比赛
16.	嗅蔷	比赛	魔鬼	憧憬	作战	憧憬	来源	作战
17.	枯燥	研究	技术	玻璃	国防	玻璃	装备	战斗
18.	枸杞	问题	必须	珀玕	海军	珀玕	美国	海军
19.	琅琊	我们	问题	羞涩	飞行	羞涩	报道	军人
20.	翡翠	装备	医疗	肆虐	来源	肆虐	技术	国防

При попарном сравнении ста максимальных результатов по каждой мере обнаружены как значительные совпадения в составе лексических единиц, так и полное отсутствие совпадений. В табл. 2 приводятся данные о количестве одинаковых биграмм, выделенных парами разных статистических мер.

Таблица 2

Количество совпадающих биграмм между разными мерами
(из первых 100)

Table 2

The number of matching bigrams between different measures
(out of the first 100)

Меры Measures	MI	MI^3	MI.log-f	logDice	t-score	MS	LL	Частота Frequency	Всего Total
MI		0	0	57	0	66	0	0	123
MI^3	0		75	20	58	21	72	56	302
MI.log-f	0	75		24	34	23	47	32	235
logDice	57	20	24		10	85	11	10	217
t-score	0	58	34	10		10	82	91	285
MS	66	21	23	85	10		12	10	227
LL	0	72	47	11	82	12		77	301
Частота Frequency	0	56	32	10	91	10	77		276

Исходя из суммарного количества совпадений с другими мерами (последняя колонка табл. 2) наиболее универсальными оказались меры MI^3 и log-likelihood, а наиболее специфической – самая популярная изначально мера MI. Она имеет совпадения лишь с двумя мерами (logDice и MS, 57 и 66 случаев соответственно), и ни одного – с остальными пятью. Все другие меры в той или иной степени разделяют результаты между собой.

Как видно, наибольшее количество совпадений – 91 из 100 – наблюдается между анализом по частоте и мерой t-score, что подтверждает приведенный выше вывод Е. В. Ягуновой и Л. М. Пивоваровой об их схожести, согласно которому t-score является «лишь несколько модифицированным ранжированием коллокаций по частоте». Второй по близости парой являются меры MS и logDice (85 совпадений из 100), что также ожидаемо [10]. При этом пересечений между двумя парами мер очень мало – всего по 10. Эти пары обозначают два «полюса», два принципиально разных статистических подхода: частота и t-score «поднимают» ранг самых частых биграмм, а MS и logDice – самых уникальных, как правило, редких, компоненты которых реже всего употребляются друг без друга.

Рассмотрим на конкретных примерах разницу между этими двумя подходами и отдельными мерами.

[10] На сайте Sketch Engine основной характеристикой меры MS является ее похожесть на меру logDice: “a statistics measure similar to logDice” (https://www.sketchengine.eu/my_keywords/minimum-sensitivity).

3. Обсуждение

3.1. Меры MI, MS и logDice: приоритет ограниченной сочетаемости

Первые места в рейтинге этих трех мер занимают в основном биграммы, представляющие собой достаточно редкое [11] для китайского языка явление: так называемые «ляньмяньцзы» (联绵字, встречается также вариант 连绵词) – слитные, нечленимые двусложные слова. Как правило, используемые в них слоги имеют либо одинаковую инициаль, либо рифмующуюся финаль (хотя могут и не иметь таких признаков), а записывающие их иероглифы нередко содержат общий семантический элемент (ключ). Например: 缥缈 *piāomiǎo* ‘туманный’, 霹雳 *pīlì* ‘удар грома’, 徜徉 *chángyáng* ‘бродить, скитаться’, 愤恨 *fènhèn* ‘негодовать’, 玻璃 *bōli* ‘стекло’ и др. Это слова из первого десятка рейтинга меры MI и второго десятка рейтингов MS и logDice (см. табл. 1). Последние две меры полностью совпадают в своих первых 32 позициях и в первой десятке, несколько отличающейся от рейтинга MI, содержат такие слова, как 魔鬼 *móguǐ* ‘нечисть’, 牺牲 *xīshēng* ‘жертвовать’, 胞胎 *bāotāi* ‘плод, младенец в утробе матери’, 橄榄 *gǎnlǎn* ‘оливки’ и др. (см. табл. 1).

Очевидно, что это очень редкие, практически случайные для военной тематики слова [12]. Бороться с завышением значимости редких сочетаний в литературе предлагается установлением порога отсечения по частоте, конкретное значение которого определяется эмпирически. Мы последовательно проверяли разные значения частот, начиная с минимальных, пока не дошли до 60 ipm – порога, выше которого мы в данной работе считаем биграммы частотными. Результат получился лучше, но по-прежнему не очень убедительным (табл. 3).

В десятку первых по значению MI с указанным порогом отсечения вошли уже упомянутое 牺牲 *xīshēng* ‘жертвовать’, слова 贡献 *gòngxiàn* ‘вклад’, 威胁 *wēixié* ‘угроза’, 伙伴 *huǒbàn* ‘партнер’, а также шесть ИС или их фрагментов (в табл. 3 обозначены звездочкой), из которых только 辽宁 *liáoníng* (название провинции и авианосца) и 浙江 *zhèjiāng* (название провинции) можно отнести к распространенным словам.

Таблица 3

Первые 20 биграмм MI, MS и logDice с частотой более 60 ipm

Table 3

The first 20 bigrams of MI, MS and logDice with a frequency of 60+ ipm

№	MI	Частота, ipm Frequency, ipm	logDice	Частота, ipm Frequency, ipm	MS	Частота, ipm Frequency, ipm
1.	牺牲	112,4	牺牲	112,4	牺牲	112,4
2.	辽宁*	70,8	介绍	341,3	介绍	341,3
3.	沃洛*	79,1	什么	359,4	记者	986,5
4.	贡献	68,0	驾驶	177,6	驾驶	177,6
5.	孟泽*	83,3	记者	986,5	世界	301,1
6.	厚鑫*	90,2	退役	560,6	训练	1540,2
7.	吴谦*	155,4	包括	177,6	退役	560,6
8.	威胁	87,4	世界	301,1	厚鑫*	90,2

[11] По данным А. А. Хаматовой все двусложные морфемы, включая «ляньмяньцзы» и иностранные заимствования, составляют около 3 % всех морфем китайского языка [Хаматова, 2003, с. 32].

[12] Языковым материалом для анализа, как указано в п. 1, служила коллекция новостных текстов военной тематики.

Окончание табл. 3

№	MI	Частота, ipm Frequency, ipm	logDice	Частота, ipm Frequency, ipm	MS	Частота, ipm Frequency, ipm
9.	浙江*	61,1	训练	1540,2	什么	359,4
10.	伙伴	61,1	潜艇	391,3	潜艇	391,3
11.	驱逐	124,9	必须	181,8	任务	958,8
12.	驾驶	177,6	指挥	593,9	问题	491,2
13.	巡逻	133,2	研究	355,2	技术	617,4
14.	侦察	120,7	厚鑫*	90,2	指挥	593,9
15.	包括	177,6	辽宁*	70,8	环境	269,2
16.	介绍	341,3	沃洛*	79,1	包括	177,6
17.	轰炸	88,8	组织	677,1	培养	166,5
18.	必须	181,8	驱逐	124,9	爷爷	156,8
19.	选择	137,4	技术	617,4	必须	181,8
20.	荣誉	117,9	威胁	87,4	解放	516,2

Меры logDice и MS при введении порога отсечения в 60 ipm показали себя лучше: в верхней двадцатке рейтинга первая мера выделяет только три ИС, вторая – лишь одно; 14 лексических единиц из 20 совпадают, причем ряд из них достаточно явно отражает тематику военных новостей: 驾驶 *jiàshǐ* ‘водить, пилотировать’, 记者 *jìzhě* ‘корреспондент’, 退役 *tuìyì* ‘уволиться со службы’, 世界 *shìjiè* ‘мир, свет’, 训练 *xùnliàn* ‘тренировка, подготовка’, 潜艇 *qiántǐng* ‘подводная лодка’, 指挥 *zhǐhuī* ‘командовать’, 技术 *jìshù* ‘техника’ и т. д.

Тем не менее, суть этих мер осталась прежней: они нацелены на поиск элементов с ограниченной сочетаемостью. Это свойство может никак не коррелировать с устойчивостью терминологических и им подобных выражений профессионального дискурса, поэтому рассматриваемые здесь статистические меры не очень хорошо подходят, например, для целей отбора наиболее употребительной лексики.

К достоинствам этих статистических мер относится (в комбинации с высоким порогом отсечения по частоте) то, что они довольно хорошо справляются с «нелексическими» биграммами: синтаксическими конструкциями, в основном структурно незавершенными, и различными синтагматическими фрагментами более крупных единиц и конструкций, которые в простом частотном списке (при ipm ≥ 60) составляют суммарно 28,8 % [Коршунов, 2020, с. 22]. Автор одной из мер, MS (minimum sensitivity), отмечал эту способность отфильтровывать незначимые слова как качество, важное для решения многих практических задач языковой обработки [13] [Pedersen, 1998].

Кроме того, к достоинствам меры logDice (и, соответственно, близкой к ней MS) можно отнести независимость от размера корпуса, что позволяет использовать их для больших корпусов и сравнений между разными корпусами [14].

[13] В оригинале: “The tendency of minimum sensitivity to filter out bigrams containing non-content words is an important quality in many practical language processing applications” [Pedersen, 1998].

[14] В оригинале: “logDice is not affected by the size of the corpus and, therefore, can be used to compare scores between different corpora. logDice is the preferred statistic measure for large corpora” (https://www.sketchengine.eu/my_keywords/logdice).

3.2. Частота, меры t-score и log-likelihood: приоритет частотности

Результаты меры t-score, как показано в табл. 2, в наибольшей степени совпадают с частотным списком (91 из 100) и результатами меры log-likelihood (82 из 100), которые между собой также ожидаемо имеют большое число общих результатов (77 из 100).

Анализ первых ста биграмм в трех вариантах рейтинга – по частоте, t-score и log-likelihood – показывает, что, во-первых, применение статистических мер улучшает качество результатов по сравнению с анализом по частоте, во-вторых, из них именно log-likelihood лучше всего выделяет знаменательные (содержательные) лексические единицы.

Так, в первой сотне результатов t-score отсутствуют такие нелексические биграммы из частотного списка, как 的一 *de yī*, 队的 *duì de*, 了一 *le yī*, 国军 *guó jūn*, 的战 *de zhàn*, 是一 *shì yī*, 上的 *shàng de*, 的军 *de jūn*. Большинство из них является сочетанием грамматического элемента с фрагментом соседнего слова, 国军 *guó jūn* представляет собой межсловную биграмму от словосочетаний 中国军人 *zhōngguó jūnrén* ‘китайские военнослужащие’ или 中国军队 *zhōngguó jūnduì* ‘китайские войска’ и т. п. В первой сотне результатов меры log-likelihood по сравнению с частотным списком помимо уже названных нелексических биграмм отсутствуют также синтаксическая конструкция 这是 *zhè shì* ‘это [есть]’, указательное местоимение 这个 *zhège* ‘этот’, глагольная форма 成为 *chéngwéi* ‘стать’, внутрисловная биграмма 防部 *fángbù* от 国防部 *guófángbù* ‘министерство обороны’.

При этом единственная нелексическая биграмма в первой сотне log-likelihood находится только на 92-м месте (*放军 fàngjūn*, часть номинации 解放军 *jiěfàngjūn* ‘[народно-]освободительная армия’), в частотном списке занимающая 61-е место, а в рейтинге t-score – 60-е. Лексические единицы «периферийных» частей речи с особой семантикой, такие как местоимения и числительные [15], также в основном занимают в результатах log-likelihood более низкие позиции (далее приводятся места в рейтингах по частоте, t-score и log-likelihood соответственно):

他们 *tāmen* ‘они’ – 9, 7, 9;

我们 *wǒmen* ‘мы’ – 10, 8, 10;

什么 *shénme* ‘что, какой’ – 90, 80, 46;

一个 *yīge* ‘один’ – 8, 10, 25;

第一 *dìyī* ‘первый’ – 29, 30, 49;

一次 *yīcì* ‘один [раз]’ – 42, 50, 84.

Качественно – с точки зрения лексического значения биграмм и его соответствия тематике коллекции текстов, послужившей языковым материалом, – три рассматриваемые меры можно считать вполне эффективными. При этом по количеству знаменательных лексических единиц лучшие результаты демонстрирует log-likelihood, за ней идет t-score. Наши данные подтверждают вывод группы китайских и британских исследователей о том, что log-likelihood является эффективной мерой для корпусов среднего размера [Piao et al., 2006, p. 19].

3.3. Меры MI^3 и MI.log-f: компромисс функциональных приоритетов

Как упоминалось выше, мера MI^3 оказалась наиболее универсальной с точки зрения совпадений с другими мерами (см. табл. 2), а мера MI.log-f имеет больше всего общих результатов с MI^3 (75 из 100). Можно предположить, что эти меры соединяют два подхода, являя собой результат компромисса между ориентацией на выявление уникальных (единичных)

[15] Ср.: «Среди периферийных частей речи действительно есть классы, выделяемые по значению. Местоимения и числительные для “экзотических” языков часто могут быть выделены только на основе семантики» [Алпатов, 2016, с. 20]. В нашем случае важно, что местоимения и числительные имеют настолько универсальные в масштабах языка значения, что по частоте употребления приближаются к служебным словам.

коллокатов и функциональным приоритетом определения частоты совместной встречаемости.

Такую комбинацию характеристик – частотные биграммы с ограниченной сочетаемостью элементов – мы уже получали искусственно, когда к результатам первых трех мер (MI, MS и logDice) применяли порог отсечения по частоте (см. п. 3.1). И действительно, в первой двадцатке рейтинга MI.log-f (см. табл. 1) имеется 14 совпадений с двадцатью лучшими результатами меры MS с порогом отсечения 60 ipm и 15 совпадений – с logDice (см. табл. 3). Из этих сравнений можно сделать вывод, что мера MI.log-f представляет собой разновидность мер MI, MS и logDice, предназначенных для поиска ограниченной сочетаемости, но с уже учтенным порогом отсечения по частоте [16].

Что касается меры MI^3 (см. табл. 1), то она имеет ровно половину совпадений (10 из 20) с лучшими частотными результатами мер MS и logDice ($ipm \geq 60$, см. табл. 3), а также 12 совпадений из 20 с log-likelihood (ср. табл. 1). Такие показатели подтверждают ее функциональную универсальность, способность усреднять противоположные тенденции [17]. Она хорошо справляется с нелексическими биграммами (в первой сотне рейтинга нет ни одной), с фрагментами ИС (всего два в первой сотне – на 88-м и 100-м местах), не завышает ранг местоимений и числительных (什么 *shénme* ‘что, какой’ – 12-е место; 我们 *wǒmen* ‘мы’ – 19-е; 他们 *tāmen* ‘они’ – 26-е; 一个 *yīge* ‘один’ – 82-е). Наши результаты подтверждают вывод о том, что мера MI^3 преодолела недостатки своего прототипа (MI) и стала одним из самых простых и эффективных статистических средств в извлечении коллокаций двухкомпонентного состава (биграмм) [全昌勤等 / Цюань Чанцинь и др., 2005, с. 56].

Заключение

Таким образом, сопоставление семи статистических мер выделения коллокаций применительно к иероглифическим биграммам на китайском языке и сравнение их с частотой совместной встречаемости биграмм позволяет сделать ряд выводов.

Во-первых, статистические меры имеют то преимущество перед простым анализом по частоте, что они понижают ранг биграмм, не совпадающих с лексическими единицами, т. е. дают более качественный результат как минимум в пределах первых десятков (сотни) мест своего рейтинга.

Во-вторых, функционал имеющихся статистических мер ориентирован на выявление разных характеристик биграмм, а их адекватное применение диктуется конкретной исследовательской задачей. Так, мера MI эффективнее других находит в коллекции текстов самые редкие сочетания; в частности, она показала себя наилучшим инструментом для выделения китайских одноморфемных двусложных слов «ляньмяньцзы». Близкие результаты демонстрируют меры MS и logDice, которые чуть лучше других справляются с ИС (т. е. дают им более низкий ранг), особенно при введении адекватного порога отсечения по частоте. Еще более хорошие результаты дает мера MI.log-f: она учитывает частотность и не требует искусственных порогов.

Все перечисленные меры ориентированы на выделение биграмм с ограниченной сочетаемостью компонентов, что является одним из важных признаков фразеологически связанных сочетаний. Соответственно если задача исследователя подразумевает поиск статистически подтвержденной идиоматической связи между слогоморфемами, то на основе наших результатов ему можно рекомендовать использование меры MI.log-f.

[16] Косвенно это подтверждается выводом В. П. Захарова о том, что мера MI.log-f является лидером по точности выделения высокоранговых коллокаций в корпусе как с примененным порогом отсечения по частоте, так и без него, а также рекомендацией в общих случаях использовать для извлечения коллокаций меры MI.log-f, logDice и MS [Zakharov, 2017a; 2017b].

[17] Отдельный вопрос – какой научный или практический интерес может представлять такое усреднение.

Если же исследовательская задача заключается в отборе наиболее частотной лексики, характерной для данной коллекции текстов, то хорошей альтернативой простому анализу по частоте является мера t-score, а лучшей – log-likelihood. Последняя особенно хорошо отделяет знаменательные лексические единицы от синтагматического «мусора» (окказиональных, хоть и частых, сочетаний фрагментов слов, грамматических элементов и синтаксических конструкций), а также не позволяет завышать ранг числительных и местоимений.

Наиболее универсальной статистической мерой в нашем исследовании оказалась MI^3, которая не искажает результаты ни по частоте совместной встречаемости биграмм, ни по выявлению ограниченности сочетаемости их компонентов. Она обладает способностью к сбалансированному учету этих противоположных параметров. Позволим себе высказать предположение, что данная мера могла бы с успехом использоваться для сравнения различных корпусов или коллекций текстов.

Таким образом, целесообразность использования статистических мер для выделения коллокаций иероглифических биграмм на китайском языке можно считать доказанной. Однако их эффективность зависит от учета исследователем их функциональной адекватности специфике стоящих перед исследователем задач. Вопрос о возможности выделять в китайском тексте этим же комплексом средств более длинные лексические единицы – предмет отдельного исследования.

Список литературы

Алпатов В. М. Части речи и семантика // Язык, сознание, коммуникация: Сб. ст. / Отв. ред. В. В. Красных, А. И. Изотов. М.: МАКС Пресс, 2016. Вып. 53. С. 11–26.

Влавацкая М. В. Типология коллокаций в комбинаторной лингвистике // Мир науки, культуры, образования. 2019. № 4 (77). С. 439–442.

Власова Е. А., Карпова Е. Л., Ольшевская М. Ю. Лексический минимум по языку специальности: сколько слов достаточно? Разработка принципов минимизации // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2019. Т. 17, № 4. С. 63–77. DOI 10.25205/ 1818-7935-2019-17-4-63-77

Гроховский П. Л., Добров А. В., Доброва А. Е., Захаров В. П., Сомс Н. Л. Компьютерный морфосинтаксический анализ несегментированного текста (на материале корпуса тибетских грамматических сочинений) // Структурная и прикладная лингвистика: Межвуз. сб. / Отв. ред. И. С. Николаев. СПб.: Изд-во СПбГУ, 2019. Вып. 12: К 60-летию отделения прикладной, компьютерной и математической лингвистики СПбГУ. С. 69–80.

Грудева Е. В., Тиханович А. Н. Лексическая функция MAGN в современном русском языке: корпусное и экспериментальное изучение: Моногр. Новосибирск: Изд-во СибАК, 2014. 264 с.

Захаров В. П., Хохлова М. В. Анализ эффективности статистических методов выявления коллокаций в текстах на русском языке // Компьютерная лингвистика и интеллектуальные технологии. 2010. № 9 (16). С. 137–143.

Иорданская Л. Н., Мельчук И. А. Смысл и сочетаемость в словаре. М.: Языки славянских культур, 2007. 673 с.

Касевич В. Б. Субморфы, слогоморфемы и слогоморфемные языки // Касевич В. Б. Труды по языкознанию: В 2 т. / Под ред. Ю. А. Клейнера. СПб.: Филол. фак. СПбГУ, 2011а. Т. 2. С. 389–394.

Касевич В. Б. О стратегиях сегментации текста (на материале китайского, японского и русского языков) // Касевич В. Б. Труды по языкознанию: В 2 т. / Под ред. Ю. А. Клейнера. СПб.: Филол. фак. СПбГУ, 2011б. С. 615–622.

Коршунов Д. С. Частота совместной встречаемости иероглифов как показатель лексичности (при отборе лексики китайского военного дискурса) // Филологические науки в МГИМО. 2020. Т. 6, № 4 (24). С. 14–24. DOI 10.24833/2410-2423-2020-4-24-14-24

Хаматова А. А. Словообразование современного китайского языка. М.: Муравей, 2003. 224 с.

Хохлова М. В. Особенности статистических мер при выделении биграмм // Тр. Международной конференции «Корпусная лингвистика – 2017». СПб.: Изд-во СПбГУ, 2017. С. 349–354.

Ягунова Е. В., Пивоварова Л. М. Природа коллокаций в русском языке. Опыт автоматического извлечения и классификации на материале новостных текстов // Сб. НТИ. Сер. 2. 2010. № 6. С. 30–40.

Chen, X. C., Shi, Z., Qiu, X. P., Huang, X. J. Adversarial multi-criteria learning for Chinese word segmentation. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017, vol. 1, pp. 1193–1203.

Church, K., Hanks, P. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 1990, no. 16 (1), pp. 22–29.

Da, Jun. Chinese text computing. 2004. (на кит., англ. яз.) URL: http://lingua.mtsu.edu/chinese-computing (дата обращения 23.03.2020).

Lan Huang, Juan Zhou, Jing Xue, Yongxing Li, Youfu Du. DACE: Extracting and Exploring Large Scale Chinese Web Collocations with Distributed Computing. *American Journal of Information Systems*, 2017, vol. 5, no. 1, pp. 27–32. DOI 10.12691/ajis-5-1-4

Li Jingyang, Sun Maosong, Zhang Xian. A Comparison and Semi-Quantitative Analysis of Words and Character-Bigrams as Features in Chinese Text Categorization. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. Sydney, 2006, pp. 545–552.

Li Shouji, Guo Shulun. Collocation Analysis Tools for Chinese Collocation Studies. *Journal of Technology and Chinese Language Teaching*, 2016, no. 7 (1), pp. 56–77.

Meng, Y., Li, X., Sun, X., Han, Q., Yuan, A., Li, J. Is Word Segmentation Necessary for Deep Learning of Chinese Representations? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 3242–3252.

Pedersen, T. Dependent Bigram Identification. *Proceedings of American Association of Artificial Intelligence*, 1998, pp. 193. URL: https://www.aaai.org/Papers/AAAI/1998/AAAI98-193.pdf

Piao, S., Sun Guangfan, Rayson, P., Yuan Qi. Automatic Extraction of Chinese Multiword Expressions with a Statistical Tool. In: Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics Workshop on Multiword Expressions in a Multilingual Context. Trento, Italy, 2006, pp. 17–24.

Sproat, R., Shih, C. A statistical method for finding word boundaries in Chinese text. *Computer Processing of Chinese and Oriental Languages*, 1990, vol. 4, no. 4, pp. 336–351.

Sun, M. S., Shen, D. Y., Benjamin, K. T. Chinese Word Segmentation without Using Lexicon and Hand-crafted Training Data. *Meeting of the Association for Computational Linguistics and International Conference on Computational Linguistics Association for Computational Linguistics*, 1998, no. 48 (2), pp. 1265–1271.

Zakharov, V. Automatic Collocation Extraction: Association Measures Evaluation and Integration. In: Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2017). Moscow, RSUH, 2017a, vol. 1, iss. 16 (23), pp. 396–407.

Zakharov, V. Comparative Evaluation and Integration of Collocation Extraction Metrics. In: Ekstein K., Matousek V. (eds.). Lecture Notes in Computer Science, vol. 10415 (Text, Speech, and Dialogue – 20th International Conference, TSD 2017, Prague, Czech Republic, August 27–31, 2017, Proceedings). Springer International Publ. AG, 2017b, pp. 255–262.

王素格, 杨军玲, 张武 (Ван Сугэ, Ян Цзюньлин, Чжан У). 自动获取汉语词语搭配 (Автоматическое извлечение коллокаций на китайском языке) // 中文信息学报, 2006. 第20卷. 第6期. 31–37页. (на кит. яз.)

邓耀臣 (Дэн Яочэнь). 词语搭配研究中的统计方法 (Статистические методы исследования коллокаций) // 大连海事大学学报（社会科学版）, 2003. 第2卷. 第4期. 74–77页. (на кит. яз.)

孙茂松, 黄昌宁, 邹嘉彦, 陆方, 沈达阳 (Сунь Маосун, Хуан Чаннин, Цзоу Цзяянь, Лу Фан, Шэнь Даян) 利用汉字二元语法关系解决汉语自动分词中的交集型歧义 (Снятие неоднозначности при автоматической сегментации китайского текста с помощью иероглифических биграмм) // 计算机研究与发展, 1997. 第34卷. 第5期. 332–339页. (на кит. яз.)

全昌勤, 刘辉, 何婷婷 (Цюань Чанцинь, Лю Хуэй, Хэ Тинтин). 基于统计模型的词语搭配自动获取方法的分析与比较 (Анализ и сопоставление методов автоматического извлечения коллокаций на основе статистических моделей) // 计算机应用研究, 2005. 第22卷. 第9期. 55–57页. (на кит. яз.)

Список источников

中国新闻网 – 中新网 [сайт китайской службы новостей «Чжунсинь»]. 军事新闻滚动新闻 [список военных новостей]. (на кит. яз.) URL: https://www.chinanews.com/mil/news (дата обращения 29.01.2019 – 28.02.2019).

References

Alpatov, V. M. Parts of Speech and Semantics. In: Krasnykh V. V., Izotov A. I. (eds.). Language, Consciousness, Communication: Collection of articles. Moscow, MAKS Press, 2016, vol. 53, pp. 11–26. (in Russ.)

Chen, X. C., Shi, Z., Qiu, X. P., Huang, X. J. Adversarial multi-criteria learning for Chinese word segmentation. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017, vol. 1, pp. 1193–1203.

Church, K., Hanks, P. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 1990, no. 16 (1), pp. 22–29.

Da, Jun. Chinese text computing. 2004. (in Chin., Engl.) URL: http://lingua.mtsu.edu/chinese-computing (accessed: 23.03.2020).

Grokhovskiy, P. L., Dobrov, A. V., Dobrova, A. E., Zakharov, V. P., Soms, N. L. Computer Morphosyntactic Analysis of the Non Segmented Text (Based on the Material of the Corpus of Tibetan Grammar Treatises). In: Nikolayev I. S. (ed.). Structural and Applied Linguistics: Interuniversity Collection. St. Petersburg, St. Petersburg State Uni. Press, 2019, vol. 12, pp. 69–80. (in Russ.)

Grudeva, E. V., Tikhanovich, A. N. Lexical function of MAGN in modern Russian: corpus and experimental study: Monograph. Novosibirsk, SibAK Publ., 2014, 264 p. (in Russ.)

Iagunova, E. V., Pivovarova, L. M. Nature of collocations in the Russian language. Experience of automatic extraction and classification on the material of news texts. *Sb. NTI. Series 2*, 2010, no. 6, pp. 30–40. (in Russ.)

Iordanskaya, L. N., Melchuk, I. A. Meaning and compatibility in the dictionary. Moscow, Languages of Slavic Cultures Publ., 2007, 673 p. (in Russ.)

Kasevich, V. B. On the strategies of text segmentation (based on the material of Chinese, Japanese and Russian languages). In: Kasevich, V. B. Works on Linguistics: In 2 vols. Ed. by Yu. A. Kleyner. St. Petersburg, Faculty of Philology, St. Petersburg State Uni. Press, 2011, vol. 2, pp. 615–622. (in Russ.)

Kasevich, V. B. Submorphs, syllomorphisms and syllable languages. In: Kasevich, V. B. Works on Linguistics: In 2 vols. Ed. by Yu. A. Kleyner. St. Petersburg, Faculty of Philology, St. Petersburg State Uni. Press, 2011, vol. 2, pp. 389–394. (in Russ.)

Khamatova, A. A. Word formation of the modern Chinese language. Moscow, Muravey Publ., 2003, 224 p. (in Russ.)

Khokhlova, M. V. Distinctive features of association measures for bigram extraction. In: Proceedings of the International Conference "Corpus Linguistics – 2017". St. Petersburg, St. Petersburg State Uni. Press, 2017, pp. 349–354. (in Russ.)

Korshunov, D. S. Frequency of Co-Occurrence of Chinese Characters as an Indicator of Lexicality (When Selecting the Vocabulary of Chinese Military Discourse). *Philological Sciences at MGIMO*, 2020, vol. 6, no 4 (24), pp. 14–24. (in Russ.) DOI 10.24833/2410-2423-2020-4-24-14-24

Lan Huang, Juan Zhou, Jing Xue, Yongxing Li, Youfu Du. DACE: Extracting and Exploring Large Scale Chinese Web Collocations with Distributed Computing. *American Journal of Information Systems*, 2017, vol. 5, no. 1, pp. 27–32. DOI 10.12691/ajis-5-1-4

Li Jingyang, Sun Maosong, Zhang Xian. A Comparison and Semi-Quantitative Analysis of Words and Character-Bigrams as Features in Chinese Text Categorization. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. Sydney, 2006, pp. 545–552.

Li Shouji, Guo Shulun. Collocation Analysis Tools for Chinese Collocation Studies. *Journal of Technology and Chinese Language Teaching*, 2016, no. 7 (1), pp. 56–77.

Meng, Y., Li, X., Sun, X., Han, Q., Yuan, A., Li, J. Is Word Segmentation Necessary for Deep Learning of Chinese Representations? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 3242–3252.

Pedersen, T. Dependent Bigram Identification. *Proceedings of American Association of Artificial Intelligence*, 1998, pp. 193. URL: https://www.aaai.org/Papers/AAAI/1998/AAAI98-193.pdf

Piao, S., Sun Guangfan, Rayson, P., Yuan Qi. Automatic Extraction of Chinese Multiword Expressions with a Statistical Tool. In: Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics Workshop on Multiword Expressions in a Multilingual Context. Trento, Italy, 2006, pp. 17–24.

Sproat, R., Shih, C. A statistical method for finding word boundaries in Chinese text. *Computer Processing of Chinese and Oriental Languages*, 1990, vol. 4, no. 4, pp. 336–351.

Sun, M. S., Shen, D. Y., Benjamin, K. T. Chinese Word Segmentation without Using Lexicon and Hand-crafted Training Data. *Meeting of the Association for Computational Linguistics and International Conference on Computational Linguistics Association for Computational Linguistics*, 1998, no. 48 (2), pp. 1265–1271.

Vlasova, E. A., Karpova, E. L., Olshevskaya, M. Yu. Vocabulary: How Many Words Are Enough? Principles of Minimizing Learners' Vocabulary. *Vestnik NSU. Series: Linguistics and Intercultural Communication*, 2019, vol. 17, no. 4, pp. 63–77. (in Russ.) DOI 10.25205/1818-7935-2019-17-4-63-77

Vlavatskaya, M. V. Typology of Collocations in Combinatorial Linguistics. *The world of science, culture and education*, 2019, no. 4 (77), pp. 439–442. (in Russ.)

Zakharov, V. Automatic Collocation Extraction: Association Measures Evaluation and Integration. In: Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2017). Moscow, RSUH, 2017a, vol. 1, iss. 16 (23), pp. 396–407.

Zakharov, V. Comparative Evaluation and Integration of Collocation Extraction Metrics. In: Ekstein K., Matousek V. (eds.). Lecture Notes in Computer Science, vol. 10415 (Text, Speech, and Dialogue – 20th International Conference, TSD 2017, Prague, Czech Republic, August 27–31, 2017, Proceedings). Springer International Publ. AG, 2017b, pp. 255–262.

Zakharov, V. P., Khokhlova, M. V. Study of effectiveness of statistical measures for collocation extraction on Russian texts. *Computational Linguistics and Intelligent Technologies*, 2010, vol. 9 (16), pp. 137–143. (in Russ.)

王素格,杨军玲,张武 (**Wang Suge, Yang Junling, Zhang Wu**). 自动获取汉语词语搭配 (Automatic Collocation Extraction in Chinese) // 中文信息学报, 2006. 第20卷. 第6期. 31–37页. (in Chin.)

邓耀臣 (**Deng Yaochen**). 词语搭配研究中的统计方法 (Collocation statistical research methods) // 大连海事大学学报（社会科学版）, 2003. 第2卷. 第4期. 74–77页. (in Chin.)

孙茂松, 黄昌宁, 邹嘉彦, 陆方, 沈达阳 (**Sun Maosong, Huang Changning, Benjamin K. Tsou, Lu Fang, Shen Dayang**) 利用汉字二元语法关系解决汉语自动分词中的交集型歧义 (Using character bigram for ambiguity resolution in Chinese word segmentation) // 计算机研究与发展, 1997. 第34卷. 第5期. 332–339页. (in Chin.)

全昌勤, 刘辉, 何婷婷 (**Quan Changqin, Liu Hui, He Tingting**). 基于统计模型的词语搭配自动获取方法的分析与比较 (Analysis and comparison of automatic collocation extraction methods based on statistical models) // 计算机应用研究, 2005. 第22卷. 第9期. 55–57页. (in Chin.)

List of Sources

中国新闻网 – 中新网 ["Zhongxin" news service site]. 军事新闻滚动新闻 [military news scroll list]. URL: https://www.chinanews.com/mil/news (accessed: 29.01.2019–28.02.2019) (in Chin.)

Информация об авторе

Дмитрий Сергеевич Коршунов, кандидат филологических наук
SPIN 7282-7336

Information about the Author

Dmitry S. Korshunov, Candidate of Sciences (Philology)
SPIN 7282-7336

Статья поступила в редакцию 14.03.2022;
одобрена после рецензирования 10.04.2022; *принята к публикации 23.04.2022*
The article was submitted 14.03.2022;
approved after reviewing 10.04.2022; *accepted for publication 23.04.2022*