Вестник Костромского государственного университета. 2022. Т. 28, № 3. С. 180–185. ISSN 1998-0817
Vestnik of Kostroma State University, 2022, vol. 28, № 3, pp. 180–185. ISSN 1998-0817
Научная статья
УДК 81’1
EDN RSMVTO
https://doi.org/10.34216/1998-0817-2022-28-3-180-185

КЛЮЧЕВЫЕ СЛОВА В АСПЕКТЕ ЧАСТОТНОСТИ И ТЕМАТИЧЕСКОЙ РЕЛЕВАНТНОСТИ

Галкина Наталия Павловна, кандидат филологических наук, Военная академия радиационной, химической и биологической защиты имени Маршала Советского Союза С.К. Тимошенко, Кострома, Россия, gnpav@mail.ru, https://orcid.org/0000-0001-7019-2413

Аннотация. Статья акцентирует внимание на значении ключевых слов, их статистических показателей для определения тематической доминанты при работе с большими массивами текстов. Описание основано на материалах типологического лингвистического исследования текстов военных песен 1939–1945 гг., англоязычных и русскоязычных. Отбор ключевых слов осуществлялся на основе семантического, лексико-синтаксического, морфологического анализа, а также с учётом частотности их употребления. Частота употребления слова не всегда может быть определяющим признаком для маркировки его в качестве ключевого. В рамках одного текста ключевыми являются слова, которые помогают открыть доступ к пониманию текста, разгадать его смысл, запомнить содержание. При объединении большого количества текстов по авторству, хронологии, тематической, стилевой или иной отнесенности частотность ключевых слов имеет значение и может послужить определяющим фактором, критерием классификации. В данной работе показано, что результаты тематического распределения текстов на основе семантического анализа их содержания соответствуют результатам статистического анализа ключевых слов и подтверждаются количественными данными частотности, полученными машинным способом. Результаты релевантны как для русскоязычных, так и для англоязычных материалов.

Ключевые слова: частотность, лексика, семантика, языковая картина, тематика, корпус, массивы текстов

Для цитирования: Галкина Н.П. Ключевые слова в аспекте частотности и тематической релевантности // Вестник Костромского государственного университета. 2022. Т. 28, № 3. С. 180–185. https://doi.org/10.34216/1998-0817-2022-28-3-180-185

Research Article

KEYWORDS IN TERMS OF FREQUENCY AND THEMATIC RELEVANCE

Natalia P. Galkina, PhD in Philology, Nuclear, Biological, Chemical Defence Military Academy named after Marshal of the Soviet Union S.K. Timoshenko, Kostroma, Russia, gnpav@mail.ru, https://orcid.org/0000-0001-7019-2413

Abstract. The article focuses on the role of keywords, their statistical data for determining the thematic dominance when working with large arrays of texts. The description is based on the materials of a typological linguistic study of the texts of military songs, the period of 1939-1945, in English and Russian. The selection of keywords was carried out on the basis of semantic, lexical-syntactic, morphological analysis, taking into account the frequency of their use. The frequency of using a word may not always be a defining feature for marking it as a keyword. Within the framework of one text, the keywords may be words that help understand the sense, unravel its deep meaning, remember the content. When combining a large number of texts, by authorship, chronology, thematic, stylistic or other relatedness, the frequency of keywords matters and can serve as a determining factor, a classification criterion. This paper shows that the results of the thematic distribution of texts based on the semantic analysis of their content correspond to the results of statistical analysis of the keywords and are confirmed by machine quantitative indicators of their frequency. The results are relevant for both Russian-language and English-language materials.

Keywords: frequency, vocabulary, semantics, language pattern, subject matter, corpus, arrays of texts.

For citation: Galkina N.P. Keywords in terms of frequency and thematic relevance. Vestnik of Kostroma State University, 2022, vol. 28, № 3, pp. 180–185 (In Russ.). https://doi.org/10.34216/1998- 0817-2022-28-3-180-185

Данная статья основана на материалах исследования, посвященного типологическому лингвистическому анализу содержания текстов популярных музыкальных произведений 1939–1945 гг. Результатом является построение, сопоставление языковой картины восприятия событий Второй мировой / Великой Отечественной войны её непосредственными участниками и свидетелями. В качестве материала использовались тексты англоязычных (англо-американских) и русскоязычных (советских) песен военных лет, отбор которых производился из печатных песенных сборников, а также интернет-ресурсов, содержащих записи и тексты песен указанного периода. Корпус исследования составил 115 англоязычных и 230 русскоязычных песен периода Второй мировой войны. Исследование имеет многоаспектный характер, и его результаты могут рассматриваться в различных сферах: общеобразовательной, исторической, социальной, филологической. Обращает на себя внимание лингвистический аспект исследования, а именно – роль ключевых слов и статистических показателей в определении тематической доминанты большого массива текстов. Эта сторона исследования и является предметом рассмотрения в данной работе.

Язык – зеркало культуры, в нём отражается не только реальный мир человека, не только реальные условия его жизни, но и общественное сознание народа, его менталитет, национальный образ жизни, традиции, обычаи, мораль, система ценностей, видение мира [Тер-Минасова: 14]. Способы и формы отражения человеком окружающего мира в языке обусловлены спецификой социокультурных, природных особенностей жизни данного речевого коллектива. В лингвистике изучается связь картины мира и языка, устанавливается их взаимоотношение. Показательными для языковой картины мира являются концептуальные, ключевые слова и их явные и «фоновые» смыслы. «Ключевые слова, обладая способностью кодировать исходную информацию, передавать её в обобщенной форме, способны послужить опорой для понимания и воспроизведения смысла» [Методы и приёмы: 4–5]. Ключевым словам свойственны следующие характеристики: они являются наиболее употребительными (частотными), представлены значимой лексикой, достаточно обобщены по своей семантике, стилистически нейтральны, не оценочны, связаны друг с другом сетью семантических связей, пересечения значений [Ванюшкин, Гращенко: 86]. Наиболее простым методом извлечения ключевых слов является статистический метод – ранжирование и отбор лексики по частоте употребления в тексте. Но признак частотности не всегда является превалирующим [Лазарева, Боломутова: 143–146]. Частота употребления слова не всегда может быть определяющим признаком, особенно в песенных, поэтических произведениях, которым свойственна образность, метафоричность, иносказательность и, соответственно, разнообразие, богатство лексики [Винарская: 27–28]. По выражению Н.Д. Арутюновой, «поэтические произведения являются естественной средой для метафоры» [Арутюнова: 169]. Извлечение ключевых слов требует умения абстрагирования, условного отделения «главного» от «неглавного». При этом осуществляется лексико-синтаксический анализ – разделение лексики текста на основные и служебные слова, удаление стоп-слов, не несущих смысловой нагрузки (артикли, предлоги, союзы, частицы, местоимения, междометия и т. д.), а также морфологический анализ, нацеленный на поиск разных форм выражения одного смысла (различные словоформы, однокоренные слова, синонимы).

Как отмечают исследователи, отбор ключевых слов часто выполняется интуитивно [Ванюшкин, Гращенко: 86]. Если речь идёт об одном тексте, частота употребления ключевых слов не столь важна: это слова, которые помогают открыть доступ к пониманию текста, разгадать его смысл, акцентировать, запомнить содержание. При объединении большого количества текстов (по авторству, хронологии, тематической, стилевой или другой отнесенности) частотность ключевых слов имеет значение. Количественные методы находят широкое применение для описания и классификации текстов, например при их тематическом распределении, анализе идиостиля авторов, установлении авторства анонимных текстов и т. д.

Выполняя исследование, мы исходили из того, что каждой группе (тематике) песен должен соответствовать определенный набор ключевых слов, поэтому на этапе тематической группировки производили отбор ключевых слов, характерных для песен данной группы (тематики). При отборе ключевых слов использовался гибридный метод – извлечение ключевых слов на основе семантического, лексико-синтаксического, морфологического анализа, с учётом частотности употребления слов.

На первоначальном этапе в результате анализа текстов русскоязычных (советских) песен Великой Отечественной войны было выделено 5 тематических направлений: 1) тема патриотизма и защиты Родины – 33 %; 2) прославление своего подразделения – 20 %; 3) поднятие боевого духа (включая шуточные песни) – 18 %; 4) тема любви и верности – 18 %; 5) тоска по родному дому, краю – 11 %. Данная тематика наглядно представлена в процентном соотношении на рисунке 1.

Анализ содержания англоязычных (англо-американских) песен [Галкина, Сорокина 2020] также позволяет определить 5 тематических направлений (не все из которых находят соответствие с указанной выше тематикой русскоязычных песен):

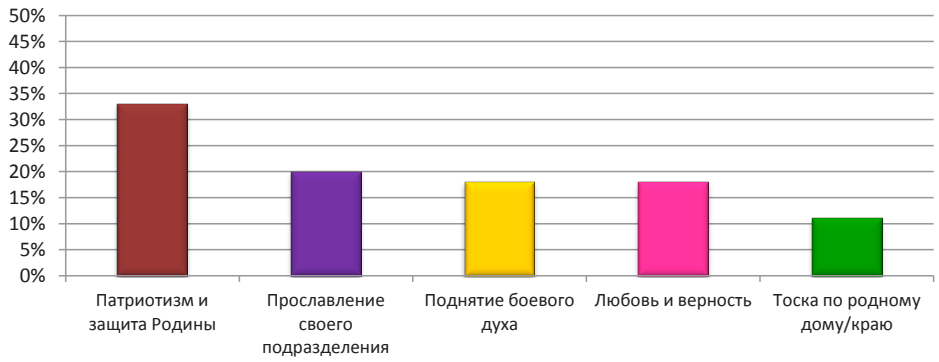

Рис. 1. Тематика русскоязычных песен 1939–1945 гг.

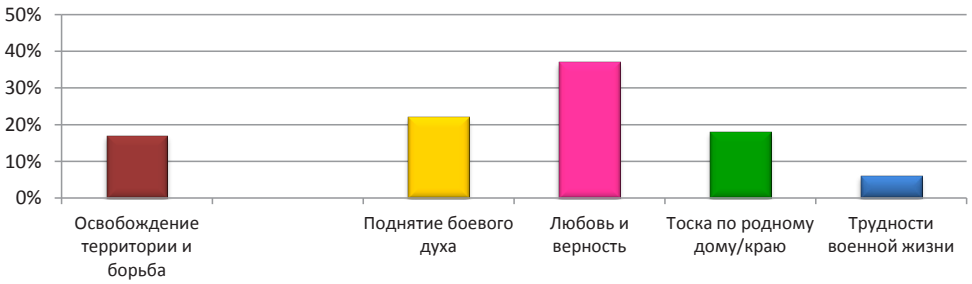

Рис. 2. Тематика русскоязычных песен 1939–1945 гг.

1) тема любви и верности – 37 %; 2) поднятие боевого духа (включая шуточные песни) – 22 %; 3) тоска по родному краю, дому, семье – 18 %; 4) освобождение территории и борьба – 17 %; 5) трудности и страдания военной жизни – 6 %. Тематика англоязычных песен 1939–1945 гг. представлена в процентном соотношении на рисунке 2.

Как видно, тематическое распределение и процентное соотношение обнаруживают различные результаты для массивов англо-американских и советских песен. Такое различие соответствует гипотезе исследования, целью которого являлось построение и сравнение языковой картины восприятия событий Второй мировой / Великой Отечественной войны различными её участниками/свидетелями. Тем не менее, чтобы сделать классификацию по темам и ключевым словам более единообразной, а также чтобы исключить/минимизировать субъективный, оценочный, эмоционально-психологический фактор, мы обратились к поисковым возможностям текстового процессора MS Word для количественного уточнения набора ключевых слов (их частотности) по каждой тематике. В результате выстроилась языковая (лексико-семантическая) картина содержания советских и англо-американских песен военного периода. Перейдем далее к её описанию и сопоставлению.

Анализ количественных данных соответствует тематическому распределению русскоязычных песен, выполненному на первом, субъективно-интуитивном этапе исследования. Количественные данные подсчета ключевых слов подтверждают, что тема патриотизма и защиты Родины, бесспорно, лидирует в корпусе русскоязычных военных песен. Лексика семантической группы *родина/отчизна* присутствует в значительном количестве во всех тематических категориях (с преобладанием в группах патриотического направления «Защита Родины» и «Прославление своего подразделения») и отчётливо лидирует в общей сводке по всем темам (359 употреблений). Боевой настрой на борьбу также пронизывает песни всех тематических направлений (с преобладанием в группах патриотической тематики), и соответствующая лексика выступает на втором месте по частоте употребления (315 употреблений в общей сводке по всем темам). Далее по частотности в общей сводке по всем темам следует лексика семантической группы *друг/товарищ* (252 употребления в целом). Данная лексика лидирует в тематической категории «Поднятие боевого духа» и занимает значительное место в других категориях («Прославление подразделения», «Любовь и верность»). Такой показатель отражает ценности дружбы и товарищества в смысловом содержании указанных категорий. Интересно, что упоминание *врага* в советских военных песнях по частотности сравнимо с упоминанием *друга/дружбы* (228 употреблений в сводке по всем темам). Это показывает степень ненависти и желания одолеть врага у советских людей периода Великой Отечест-

венной войны. В тематической категории «Любовь и верность», как и следовало ожидать, по частотности лидирует лексика с семантикой *любовь/друг/милый/ждать* (в целом 184 употребления в данной группе). В проекции на общую сводку данная лексика обнаруживает четвертый результат, что соответствует процентному соотношению песен данной группы в тематическом распределении (см. рис. 1). В пятой семантической группе «Тоска по родному краю/дому» преобладает лексика с семантикой *дом/край/город/сад/луг* (70 употреблений), и при сложении с близкой по семантике группой слов *родина/родной* в общей сводке частотности по всем темам получается пятый результат (103 употребления). Особо следует отметить тему *победы* в советских песнях военного периода. Мы не стали выделять их в отдельную группу, так как победный настрой пронизывает многие песни Великой Отечественной войны. Коннотативная лексика победы прослеживается в произведениях всех указанных категорий: *победа впереди / радость победы / солнце победы / поле победы / знамя побед / слава суворовских побед / ускорять победу / биться до победы / иль умереть, иль победить / вернемся с победой / приеду с победой / выпьем за победу / отпор / расплата / одолеть / сокрушить / разгромить / уничтожить / изгнать* и т. п. Частотный анализ слова *победа* (включая однокоренные) выдает результат 86, оно встречается во всех тематических группах нашей выборки, более всего – в первых двух категориях.

Таким образом, мы с удовлетворением отмечаем, что результаты семантического и статистического анализа содержания русскоязычных песен подтверждаются количественными данными частотности, полученными машинным способом.

В целях применения единого подхода к исследованию англоязычного и русскоязычного материала мы таким же образом проанализировали тематические группы англоязычных песен, вводя на их материале английские эквиваленты ключевых слов в поисковую систему MS Word. Результаты подсчета английских ключевых слов также соответствуют тематическому распределению англо-американских песен, выполненному на первоначальном этапе исследования (см. рис. 2). В корпусе англоязычных источников на первом месте по частотности – лексика семантической группы *любовь/верность/расставание* – в 308 употреблений в тематической группе «Любовь и верность» и 481 употребление в общем по всем темам. Второе место по частоте занимает лексика с семантикой *родной дом/край, помнить/покинуть/вернуться* – 62 употребления в тематической группе «Тоска по земле/дому» и 250 употреблений по всем темам. Тема «Поднятие боевого духа» разнообразна по содержанию и лексике, поэтому частотный анализ не дает заметного преобладания какой-либо лексики. Тем не менее объединение показателей, связанных с семантикой *дом* и *любовь,* выдает в данной тематической группе наибольший результат (в целом 52 употребления). На третье место по общему количеству выходит лексическая группа, охватывающая семантику *земля/свобода/ борьба/сражаться* – 145 употреблений по всем темам. Поиск слов, соответствующих родам войск, обнаружил примерно равную частоту употребления слов *sea, ship, sail* (37) и *sky, plane, air* (32). Это и понятно, так как боевые действия Великобритании и США во время Второй мировой войны осуществлялись в основном в воздухе и на море.

Здесь уместно привести высказывание американского математика Норберта Винера, одного из основоположников кибернетики и теории искусственного интеллекта: «Высшее назначение математики состоит в том, чтобы находить скрытый порядок в хаосе, который нас окружает» [Социологическая энциклопедия]. Порядок этот в данном случае заключается в том, что результаты тематического анализа текстов по содержанию (субъективного) совпадают (или близки, сопоставимы) со статистическими результатами «языка цифр» – машинного (объективного) подсчета ключевых слов и анализа их употребления по темам. Как видно из таблицы 1, совпадает не только распределение по темам, но и процентное соотношение! Отметим, что согласно выводу Б. Н. Головина, автора работы «Язык и статистика», для лингвистических исследований величина относительной ошибки допустима в 5–10 %, а в ряде случаев может быть и большей [Головин: 56]. В нашем случае погрешность остается в пределах 2 %. Результаты машинной обработки ключевых слов на англоязычном материале также подтверждают выводы, полученные на этапе интуитивного, субъективного анализа тематического распределения текстов с погрешностью менее 5 %.

Лексико-семантическая картина содержания военных песен отражает смысловые ценности народов в военный период. Частотность ключевых слов песен военных лет свидетельствует о том, что переживания советских людей были сосредоточены главным образом на боевых сражениях, защите Родины, самоотверженных совместных действиях, победе, уничтожении/изгнании врага, надёжном плече боевого товарища и через всё это – на возвращении к семье, любимой, в родной дом. Общая картина переживаний представителей Великобритании и США в период WWII несколько иная. Прежде всего, она складывается через тему расставания с любимой, близкими, домом, тоски и стремления к возвращению на родную землю. Тема сражения, освобождения, героизма, победы оказывается на втором плане. Об этом свидетельствует и столь незначительная частотность

Таблица 1

Частотность употребления ключевых слов по темам русскоязычных/советских песен Великой Отечественной войны

Ключевые слова	Тема					
	Защита Родины + Прославление подразделения 33 % + 20 % = 53 %	Поднятие боевого духа 18 %	Любовь и верность 18 %	Родной край/дом 11 %	По всем темам	
					кол-во	%
Родина Родной Родимый Отчизна Отчий	232	43	51	33	359	53 %
Бой Боевой Боец	223	44	37	11	315	
Смело, смелый Отважный	44	28				
Смерть Бессмертный	47 — 149	8 — 47				
Слава Славный	58	11				
Друг, дружба Товарищ	107	59	69	17	252	20 %
Милый Дорогой			29	18		
Любовь Любимая Любить			64 — 220 +	15	235	18 %
Ждать			22			
Помнить, память Прощай			36	11		
Дом, край, город, Сад, луг Река, дорога				70 — 81	114	9 %
Итого					1 275	100 %

частотность употребления слов *fight* – 22; *war* – 11; *enemy* – 10; *battle* – 4; *victory* – 4; *overtake/win* – 7.

Ключевые слова отражают содержание контента, описывают общую тему публикуемых материалов, например статьи, рубрики, сайта или блога. Работа поисковых систем связана с частотностью употребления ключевых слов. Данное исследование указывает на значение ключевых слов в процессе обработки больших массивов текстов. При объединении текстов в массивы, корпуса ключевые слова играют заметную роль, и частотность их может послужить определяющим фактором, маркером, критерием классификации.

Список литературы

Арутюнова Н.Д. Языковая метафора (синтаксис и лексика) // Лингвистика и поэтика. Москва: Наука, 1979. С. 147–173.

Ванюшкин А.С., Гращенко Л.А. Методы и алгоритмы извлечения ключевых слов // Новые информационные технологии в автоматизированных системах. 2016. № 19. С. 85–93.

Винарская Е.Н. Выразительные средства текста (на материале русской поэзии). Москва: Высшая школа, 1989. 136 с.

Галкина Н.П., Сорокина О.В. Произведения массовой культуры – реальные свидетели истории Второй мировой войны (на материале англоязычных песен) // Актуальные вопросы современного языкознания и тенденции преподавания иностранных языков в неязыковом вузе: теория и практика. Кострома: ВА РХБЗ, 2020. С. 93–100.

Головин Б.Н. Язык и статистика. Москва: Просвещение, 1970. 190 с.

Лазарева О.Ю., Боломутова М.С. Методы выделения ключевых слов в контексте электронных обучающих систем // Молодой учёный. 2016. № 26 (130). С. 143–146.

Методы и приёмы работы с ключевыми словами текста на уроках русского языка и литературы: сборник методических рекомендаций. Биробиджан: ОГАОУ ДПО «ИПКПР», 2017. 48 с.

Социологическая энциклопедия. URL: https://voluntary.ru/termin/viner-norbert.html (дата обращения: 10.03.2022).

Тер-Минасова С.Г. Язык и межкультурная коммуникация. Москва: Высшая школа, 2000. 259 с.

References

Arutyunova N.D. *Yazykovaya metafora (sintaksis i leksika)* [Language metaphor (syntax and vocabulary)]. *Lingvistika i poetika* [Linguistics and poetics]. Moscow, Nauka Publ., 1979, pp. 147–173. (In Russ.)

Vanyushkin A.S., Grashchenko L.A. *Metody i algoritmy izvlecheniya klyuchevyh slov* [Methods and algorithms for extracting keywords]. *Novye informacionnye tekhnologii v avtomatizirovannyh sistemah* [New information technologies in automated systems], 2016, No 19, pp. 85–93. (In Russ.)

Vinarskaya E.N. *Vyrazitel'nye sredstva teksta (na materiale russkoj poezii)* [Expressive means of the text (based on Russian poetry)]. Moscow, Vysshaya shkola Publ., 1989, 136 p. (In Russ.)

Galkina N.P., Sorokina O.V. *Proizvedeniya massovoj kul'tury – real'nye svideteli istorii Vtoroj mirovoj vojny (na materiale angloyazychnyh pesen)* [Works of mass culture - real witnesses of the history of the Second World War (on the material of English-language songs)]. *Aktual'nye voprosy sovremennogo yazykoznaniya i tendencii prepodavaniya inostrannyh yazykov v neyazykovom vuze: teoriya i praktika* [Topical issues of modern linguistics and trends in teaching foreign languages in a non-linguistic university: theory and practice]. Kostroma, VA RHBZ Publ., 2020, pp. 93–100. (In Russ.)

Golovin B.N. *Yazyk i statistika* [Language and statistics]. Moscow, Prosveshchenie Publ., 1970, 190 p. (In Russ.)

Lazareva O.YU., Bolomutova M.S. *Metody vydeleniya klyuchevyh slov v kontekste elektronnyh obuchayushchih sistem* [Methods for extracting keywords in the context of e-learning systems]. *Molodoj uchyonyj* [Young Scientist], 2016, No 26 (130), pp. 143–146. (In Russ.)

Metody i priyomy raboty s klyuchevymi slovami teksta na urokah russkogo yazyka i literatury: sbornik metodicheskih rekomendacij [Methods and techniques for working with text keywords in the lessons of the Russian language and literature: a collection of methodological recommendations]. Birobidzhan, OGAOU DPO «IPKPR» Publ., 2017, 48 p.

Sociologicheskaya enciklopediya [Sociological Encyclopedia]. URL: https://voluntary.ru/termin/viner-norbert.html (access date: 10.03.2022). (In Russ.)

Ter-Minasova S.G. *Yazyk i mezhkul'turnaya kommunikaciya* [Language and intercultural communication]. Moscow, Vysshaya shkola Publ., 2000, 259 p. (In Russ.)

Статья поступила в редакцию 11.06.2022; одобрена после рецензирования 31.08.2022; принята к публикации 05.09.2022.

The article was submitted 11.06.2022; approved after reviewing 31.08.2022; accepted for publication 05.09.2022.