Марина Анатольевна КОВЯЗИНА[1]

УДК 81'32

ИЗВЛЕЧЕНИЕ КЛЮЧЕВЫХ ТЕРМИНОВ НА БАЗЕ КОРПУСА ТЕКСТОВ О РАЗРАБОТКЕ НЕФТЯНЫХ И ГАЗОВЫХ МЕСТОРОЖДЕНИЙ

[1] кандидат филологических наук,
доцент кафедры английской филологии и перевода,
Институт филологии и журналистики,
Тюменский государственный университет
m.a.kovyazina@utmn.ru

Аннотация

Статья излагает результаты исследования, посвященного извлечению терминологии на базе текстового корпуса. Автор применяет программное приложение AntConc и корпусную поисковую систему Sketch Engine для формирования корпуса специальных текстов, рассматривающих основные этапы и методы разработки месторождений нефти и газа, и выявления терминологии, являющейся ключевой для данной предметной области. Основная терминология, описывающая область разработки месторождений нефти и газа, извлекается с использованием нескольких корпусных инструментов: построение частотных списков слов, вычисление относительной частоты (ipm) для единиц корпуса, выявление ключевых слов и терминов с применением статистической меры ключевого слова (keyness score), построение дистрибутивного тезауруса на основе меры ассоциации logDice. В результате анализа на базе корпуса выделены единицы, семантически близкие термину «разработка», а также отраслевые и общенаучные термины, ключевые для исследуемой предметной области.

Ключевые слова

Терминология, извлечение терминов, корпус, показатель ключевого слова.

DOI: 10.21684/2411-197X-2016-2-3-61-69

Цитирование: Ковязина М. А. Извлечение ключевых терминов на базе корпуса текстов о разработке нефтяных и газовых месторождений / М. А. Ковязина // Вестник Тюменского государственного университета. Гуманитарные исследования. Humanitates. 2016. Том 2. № 3. С. 61–69.
DOI: 10.21684/2411-197X-2016-2-3-61-69

Исследование отраслевых терминологий на протяжении долгого времени продолжает вызывать интерес как терминоведов, так и отраслевых специалистов и переводчиков. Однако в последние несколько лет одним из ведущих направлений терминологических исследований стало изучение терминологий с применением технологий корпусной лингвистики, автоматического выделения терминологических единиц и их сочетаний на базе текстовых корпусов. В частности, к подобным исследованиям относятся работы В. П. Захарова [4], В. П. Захарова и М. В. Хохловой [1; 2], G. Andersen [7], J. Kast-Aigner [9], A. Kilgarriff и др. [10].

Работы, посвященные анализу нефтегазовой терминологии и ее функционированию в текстах разных типов, представляют особую важность для лингвистов и переводчиков Тюмени и Тюменской области, так как нефтегазовая отрасль имеет первостепенное значение для Тюменского региона.

В данном исследовании выявление ключевой терминологии, описывающей предметную область разработки месторождений нефти и газа, проводится на базе корпуса специальных текстов. Под *языковым корпусом текстов* В. П. Захаров и С. Ю. Богданова понимают «большой, представленный в машиночитаемом виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач» [3: 7].

Сформированный нами корпус текстов, посвященных разработке нефтяных и газовых месторождений, включает тексты трех типов: научный, учебный и технический (см. Рис. 1). Доля учебного текста составляет почти половину всех текстов корпуса (49%) и включает учебные пособия о разработке нефтяных и газовых месторождений. Научный текст (41%) — научные статьи и авторефераты диссертаций, рассматривающие различные аспекты и методы разработки месторождений нефти и газа. Технический текст (10%) — текст инструкции о правилах безопасности при разведке и разработке нефтяных и газовых месторождений. Включенные тексты опубликованы с 2003 по 2014 гг. Общий объем корпуса составил 167 453 слов.

Для получения данных о частотности и сочетаемости единиц корпуса и извлечения ключевых слов и терминов исследуемой предметной области нами

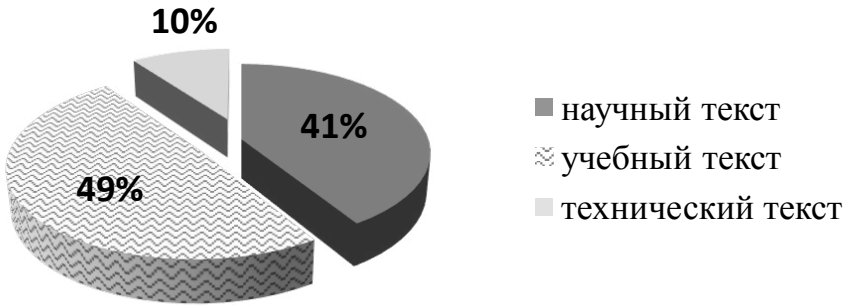

Рис. 1. Типы текстов, включенных в корпус

Fig. 1. The types of texts included into the corpus

были использованы инструменты анализа текстового корпуса — программное приложение AntConc (Version 3.4.3), а также система Sketch Engine. Программа AntConc позволяет составлять частотный список единиц корпуса, формировать конкорданс для искомого слова в формате KWIC (ключевое слово в контексте), выявлять ключевые слова корпуса, а также анализировать коллокаты и кластеры. Sketch Engine — это корпусная поисковая система (Corpus Query System), позволяющая пользователю просматривать сочетаемость слова, получать список семантически связанных слов при помощи функции «Тезаурус», а также сравнивать слова, наряду с другими функциями, типичными для поисковых систем данного типа. Данные о сочетаемости единицы интегрированы с функцией конкорданса [11: 4]. Проект системы Sketch Engine был запущен в 2004 г. компанией Lexical Computing Limited, основанной британским лексикографом, корпусным лингвистом А. Килгарифом. Основные функции данной системы разрабатывались А. Килгарифом в сотрудничестве с П. Рыхли, разработчиком корпус-менеджера Manatee, сотрудником Лаборатории обработки естественного языка (Natural Language Processing Laboratory) факультета информатики Университета им. Масарика в г. Брно, Чехия [15: 6]. Сетевая версия системы Sketch Engine (the.sketchengine.co.uk) включает большое количество готовых к использованию корпусов на нескольких десятках языков, а также инструменты для создания, установки и управления собственным корпусом [12: 17-18].

На первом этапе корпусного анализа текстов о разработке месторождений нефти и газа нами была проведена оценка частотности единиц корпуса с целью выделить термины, которые являются ключевыми для рассматриваемой области. Данная задача была выполнена благодаря применению нескольких корпусных инструментов.

С помощью приложения AntConc нами был получен список, включающий сотню наиболее частотных единиц корпуса. На данном этапе мы использовали в приложении список запрещенных слов (stop-words) с целью отсеять служебную лексику (предлоги, союзы и др.). Для оценки распространенности соответствующих единиц в корпусе рекомендуется опираться не на абсолютные частоты, а на относительные [5: 671]. Относительная частота, или «частота на миллион» (ipm), — это стандартное представление частоты токена или леммы, которое вычисляется относительно условного корпуса в миллион единиц независимо от объема реального корпуса [6: 86]. С помощью системы Sketch Engine была получена относительная частота лексем из первой сотни. Также мы применили функцию системы Sketch Engine *Keywords and Terms* для выявления ключевых слов и терминов на базе корпуса. Система сравнивает частоту единиц корпуса с их частотой в референтном корпусе, в частности, для русского языка используется корпус Russian Web 2011 (ruTenTen11) объемом более 14,5 млрд. слов, скомпилированный на базе веб-текстов [8]. В результате сравнения частот употребления система присваивает единицам, отличающимся неожиданно высокой частотностью, статистическую меру ключевого слова (keyness score), вычисляемую по следующей формуле:

$$\frac{f_{pm_{focus}}+n}{f_{pm_{ref}}+n},$$

где $f_{pm_{focus}}$ — относительная частота слова в изучаемом корпусе, $f_{pm_{ref}}$ — относительная частота слова в референтном корпусе, *n* — сглаживающий параметр (по умолчанию *n* = 1) [14; 3].

Сопоставление показателей частотности единиц корпуса позволило нам выделить следующий перечень ключевых отраслевых терминов, используемых в текстах о разработке месторождений нефти и газа (см. Таблицу 1). Наиболее высокие статистические показатели отмечены серым цветом.

Как видно из таблицы, высокие показатели относительной частоты и ключевого слова совпадают у терминов «скважина» и «пласт». Несмотря на разброс статистических показателей других элементов списка, они, несомненно, являются основными терминами данной области знания.

Таблица 1

Наиболее частотные терминологические единицы корпуса

Table 1

The most frequent corpus terminological units

№	Лемма	Относительная частота (ipm)	Показатель ключевого слова (keyness score)
1	скважина	13 914,24	695,14
2	нефть	9 216,11	152,72
3	пласт	9 135,52	801,95
4	разработка	8 026,17	41,84
5	давление	5 978,14	57,00
6	месторождение	5 878,59	210,91
7	газ	5 376,06	46,81
8	залежь	4 707,61	1 326,38
9	добыча	2 631,14	71,32
10	жидкость	1 938,99	38,91
11	проницаемость	1 332,16	381,76
12	дебит	1 303,72	905,35
13	вытеснение	1 189,94	322,63
14	закачка	1 024,01	353,32
15	коллектор	990,83	107,30
16	заводнение	327,11	792,28

В отдельную группу наиболее частотных единиц корпуса нами были выделены общенаучные термины и термины, обозначающие меры (в скобках указана ipm): *исследование* (1 924,76), *количество* (1 142,53), *коэффициент* (1 924,76), *метод* (2 688,03), *модель* (1 938,99), *объект* (2 588,48), *объем* (1 910,54), *отбор* (1 185,20), *параметр* (1 152,01), *период* (1 332,16), *плотность* (1 080,90), *площадь* (905,49), *показатель* (1 218,38), *процесс* (2 446,25), *режим* (2 450,99), *свойство* (1 341,65), *система* (3 721,52), *стадия* (1 128,31), *технология* (1 517,06), *условие* (2 417,81). Высокая частотность перечисленных единиц обусловлена типами текстов, включенных в корпус.

Также необходимо отметить, что как перечень единиц с высокой частотностью, так и список ключевых слов, сгенерированный в системе Sketch Engine, включают значительное количество прилагательных. Очевидно, что они участвуют в формировании коллокаций атрибутивного типа, которые часто можно отнести к терминам, видовым по отношению к единицам, включенным в Таблицу 1. Например, прилагательное «пластовый», отмеченное высоким показателем ключевого слова в корпусе (1 914,90), формирует такие сочетания, как «пластовая вода», «пластовая жидкость», «пластовая микрофлора», «пластовая нефть», «пластовая система», «пластовая температура», «пластовая энергия», «пластовое давление», «пластовые условия», «пластовый газ», «пластовый рассол», «пластовый флюид». На базе корпуса текстов о разработке месторождений нефти и газа были выявлены следующие прилагательные (указаны в исходной форме и перечислены по убыванию показателя относительной частоты — ipm): *пластовый* (3 194,92), *нефтяной* (2 460,47), *газовый* (2 247,14), *добывающий* (2 176,03), *нагнетательный* (1 730,39), *геологический* (1 014,53), *гидродинамический* (1 009,79), *продуктивный* (834,38), *фильтрационный* (677,93), *промысловый* (483,56), *водонапорный* (464,60), *забойный* (455,12), *газоконденсатный* (379,26), *призабойный* (355,56), *углеводородный* (284,45), *геофизический* (260,74), *площадной* (256,78), *водонефтяной* (256,00), *карбонатный* (251,26), *гидрохимический* (237,04), *безводный* (189,63), *низкопроницаемый* (151,71), *нефтегазопромысловый* (137,48), *разведочный* (137,48), *гидростатический* (118,52) и др. Анализ сочетаемости данных прилагательных позволяет выделить терминологию, периферийную по отношению к ключевой.

К инструментам извлечения ключевых единиц корпуса также можно отнести функцию «Тезаурус» системы Sketch Engine, которая демонстрирует, какие единицы имеют схожую дистрибуцию с заданным словом. Схожесть дистрибуции слов высчитывается статистически на основе меры ассоциации logDice и с учетом лексико-синтаксических шаблонов [4: 128-129]. logDice — нормализованная форма меры Dice — вычисляется по формуле

$$logDice = 14 + \log_2 D = 14 + \log_2 \frac{2f_{xy}}{f_x + f_y},$$

где x — ключевое слово; y — коллокат; f_{xy} — частота встречаемости ключевого слова x в паре с коллокатом y; f_x, f_y — абсолютные (независимые) частоты ключевого слова x и коллоката y в корпусе (тексте) [13: 9].

Дистрибутивный тезаурус группирует слова, встречающиеся в контекстах, подобных контекстам употребления искомого слова. В результате применения данной функции генерируется так называемое «облако слов» (word cloud), включающее единицы тезауруса. Чем крупнее шрифт включенного в облако слова, тем выше его значение статистической меры, отражающей степень семантической близости данного слова к ключевому [15: 71]. Также результатом обработки корпуса при помощи инструмента «Тезаурус» является сводная таблица лексем, включенных в облако в лемматизированной форме с указанием их ранга и частоты [12: 14].

В рамках данного исследования мы применили функцию «Тезаурус» к термину «разработка» (мы ограничили количество единиц тезауруса до 50-и и включили ключевое слово в облако; см. Рис. 2). Как видно на Рис. 2, ряд единиц имеют дистрибуцию, максимально приближенную к дистрибуции ключевого слова (в скобках указаны значения статистической меры logDice): *эксплуатация* (0,308), *система* (0,204), *добыча* (0,180), *применение* (0,164), *работа* (0,163), *исследование* (0,151), *процесс* (0,151), *моделирование* (0,117) и др. Построение дистрибутивного тезауруса для единицы корпуса помогает анализировать, какие слова или термины находятся в синонимических отношениях с данной единицей.

Рис. 2. «Облако слов» для термина «разработка» на базе корпуса текстов о разработке месторождений нефти и газа

Fig. 2. The “word cloud” for the term “development” based on the text corpus of oil and gas reservoirs development

Представленные в статье инструменты корпусного анализа помогают исследователю извлекать как на базе уже существующих корпусов, так и на базе специально созданного корпуса данные о частотности единиц, формировать перечни ключевых слов и терминов, анализировать единицы со сходной дистрибуцией. Интерпретация полученных нами статистических данных дает представление о терминологии, описывающей предметную область «Разработка месторождений нефти и газа».

СПИСОК ЛИТЕРАТУРЫ

1. Захаров В. П. Автоматическое выявление терминологических словосочетаний / В. П. Захаров, М. В. Хохлова // Структурная и прикладная лингвистика. Вып. 10. Изд-во С.-Петерб. ун-та, 2014. С. 182-200.

2. Захаров В. П. Автоматическое извлечение терминов из специальных текстов с использованием дистрибутивно-статистического метода как инструмент создания тезаурусов / В. П. Захаров, М. В. Хохлова // Структурная и прикладная лингвистика. Вып. 9. Изд-во С.-Петерб. ун-та, 2012. С. 222-233.
3. Захаров В. П. Корпусная лингвистика / В. П. Захаров, С. Ю. Богданова. Иркутск: ИГЛУ, 2011. 161 с.
4. Захаров В. П. Корпусно-ориентированный подход к построению тезаурусов и онтологий / В. П. Захаров // Структурная и прикладная лингвистика. Вып. 11. СПб.: Изд-во С.-Петерб. ун-та, 2015. С. 123-141.
5. Захаров В. П. Сочетаемость через призму корпусов / В. П. Захаров // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог», Москва, 27–30 мая 2015 г. Вып. 14 (21): В 2 т. Т. 1: Основная программа конференции. М.: Изд-во РГГУ, 2015. С. 667-682.
6. Копотев М. В. Введение в корпусную лингвистику / М. В. Копотев. Прага: Animedia Company, 2014. 218 с.
7. Andersen G. Evaluation of Alternative Association Measures for Extraction of Terminology Based on a Large Norwegian Corpus / G. Andersen // SYNAPS – A Journal of Professional Communication. 2011. Vol. 26. Pp. 62-68.
8. Jakubíček M. The TenTen Corpus Family / M. Jakubíček, A. Kilgarriff, V. Kovář, P. Rychlý, V. Suchomel // 7th International Corpus Linguistics Conference, Lancaster, July 2013. URL: https://www.sketchengine.co.uk/wp-content/uploads/The_TenTen_Corpus_2013.pdf
9. Kast-Aigner J. Terms in Context: A Corpus-Based Analysis of the Terminology of the European Union's Development Cooperation Policy / J. Kast-Aigner // Fachsprache – International Journal of LSP. 2009. No 3-4. Pp. 139-152.
10. Kilgarriff A. Finding Terms in Corpora for Many Languages with the Sketch Engine / A. Kilgarriff, M. Jakubíček, V. Kovář, P. Rychlý, V. Suchomel // Proceedings of the Demonstrations at the 14th Conference the European Chapter of the Association for Computational Linguistics. Sweden, April 2014. Pp. 53-56. URL: https://www.sketchengine.co.uk/wp-content/uploads/Finding_Terms_2014.pdf
11. Kilgarriff A. The Sketch Engine / A. Kilgarriff, P. Rychlý, P. Smrž, D. Tugwell // Proceedings of the XI EURALEX International Congress. Lorient: Universite de Bretagne-Sud, 2004. Pp. 105-116. URL: https://www.sketchengine.co.uk/wp-content/uploads/The_Sketch_Engine_2004.pdf
12. Kilgarriff A. The Sketch Engine: Ten Years On / A. Kilgarriff, V. Baisa, J. Bušta, M. Jakubíček, V. Kovář, J. Michelfeit, P. Rychlý, V. Suchomel // Lexicography ASIALEX. 2014. Vol. 1. Pp. 7-36. URL: http://link.springer.com/article/10.1007/s40607-014-0009-9
13. Rychlý P. A Lexicographer-Friendly Association Score / P. Rychlý // Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2008, Brno, Masaryk University, 2008. Pp. 6–9. URL: https://nlp.fi.muni.cz/raslan/2008/papers/13.pdf
14. Statistics Used in the Sketch Engine. Lexical Computing Ltd., 2015. URL: https://www.sketchengine.co.uk/wp-content/uploads/ske-stat.pdf
15. Thomas J. Discovering English with Sketch Engine: A Corpus-Based Approach to Language Exploration / J. Thomas. Versatile, 2016. 228 pp.

Marina A. KOVYAZINA[1]

KEY TERM EXTRACTION BASED ON A CORPUS OF OIL AND GAS FIELD DEVELOPMENT DISCOURSE

[1] Cand. Sci. (Philol.), Associate Professor,
Department of English Philology and Translation,
Institute of Philology and Journalism,
Tyumen State University
m.a.kovyazina@utmn.ru

Abstract

The paper presents a research targeted at term extraction based on a text corpus. The author of the research uses the corpus analysis toolkit "AntConc" and the corpus query system "Sketch Engine" to compile the corpus of texts devoted to oil and gas field development processes, stages, and methods, as well as to extract the key terminology of the domain. Several corpus methods are used to identify the terminology inherent in oil and gas field development discourse: analysing word frequency lists, generating a list of key words and terms based on keyness score, and building a distributional thesaurus with the application of the logDice coefficient. As a result of the corpus-based research, the terms synonymous with the key notion "field development" have been grouped, as well as the key domain-specific and general scientific terminology has been extracted.

Keywords

Terminology, term extraction, corpus, keyness score.

DOI: 10.21684/2411-197X-2016-2-3-61-69

REFERENCES

1. Andersen G. 2011. "Evaluation of Alternative Association Measures for Extraction of Terminology Based on a Large Norwegian Corpus". SYNAPS – A Journal of Professional Communication, vol. 26, pp. 62-68.

Citation: Kovyazina M. A. 2016. "Key Term Extraction Based on a Corpus of Oil and Gas Field Development Discourse". Tyumen State University Herald. Humanities Research. Humanitates, vol. 2, no 3, pp. 61–69.
DOI: 10.21684/2411-197X-2016-2-3-61-69

2. Jakubíček M., Kilgarriff A., Kovář V., Rychlý P., Suchomel V. 2013. "The TenTen Corpus Family". 7th International Corpus Linguistics Conference, Lancaster, July. https://www.sketchengine.co.uk/wp-content/uploads/The_TenTen_Corpus_2013.pdf
3. Kast-Aigner J. 2009. "Terms in Context: A Corpus-Based Analysis of the Terminology of the European Union's Development Cooperation Policy". Fachsprache – International Journal of LSP, no. 3-4, pp. 139-152.
4. Kilgarriff A., Baisa V., Bušta J., Jakubíček M., Kovář V., Michelfeit J., Rychlý P., Suchomel V. 2014. "The Sketch Engine: Ten Years On". Lexicography ASIALEX, vol. 1, pp. 7-36. http://link.springer.com/article/10.1007/s40607-014-0009-9
5. Kilgarriff A., Jakubíček M., Kovář V., Rychlý P., Suchomel V. 2014. "Finding Terms in Corpora for Many Languages with the Sketch Engine". Proceedings of the Demonstrations at the 14th Conference the European Chapter of the Association for Computational Linguistics, Sweden, April, pp. 53–56. https://www.sketchengine.co.uk/wp-content/uploads/Finding_Terms_2014.pdf
6. Kilgarriff A., Rychlý P., Smrž P., Tugwell D. 2004. "The Sketch Engine". Proceedings of the XI EURALEX International Congress, Lorient, pp. 105–116. https://www.sketchengine.co.uk/wp-content/uploads/The_Sketch_Engine_2004.pdf.
7. Kopotev M. V. 2014. Vvedenie v korpusnuyu lingvistiku [Introduction to Corpus Linguistics]. Prague: Animedia Company.
8. Rychlý P. 2008. "A Lexicographer-Friendly Association Score". Proceedings of Recent Advances in Slavonic Natural Language Processing, Brno, Masaryk University, pp. 6–9. https://nlp.fi.muni.cz/raslan/2008/papers/13.pdf
9. Sketch Engine. "Statistics Used in the Sketch Engine". https://www.sketchengine.co.uk/wp-content/uploads/ske-stat.pdf
10. Thomas J. 2016. Discovering English with Sketch Engine: A Corpus-Based Approach to Language Exploration. Versatile, 228 p.
11. Zakharov V. P. 2015. "Korpusno-orientirovannyiy podhod k postroeniyu tezaurusov i ontologiy" [Corpus-Based Approach to Thesaurus and Ontology Construction]. Structural and Applied Linguistics, no. 11, pp. 123-141.
12. Zakharov V. P. 2015. "Sochetaemost cherez prizmu korpusov" [Set Phrases: a View through Corpora]. Proceedings of the International Conference "Dialog 2015: Computational Linguistics and Intellectual Technologies", vol. 1, no 14 (21), pp. 667-682. Moscow: RGGU.
13. Zakharov V. P., Bogdanova S. Yu. 2011. Korpusnaya lingvistika [Corpus Linguistics]. Irkutsk: Irkutsk State Linguistic University.
14. Zakharov V. P., Khokhlova M. V. 2012. "Avtomaticheskoe izvlechenie terminov iz spetsialnyih tekstov s ispolzovaniem distributivno-statisticheskogo metoda kak instrument sozdaniya tezaurusov" [Automatic Term Extraction and Statistical Analysis in a Special Text Corpus as a Tool for Thesaurus Construction]. Structural and Applied Linguistics, no. 9, pp. 222-233.
15. Zakharov V. P., Khokhlova M. V. 2014. "Avtomaticheskoe vyiyavlenie terminologicheskih slovosochetaniy" [Automatic Extracting Terminological Phrases]. Structural and Applied Linguistics, no. 10, pp. 182-200.