Сравнение инструментов Sketch Engine и TermoStat для извлечения терминологии

А. А. Новикова

***Аннотация*—В статье приводятся результаты использования функции term extraction на базе небольшого англоязычного корпуса технических текстов предметной области «Водоснабжение». На данный момент существует большое количество специальных программных инструментов для извлечения терминологии и работы с текстами разного объема и тематики. Автоматическая обработка текстов на естественном языке предполагает использование методов корпусной лингвистики. Корпус текстов – собрание текстов, объединенных одной тематикой, представляющее своеобразную модель какого-либо языка, и используемое в качестве базы исследования языка. Существует множество специальных программ, позволяющих пользователю создавать корпуса текстов и с их помощью решать узкоспециализированные лингвистические задачи. Извлечение терминов – важная задача в работе лингвиста-лексикографа или переводчика, поскольку термины составляют огромный пласт лексики в языках для специальных целей. По причине того, что новые термины появляются постоянно, необходимо периодически дополнять и корректировать госты, словари, глоссарии и т.д. Своевременно фиксировать термины в словарях и глоссариях также необходимо для корректной работы онлайн-словарей и систем машинного перевода, более того, это значительно облегчает работу специалисту-переводчику. В этом ключе представляют интерес программы Sketch Engine и TermoStat, применительно к решению задач по извлечению терминологии.**

***Ключевые слова*—терминология, термин, извлечение терминов, корпус текстов**

I. Введение

Задача обработки текстов на естественном языке (natural language processing) возникла еще с появлением первых компьютеров. На данный момент в этой области достигнуты значительные успехи, например, в машинном переводе, информационном поиске, реферировании текстов, распознавании речи и др. С переходом от традиционных методов обработки языкового материала к компьютерным, появился новый инструментарий, позволяющий решать различные лингвистические задачи. В области машинного перевода, компьютерной лексикографии, терминоведения, индексирования и информационного поиска важной стала задача извлечения ключевых слов и терминов.

В языках для специальных целей (language for special purposes, далее – LSP), обслуживающих определенную область знаний, термины и ключевые слова встречаются в большом количестве. Такой язык очень терминологичен, иными словами, он включает огромный пласт терминологической лексики, поэтому часто LSP называют «профессиональным подъязыком», «профессиональным диалектом», «специальным подъязыком» и т. д.

В процессе развития технологий, в любой отрасли появляются новые термины и понятия, которые необходимо своевременно фиксировать в словарях или глоссариях. Стремление к унификации терминов особенно важно, так как их разобщенность представляет проблему в решении задач перевода (и машинного в том числе) с одного языка на другой. В условиях глобализации и постоянного обмена опытом растет необходимость в мгновенной передаче/обработке информации. Сейчас подобные задачи решают достаточно быстро с помощью специального программного обеспечения. К таким программным средствам относят словари (электронные и онлайн), базы данных, тезаурусы, системы машинного перевода, корпуса текстов.

Использование корпусов текстов в решении задач перевода существенно облегчает работу. Более того, корпуса текстов являются лингвистическими информационными ресурсами, хранящими знания, которые не всегда можно найти в словарях или других лексикографических ресурсах.

В многообразии жанров корпусов текстов особое место занимают корпусы специальных, прежде всего, научных текстов, отражающие знания по отдельным предметным областям. Особенности данных корпусов — наличие жёстких ограничений по типу и тематике текстов, входящих в их состав; формализованность содержания текстов, опирающегося на логико-понятийную схему предметной области; высокая структурированность словаря текстов за счёт насыщенности терминами; очевидное влияние научного стиля на лексико-семантические, морфологические, синтаксические параметры текстов в корпусе [2]. В обучении языкам, а также, переводческой деятельности, широко используют корпусные инструменты. Сейчас каждый желающий может создать свой корпус для решения специфических задач, используя различные корпус-менеджеры. К таким программам, например, относят корпусный инструмент Sketch Engine [18]. С его помощью можно сформировать корпус текстов, получить данные о сочетаемости слов, создать списки семантически связанных лексем, а также извлечь ключевые слова и термины. Также интерес представляет программа TermoStat [20]. Этот программный инструмент разработан канадскими учёными для извлечения терминов из корпуса текстов. В отличие от Sketch Engine в этой программе можно работать только с корпусом, сделанным самостоятельно.

Пользователь может загрузить свой корпус в систему, и если Sketch Engine работает с файлами разных форматов (например, *.doc, .docx, .htm, .html, .tei, .tmx,*

.txt, .vert, .xml, .pdf, .xls, .xlsx, .tmx, .xlf/.xliff, .ods, .zip, .tar.gz), то TermoStat работает только с файлами форматов *.txt* и *.rtf*, что иногда может показаться не очень удобным.

Большим преимуществом Sketch Engine является работа с большим количеством языков (более 85), в то время как TermoStat работает только с текстами на английском, французском, итальянском, испанском и португальском языках. Статистические меры, с использованием которых извлекаются термины, различны для обеих программ. Так, для Sketch Engine используются мера logDice, метод Simple Maths и различные комбинации мер [19]. Для TermoStat характерно использование меры logLikelihood, статистического критерия Chi2, функции отношения шансов Log Odds Ratio [20]. Утверждать, какая статистическая мера (или комбинация мер) работает точнее для извлечения терминов – затруднительно, во многом из-за специфики (объема и наполнения) фоновых корпусов, которые используются этими программами.

Использование таких программных средств ускоряет процесс решения соответствующих лингвистических задач.

II. Методы извлечения терминологии

Прежде чем начинать работу с терминологией, необходимо определить понятие термина. Дать универсальное определение этому понятию сложно, о чем свидетельствует наличие большого количества исследований, посвященных этой проблеме [1, 3, 4, 9, 12].

Под *термином* чаще всего понимают слово или словосочетание, являющееся точным обозначением определенного понятия какой-либо области знания [6].

Можно выделить некоторые основные особенности, которыми термин должен обладать: наличие дефиниции, однозначность, отсутствие синонимов и экспрессивности. В исследовании используется нестрогое понимание термина. Таким образом, термин представляется лексической единицей, характерной для некоего текста или множества текстов [10]. Иными словами, каждая лексическая единица текста рассматривается как потенциальный термин. Отметим, что техническая терминология имеет преимущественно именной характер, поскольку основу любой технической терминосистемы, составляют термины-существительные, представляющие статическую часть словарного состава и передающие родовидовые отношения как отношения системы понятий [1]. К методам извлечения терминологии обычно относят: лингвистические, статистические и гибридные методы.

Для лингвистических методов характерна ручная обработка текстов и документов в специальном корпусе. Эксперты должны выявить выражения, которые рассматриваются как термины/терминосочетания. Для выделения терминосочетаний рекомендуется использовать лексико-грамматические шаблоны однословных и неоднословных терминов. Целесообразно также использовать систему фильтров (стоп-словарь) для отсеивания нетерминов. Лексико-грамматический шаблон – это структурный образец (модель) языковой конструкции, в котором указываются существенные грамматические характеристики множества лексем, которые входят в языковые выражения, принадлежащие данному классу, и синтаксические условия употребления языкового выражения, построенного в соответствии с шаблоном (например, правила согласования морфологических признаков лексем) [8].

Использование статистических методов предполагает понимание терминов как наиболее частотных слов и словосочетаний, встречающихся в специальных текстах и выражающих понятия предметной области. Терминологические сочетания чаще всего соотносят с *n*-граммами (двух-, трех-, четырехчленными сочетаниями). Терминологическим n-граммам свойственна высокая степень устойчивости. Для оценки устойчивости словосочетаний в специальных текстах применяют статистические меры: MI-score, t-score, Log-Likelihood, C-value, критерий χ^2 и др. [10, 13].

Гибридные методы предполагают использование лингвистических и статистических методов и дают наиболее точный результат.

III. Подготовка

Исследование выполняется в сфере терминоведения и научно-технического перевода, поэтому внимание уделяется именно лексическому составу предметной области «Водоснабжение». Терминология этой предметной области интересна тем, что объединяет термины и понятия предметных областей строительства, машиностроения, химии, биологии, экологии, гидротехники и многих других. Более того, эта терминология является довольно специфичной.

Подготовительный этап работы предполагает определение источников, в которых содержится терминология указанной предметной области. К таким источникам относят: учебники, энциклопедии, технические документы, словари и журналы. Анализ терминологии также предполагает анализ моделей терминообразования. Это дает возможность выработать рекомендации по образованию новых терминов, а также вносит существенный вклад в разработку программ типа Automatic Term Extraction [11]. Для исследования выбраны программные инструменты Sketch Engine и TermoStat

В целях сравнения функционала двух программных средств Sketch Engine и TermoStat, был создан небольшой корпус англоязычных текстов по тематике «Водоснабжение» объёмом 311 643 слов (токенов).

Из созданного корпуса были извлечены термины-кандидаты с помощью программ Sketch Engine и TermoStat, получены ранжированные списки с указанием частоты вхождений кандидата в тексте.

IV. Извлечение терминов с помощью Sketch Engine

Sketch Engine – программное обеспечение, созданное в 2003 году командой разработчиков во главе с А. Килгарифом, для анализа текстовых данных и управления корпусами текстов [18]. Эта система содержит более 500 текстовых корпусов, созданных более чем на 85 языках. Для исследования языковых процессов эту базу данных (состоящую из миллиардов

слов) используют различные компании, от университетов до IT-компаний по всему миру. Система обладает дружелюбным интерфейсом и возможностью всестороннего анализа языкового материала, загруженного пользователем.

Разработчики программы указывают, что термины-кандидаты могут быть выделены, учитывая следующие условия [16]:

• созданный пользователем корпус является главным (для извлечения терминов-кандидатов), а фоновый корпус представляет собой модель языка;
• определение грамматической формы термина-кандидата;
• токенизация, лемматизация и разметка по частям речи необходима для обоих корпусов;
• определение (и подсчёт) языковых единиц, совпадающих по грамматической форме, в каждом корпусе;
• сравнение частоты каждой языковой единицы в главном корпусе с её частотой в фоновом корпусе.

Термин рассматривается, с одной стороны, как концепт, с другой, как многословное выражение, состоящее из нескольких токенов. При этом, частота в главном корпусе будет выше, чем в фоновом. В языке такое многословное выражение как раз выполняет функцию термина. Как правило, это определяется терминологической грамматикой (term grammar), уникальной для каждого языка. Term grammar – это свод правил, написанных на специальном на основе языка регулярных выражений, который определяет лексические структуры и именные группы и использует набор тегов для морфологической разметки [19].

Говоря о извлечении терминологии из корпуса с использованием Sketch Engine, отметим, что в этой системе для этих целей создана функция «Keywords and terms». Методом определения ключевых слов и терминов является метод «Simple maths», который предполагает выделение ключевых слов на базе двух корпусов – главного (созданного пользователем) и фонового (встроенного в систему). Пользователь системы может переключаться между корпусами, таким образом, фокусируясь на словах с высокой или низкой частотой. Единицам с высокой частотой система присваивает статистическую меру ключевого слова (keyness score), которая вычисляется по формуле:

$$\text{fpm } \{\text{rm focus}\} + N/ \text{ fpm\_}\{\text{rm ref}\} + N \qquad (1)$$

где fpm_{rm focus} – нормализованная частота слова в главном корпусе, fpm_{rm ref} – нормализованная частота слова в фоновом корпусе, а N – коэффициент сглаживания, по умолчанию равный 1 [18].

После загрузки созданного корпуса в систему и применения функции «Keywords and terms» получаем ранжированный список терминов-кандидатов (таблица 1) с указанием абсолютной и относительной частоты в главном и фоновом корпусах, а также коэффициентом семантической близости (Score).

Таб. 1. Термины-кандидаты (Sketch Engine)

Item	Score	Freq	Ref_ freq	Rel_ freq	Rel_ref_freq
raw water	305,530	132	28	340,273	0,118
distribution system	215,020	149	188	384,096	0,791
water level	182,400	132	207	340,273	0,871
slow sand	178,870	69	0	177,870	0,000
rapid gravity	151,280	59	3	152,092	0,013
water treatment	147,200	135	326	348,006	1,372
ductile iron	141,650	55	2	141,780	0,008
critical depth	129,370	50	1	128,891	0,004
public water	125,020	63	73	162,403	0,307
mm diameter	124,580	52	20	134,047	0,084
head loss	106,270	41	1	105,691	0,004
calcium carbonate	105,470	51	61	131,469	0,257
service reservoir	104,110	40	0	103,113	0,000
chlorine dioxide	102,790	42	15	108,269	0,063
water supply	96,760	151	721	389,252	3,034

Функция «Thesaurus» также позволяет извлекать ключевые слова из корпуса и показывает единицы, имеющие схожую дистрибуцию с заданным словом. Схожесть дистрибуции слов высчитывается статистически на основе меры ассоциации logDice (нормализованной меры Dice) и с учетом лексико-синтаксических шаблонов [8]. Формула выглядит следующим образом:

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

где x — ключевое слово; y — коллокат; f_{xy} — частота встречаемости ключевого слова x в паре с коллокатом y; f_x, f_y — абсолютные частоты ключевого слова x и коллоката y в корпусе [17]. В таблице 2 приведен список терминов-кандидатов, полученных с использованием функции «Thesaurus». Дистрибутивный тезаурус строится на основе общих коллокатов – если два слова в корпусе имеют много общих коллокатов, они будут включены в тезаурус для ключевого слова [17]. На примере ключевого слова *water* был построен тезаурус.

Таб. 2. Результат использования функции «Thesaurus»

Word	Similarity score	Frequency
flow	0,180	1267
supply	0,177	761

pressure	0,129	717
level	0,117	688
system	0,114	747
pipe	0,112	1186
reservoir	0,097	543
quality	0,081	295
filter	0,080	600
pump	0,079	535

В итоге получен список слов, имеющих схожую дистрибуцию со словом *water*, которые также являются терминами-кандидатами.

С точки зрения эксперта, можно отметить, что, в целом, Sketch Engine извлекает термины достаточно точно.

V. Извлечение терминов с помощью TermoStat

TermoStat – проект, разработанный под руководством канадского исследователя Патрика Друэна [20]. Программа позволяет получать статистическую информацию о извлеченных терминах, список биграмм, список семантически связанных слов. При этом указывается терминологическая модель (pattern). Функционал программы более скромный по сравнению со Sketch Engine. Термины-кандидаты и схожесть дистрибуции слов определяются с помощью статистического критерия Chi2, меры logLikelihood, функции отношения шансов Log Odds Ratio. Вычисление разницы между относительными частотами терминов-кандидатов в главном и фоновом корпусах определяется функцией Specificity (определенность). Это значение по умолчанию, используемое для идентификации терминов-кандидатов [15]. Для каждого кандидата программа позволяет посмотреть значения Score (коэффициент семантической близости), используя встроенную функцию переключения указанных статистических критериев.

После загрузки корпуса в систему, были получены списки терминов-кандидатов (Candidates) с указанием частоты, вариантов употребления и терминологической модели (Pattern). Программа учитывает не только структуру потенциальных терминов-кандидатов, но и относительные частоты этих потенциальных кандидатов в обрабатываемом тексте (главный корпус) и очень большой коллекции газетных статей (фоновый корпус). TermoStat, тем не менее, допускает использование других мер для идентификации терминов-кандидатов, что позволяет сравнивать результаты различных подходов к извлечению терминов, исходя из задач исследования.

Для определения точного количества вхождений терминов-кандидатов в тексте, используется лемматизация – приведение склоняемых форм слов к базовым (например, преобразование формы множественного числа существительных в форму единственного числа). После этого все формы можно считать вхождениями одного термина-кандидата, а не разными терминами. В таблице 3 представлен список терминов-кандидатов.

Таб. 3. Термины-кандидаты (TermoStat)

Candidates	Freq	Score
water	3957	217.66
pipe	1178	144.96
flow	1267	137.68
filter	598	100.39
reservoir	541	97.73
valve	541	97.11
pump	534	95.36
chlorine	451	90.67
supply	757	85.59
main	349	81.33
dam	386	77.8
filtration	295	74.76

TermoStat показывает информацию о количестве терминов-кандидатов, построенных по определенной модели, и процент таких терминов-кандидатов от общего числа терминов в корпусе.

Термины-кандидаты с наибольшей частотой имеют следующие терминологические модели:

- Noun + Noun = 2013 (34 %);
- Adjective + Noun = 1492 (25 %);
- Noun = 1405 (24 %).

Например, термин *water level* образован по модели Noun + Noun, термин *raw water* по модели Adjective + Noun, а термин *water* – по модели Noun, соответственно.

Наряду с однословными терминами, с помощью TermoStat можно выделить биграммы и многословные термины. Также система позволяет построить тезаурус с помощью функции «Structuration» (рис. 1), где программа учитывает информацию о частоте термина и показывает все включения термина-кандидата в корпусе (term inclusion).

Рисунок 1. Результат использования функции «Structuration».

Термин-кандидат с самыми высокими показателями – *water* (3957). В третьем столбце на рисунке 1 показаны все слова, сочетающиеся с термином *water,* и при этом также являющиеся терминами-кандидатами. Отметим, что в столбце «term inclusion» показаны термины-кандидаты разного ранга, что позволяет обратить внимание на термины-кандидаты, которым присвоена небольшая частота.

Отметим, что программа позволяет посмотреть все контексты употребления того или иного термина-кандидата, то есть, построить конкорданс.

VI. Сравнение результатов работы Sketch Engine и TermoStat

Sketch Engine и TermoStat являются популярными программами для работы с корпусами текстов и извлечения терминологии. С их помощью можно извлекать термины-кандидаты разной длины (например, 1-компонентные, 2-компонентные и т.д.), составлять ранжированные списки семантически связанных слов, строить тезаурус и просматривать контексты употребления слов. Программы доступны в онлайн-режиме и обладают дружелюбным интерфейсом. Однако у каждой программы есть свои особенности (перечислим некоторые из них):

- для идентификации ключевых слов и терминов-кандидатов используются разные статистические меры;
- объем и наполнение фонового корпуса для каждой программы разный;
- программы работают с разным количеством языков (например, TermoStat работает с ограниченным количеством языков (5), а Sketch Engine – более, чем с 85 языками).

В таблице 4 приведены первые десять терминов из первой тысячи терминов. Любопытно, что термину *pump* в Sketch Engine присвоен ранг 6, в то время как TermoStat показывает, что в первые десять терминов этот термин не попал. Разница между рангами у терминов *filter* и *reservoir* составляет 1.

Таб. 4. Однокомпонентные термины

Rank	**Sketch Engine**	Freq	**TermoStat**	Freq
1	water	3990	water	3957
2	flow	1347	flow	1267
3	pipe	1216	pipe	1178
4	m	939	m	771
5	supply	890	supply	757
6	pump	800	system	747
7	pressure	724	pressure	715
8	filter	676	level	683
9	reservoir	543	filter	598
10	valve	542	reservoir	541

Далее, в таблице 5 сравниваются результаты извлечения двухкомпонентных терминов.

Таб. 5. Двухкомпонентные термины

Rank	**Sketch Engine**	Freq	**TermoStat**	Freq
1	water supply	151	water supply	200
2	distribution system	149	distribution system	167
3	water quality	146	raw water	153
4	water treatment	135	water treatment	134
5	water level	132	water level	132
6	raw water	132	water quality	130
7	drinking water	124	surface water	79
8	slow sand	69	sand filter	77
9	public water	63	slow sand	70
10	rapid gravity	59	service reservoir	69

У двухкомпонентных терминов ранги различаются – термину *raw water* присвоены ранги 6 и 3 соответственно. Это связано с тем, что программы присваивают разную информацию о частоте термина в корпусе.

Для оценки результатов работы программ вычислена точность и полнота.

Полнота – способность системы извлекать все термины из корпуса – вычисляется как отношение количества извлеченных кандидатов в термины к общему количества терминов в корпусе. Точность – способность системы отличать термины от нетерминов – рассчитывается через отношение количества извлеченных терминов к количеству извлеченных кандидатов в термины [14].

TermoStat автоматически выделил 5968 терминов-кандидатов. После проведенной экспертной оценки, установлено, что программа неверно выделила 700 терминов. Таким образом, точность составляет 85%, а полнота – 88%.

Sketch Engine не показывает количество выделенных терминов, поэтому было проанализировано так же 5968 терминов-кандидатов. Экспертная оценка показала, что программа неверно выделила 840 терминов. Соответственно, для Sketch Engine точность составляет 71%, а полнота 85%.

Таким образом, можно утверждать, что обе программы, несмотря на различия в своей архитектуре, точно извлекают термины, и охарактеризовать полученные результаты как положительные.

VII. Заключение

В исследовании сравнивались инструменты Sketch Engine и TermoStat применительно к задачам по извлечению терминологии.

В целом, обе программы хорошо зарекомендовали себя и являются довольно популярными инструментами обработки и извлечения лингвистической информации. Программы позволяют получить ранжированные списки терминов-кандидатов для дальнейшей оценки экспертом в данной предметной области.

Обе программы являются хорошими инструментами для решения соответствующих лингвистических задач, во многом благодаря использованию гибридных методов (комбинации статистических и лингвистических методов соответственно), несмотря на различия по показателям точности и полноты. TermoStat выделяет термины с большей точностью (85%).

Интересно отметить, что при использовании одного и того же корпуса, Sketch Engine и TermoStat присваивают идентичным терминам разные данные о частоте (и, соответственно, разные данные о ранге). Определение причин этой особенности может являться одним из направлений дальнейшего исследования.

Для получения наиболее точных результатов корпус будет дополнен, а также в него будут включены еще два подкорпуса – на немецком и на русском языках, что позволит проводить сопоставительные исследования, находить и анализировать переводные эквиваленты для трех языков.

Полученные данные будут использованы для составления многоязычного глоссария терминов предметной области «Водоснабжение», который будет включать термины на английском, немецком и русском языках.

Библиография

[1] Гаврилова И. А. К вопросу определения сущности термина (на материале английской полиграфической терминологии) // В мире науки и искусства: вопросы филологии, искусствоведения и культурологии: сб. ст. по матер. V междунар. науч.-практ. конф. – Новосибирск: СибАК, 2011. [Электронный ресурс]. URL: https://sibac.info/conf/philolog/v/27647 (дата обращения: 20.09.2020).

[2] Герд А. С. Прикладная лингвистика / А. С. Герд. - СПб.: Изд-во СПбГУ, 2005. - 267 с.

[3] Гринев С. В. Введение в терминоведение. – М.: Московский лицей, 1993. – 309 с.

[4] Гринев-Гриневич, С.В. Терминоведение: Учеб. пособие. – М.: Академия, 2008. – 303 с.

[5] ГОСТ 30813-200. Вода и водоподготовка. Термины и определения. [Электронный ресурс]. URL: https://files.stroyinf.ru/Data2/1/4294817/4294817020.htm (дата обращения: 20.09.2020).

[6] ГОСТ 7.0-99. Система стандартов по информации, библиотечному и издательскому делу. Информационно-библиотечная деятельность, библиография. Термины и определения. [Электронный ресурс]. URL: http://docs.cntd.ru/document/gost-7-0-99 (дата обращения: 20.09.2020).

[7] Захаров В. П., Хохлова М. В. (2014). Автоматическое выявление терминологических словосочетаний. Структурная и прикладная лингвистика, (Вып.10), 182–200.

[8] 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.

[9] Лейчик В.М. Терминоведение: предмет, методы, структура. – М.: Изд-во ЛКИ, 2007. – 256 с.

[10] Митрофанова О. А., Захаров В. П. Автоматизированный анализ терминологии в русскоязычном корпусе текстов по корпусной лингвистике // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009». – М.: РГГУ, 2009. – С. 321–328. – [Электронный ресурс]. URL: http://www.dialog-21.ru/dialog2009/materials/pdf/49.pdf (дата обращения: 20.09.2020).

[11] Соловьева А. Е. Терминология военной вертолетной авиации как объект лингвистического исследования (на примере английского, русского и турецкого языков). Филологические науки. Вопросы теории и практики. Тамбов: Грамота, 2018. No 4(82). Ч. 1. С. 172-176. – [Электронный ресурс]. URL: https://www.gramota.net/materials/2/2018/4-1/40.html (дата обращения: 20.09.2020).

[12] Суперанская А. В., Подольская Н. В., Васильева Н. В. Общая терминология: Вопросы теории. М.: ЛИБРОКОМ, 2012. — 248 с.

[13] Хохлова М. В. Сопоставительный анализ статистических мер на примере частеречных преференций сочетаемости существительных // Компьютерная лингвистика и вычислительные онтологии. Выпуск 1 (Труды XX Международной объединенной конференции «Интернет и современное общество, IMS-2017, Санкт-Петербург, 21 - 23 июня 2017 г. Сборник научных статей). СПб: Университет ИТМО, 2017. С. 165-174.

[14] Cabré, M. T., Estopà, R., Vivaldi, J. Automatic term detection: a review of current systems // Bourigault, D.; Jacquemin, C.; L'Homme, M-C. (2001) Recent Advances in Computational Terminology, p. 53-88.

[15] Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology*, 1(9):99– 115.

[16] Kilgariff A., Jakubíček M., Kovář V., Rychlý P., Suchomel V. Finding Terms in Corpora for Many Languages with the Sketch Engine // Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics. – Goethenburg, 2014. – [Электронный ресурс]. URL: https://www.sketchengine.co.uk/wpcontent/uploads/Finding_Terms_2014.pdf (дата обращения: 20.09.2020).

[17] 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 (дата обращения: 20.09.2020).

[18] Sketch Engine. – [Электронный ресурс]. URL: https://www.sketchengine.eu/ (дата обращения: 20.09.2020).

[19] Statistics Used in the Sketch Engine. Lexical Computing Ltd., 2015. URL: https://www.sketchengine.co.uk/wp-content/uploads/ske-stat.pdf (дата обращения: 20.09.2020).

[20] TermoStat. – [Электронный ресурс]. URL: http://termostat.ling.umontreal.ca/index.php (дата обращения: 20.09.2020).

Статья получена 9 октября 2020.

А. А. Новикова, аспирант кафедры математической лингвистики Санкт-Петербургского государственного университета (e-mail: alexa.novikova707@gmail.com).

Sketch Engine and TermoStat tools for automatic term extraction

A. A. Novikova

***Abstract* – The paper considers comparison of Sketch Engine and TermoStat tools in automatic term extraction for a small English-language corpus. The corpus includes technical texts of "Water supply" subject domain. Nowadays there are a lot of specialized programme tools for term extraction purposes. These tools also allow to work with large text data. The methods of corpus linguistics are often used in natural language processing. Text corpus is a collection of texts with specific genre. It could represent a so-called "language model" in general and is often used as a basis for language research. There are many special programme tools which allow to create different corpora. Different specific linguistic problems could be solved using text corpora, especially term extraction which is very important task for linguists, lexicographers and terminologists to solve. It is very important to correct and improve terminological standarts, glossaries, dictionaries, etc. because of new terms which appear permanently. It is also important for stable work of online-dictionaries and machine translation systems, and it also helps translation specialists. The results of using term extraction tools with small English text corpus of water supply subject domain are discussed in the paper.**

***Keywords* – terminology, term, automatic term extraction, text corpus**

REFERENCES

[1] Gavrilova I. A. K voprosu opredelenija sushhnosti termina (na materiale anglijskoj poligraficheskoj terminologii) // V mire nauki i iskusstva: voprosy filologii, iskusstvovedenija i kul'turologii: sb. st. po mater. V mezhdunar. nauch.-prakt. konf. – Novosibirsk: SibAK, 2011. [Online]. URL: https://sibac.info/conf/philolog/v/27647 (request date: 20.09.2020).
[2] Gerd A. S. Prikladnaja lingvistika [Applied linguistics]. Saint-Petersburg: izd. SPbGU Publ., 2005. 267 p.
[3] Grinev C. B. Vvedenie v terminovedenie. – M.: Moskovskij licej, 1993. – 309 p.
[4] Grinev-Grinevich, S.V. Terminovedenie: Ucheb. posobie. – M.: Akademiya, 2008. – 303 p.
[5] GOST 30813-200. Voda i vodopodgotovka. Terminy i opredeleniya. [Online]. URL: https://files.stroyinf.ru/Data2/1/4294817/4294817020.htm (Request date: 20.09.2020).
[6] GOST 7.0-99. Sistema standartov po informacii, bibliotechnomu i izdatel'skomu delu. Informacionno-bibliotechnaya deyatel'nost', bibliografiya. Terminy i opredeleniya. [Online]. URL: http://docs.cntd.ru/document/gost-7-0-99 (Request date: 20.09.2020).
[7] Zakharov, V. P., Khokhlova, M. V. (2014). Avtomaticheskoe vyjavlenie terminologicheskih slovosochetanij. Strukturnaja i prikladnaja lingvistika, (Vyp.10), 182–200.
[8] Zakharov V. P. (2015). Korpusno-orientirovannji podhod k postrojeniju tezaurusov i ontologij [Corpus-based approach to thesaurus and ontology construction]. Strukturnaja i prikladnaja lingvistika. Vip. 11. SPb. 2015. P. 123-141.
[9] Lejchik V.M. Terminovedenie: predmet, metody, struktura. – M.: Izd-vo LKI, 2007. – 256 p.
[10] Mitrofanova O.A., Zakharov V.P. Avtomatizirovannyj analiz terminologii v russkojazychnom korpuse tekstov po korpusnoj lingvistike // Komp'juternaja lingvistika i intellektual'nye tehnologii: Po materialam ezhegodnoj Mezhdunarodnoj konferencii «Dialog 2009». – M.: RGGU, 2009. – S. 321–328. – [Online]. URL: http://www.dialog-21.ru/dialog2009/materials/pdf/49.pdf (request date: 20.08.2020).
[11] Solov'eva A. E. Terminology of military helicopter aviation as an object of linguistic study (by the example of the English, Russian and Turkish languages) // Philology. Theory and Practice. – Tambov: Gramota, 2018. №4(82). Vol. 1. P. 172-176. [Online]. URL: https://www.gramota.net/materials/2/2018/4-1/40.html (request date: 20.09.2020).
[12] Superanskaya A. V., Podol'skaya N. V., Vasil'eva N. V. Obshchaya terminologiya: Voprosy teorii. M.: LIBROKOM, 2012. — 248 p.
[13] Khokhlova M. V. Sopostavitel'nyj analiz statisticheskih mer na primere chasterechnyh preferencij sochetaemosti sushchestvitel'nyh // Komp'yuternaya lingvistika i vychislitel'nye ontologii. Vypusk 1 (Trudy XX Mezhdunarodnoj ob"edinennoj konferencii «Internet i sovremennoe obshchestvo, IMS-2017, Sankt-Peterburg, 21 - 23 iyunya 2017 g. Sbornik nauchnyh statej). SPb: Universitet ITMO, 2017. S. 165-174.
[14] Cabré, M. T., Estopà, R., Vivaldi, J. Automatic term detection: a review of current systems // Bourigault, D.; Jacquemin, C.; L'Homme, M-C. (2001) Recent Advances in Computational Terminology, p. 53-88.
[15] Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology*, 1(9):99– 115.
[16] Kilgariff A., Jakubíček M., Kovář V., Rychlý P., Suchomel V. Finding Terms in Corpora for Many Languages with the Sketch Engine // Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics. – Goethenburg, 2014. – [Online]. URL: https://www.sketchengine.co.uk/wpcontent/uploads/Finding_Terms_2014.pdf (request date: 20.09.2020).
[17] 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. [Online]. URL: https://nlp.fi.muni.cz/raslan/2008/papers/13.pdf (request date: 20.09.2020).
[18] Sketch Engine. [Online]. URL: https://www.sketchengine.eu/ (request date: 20.09.2020).
[19] Statistics Used in the Sketch Engine. Lexical Computing Ltd., 2015. – [Online]. URL: https://www.sketchengine.co.uk/wp-content/uploads/ske-stat.pdf (request date: 20.08.2020).
[20] TermoStat. [Online]. URL: http://termostat.ling.umontreal.ca/index.php (request date: 20.09.2020).