

Научная статья

УДК 81'33

DOI: <https://doi.org/10.18721/JHSS.12405>

ВЫБОР ТЕРМИНОЛОГИЧЕСКИХ ЭКСТРАКТОРОВ ДЛЯ ВЫЯВЛЕНИЯ НОМИНАЦИЙ ПОНЯТИЙ ЯЗЫКОВОЙ ПОЛИТИКИ В ТЕКСТАХ ОФИЦИАЛЬНЫХ ДОКУМЕНТОВ ЕВРОПЕЙСКОГО СОЮЗА

Е.Ю. Гацук  

Гродненский государственный университет им. Янки Купалы,
г. Гродно, Республика Беларусь

 kadam@tut.by

Аннотация. Инвентаризация терминологии реализуется, как правило, в автоматизированном режиме с использованием специализированного программного обеспечения. Многообразие терминологических экстракторов требует разработки критериев, позволяющих осуществлять их выбор для решения конкретных исследовательских задач. Цель данной статьи – осуществить сопоставительный анализ терминологических экстракторов с точки зрения их доступности и результативности извлечения потенциальных терминов из специального текста для решения конкретной исследовательской задачи, а именно инвентаризации номинаций понятий языковой политики из текстов официальных документов Европейского Союза. В качестве методологической базы использован комплекс современных научных методов: таксономический, объяснительное описание, обобщение, сопоставительный анализ. Осуществлено сопоставление 4 свободно доступных терминологических экстракторов, рекомендованных для работы отделом по координации терминологии Европейского союза TermCoord, и 1 программного инструмента, выбранного на основании сведений об его эффективности, представленных в научных публикациях. Анализ критериев, заявленных разработчиками анализируемого программного обеспечения, позволил выделить 5 таксонов для сопоставления терминологических экстракторов. Таксономический анализ позволил выявить оптимальные по критериям инструменты: онлайн-экстрактор OneClick Terms и инструмент корпусного запроса Sketch Engine, которые затем были сопоставлены с точки зрения их результативности в решении исследовательской задачи. Для проверки терминологических экстракторов на эффективность результаты их работы были сопоставлены со списком терминов, извлеченных вручную, что затем позволило применить критерии полноты и точности, традиционно используемые в информационном поиске, для сравнения результативности экстракторов. С учетом конкретной исследовательской задачи наиболее важной характеристикой является полнота, и по этому показателю оптимальным экстрактором оказался инструмент корпусного запроса Sketch Engine. Таким образом, в данной статье представлен комплексный подход к определению эффективности терминологических экстракторов не с точки зрения извлечения терминов, отражающих понятия определенной предметной области, а с точки зрения их результативности для решения конкретной исследовательской задачи.

Ключевые слова: терминологический экстрактор, специальная номинация, таксон, универбы, полилексы, Sketch Engine, OneClick Terms.

Для цитирования: Гацук Е.Ю. Выбор терминологических экстракторов для выявления номинаций понятий языковой политики в текстах официальных документов Европейского Союза // Общество. Коммуникация. Образование. 2021. Т. 12. № 4. С. 60–80. DOI: 10.18721/JHSS.12405

Статья открытого доступа, распространяемая по лицензии CC BY-NC 4.0 (<https://creativecommons.org/licenses/by-nc/4.0/>).

Scientific article

DOI: <https://doi.org/10.18721/JHSS.12405>

SELECTION OF TERM EXTRACTORS TO IDENTIFY NOMINATIONS OF LANGUAGE POLICY CONCEPTS IN THE TEXTS OF OFFICIAL DOCUMENTS OF THE EUROPEAN UNION

E.Yu. Hatsuk ✉ Yanka Kupala State University of Grodno,
Grodno, Republic of Belarus✉ kadam@tut.by

Abstract. Term extractors are automatic tools that help identify term candidates in a corpus. The diversity of term extractors requires the development of criteria that allow their selection for specific research tasks. The purpose of this article is to carry out a comparative analysis of term extractors in terms of their accessibility and effectiveness when extracting term candidates from the corpora to solve a specific research problem, namely the inventory of nominations of language policy concepts from the texts of official documents of the European Union. The study is based on a set of modern scientific methods, namely taxonomic method, explanatory description, generalization, comparative analysis. The study analyses 5 term extractors, namely, AntConc, fivefilters.org, OneClick Terms, TerMine, Terminology Extraction and corpus query tool Sketch Engine. The taxonomic analysis identified the optimal tools according to the criteria: the online extractor OneClick Terms and the corpus query tool Sketch Engine. These tools were then compared in terms of solving the research problem mentioned above. In order to test the term extractors in terms of their effectiveness, the results were compared with a list of manually extracted terms, which then allowed the application of the criteria of completeness and accuracy traditionally used in information retrieval to compare the performance of the extractors. Given the specific research objective of the term inventory, completeness was the most important characteristic and in this respect the corpus query tool Sketch Engine proved to be the optimal extractor. Thus, this paper presents a comprehensive approach to determining the effectiveness of terminological extractors not in terms of extracting terms that reflect the concepts of a particular subject area, but in terms of their effectiveness for solving a specific research problem.

Keywords: corpora, term extractor, term candidate, taxonomy, Sketch Engine, OneClick Terms.

Citation: E.Yu. Hatsuk, Selection of term extractors to identify nominations of language policy concepts in the texts of official documents of the European Union, *Society. Communication. Education*, 12 (4) (2021) 60–80. DOI: 10.18721/JHSS.12405

This is an open access article under the CC BY-NC 4.0 license (<https://creativecommons.org/licenses/by-nc/4.0/>).

Введение

С 1980-х годов XX века автоматическое извлечение терминов¹, или обозначений, которые с помощью лингвистических средств представляют общее понятие², является важным этапом при обработке текстов, принадлежащих предметной области. По мнению исследователей, автоматическое извлечение терминов, под которым понимают «процессы для систематического извлечения соответствующих терминов и их вариантов из корпуса документов» [1, с. 120] (здесь и далее перевод с английского языка наш – Е.Г.), «может улучшить широкий спектр таких задач, как изучение онтологий, машинный перевод, перевод с помощью компьютера, построение тезауруса,

¹ Очевидно, что весь процесс извлечения терминов следует считать полуавтоматическим, поскольку результат работы программного обеспечения – это всего лишь список кандидатов в термины, а реальные термины должны быть подтверждены терминологами [4, с. 726].

² Terminology work and terminology science – Vocabulary [Electronic resource] : ISO 1087:2019(en). 2019. URL: <https://www.iso.org/obp/ui/#iso:std:iso:1087:ed-2:v1:en:sec:3.4.2> (дата обращения: 20.12.2019).

классификация, индексирование, поиск информации, а также анализ текста и автоматическое обобщение» [2, с. 122–123].

Автоматическое извлечение терминологии реализуется с использованием терминологических экстракторов, или специализированного программного обеспечения, которое «пытается автоматически идентифицировать все потенциальные термины в корпусе и представить список кандидатов пользователю для верификации» [3, с. 222]. Существенное преимущество терминологических экстракторов состоит не столько в экономии времени при извлечении максимального количества специальных единиц, сколько в «возможности задавать различные критерии поиска, что позволяет адаптировать поисковый запрос к конкретной задаче..., а также сузить поиск и отфильтровать результаты в зависимости от того, что ищут пользователи» [5, с. 247].

Анализ научной литературы, посвященной рассмотрению различных терминологических экстракторов, позволил выявить 3 направления исследований. Представители первого направления [6–9], описывают возможности и алгоритм работы с терминологическими экстракторами для извлечения универбов, или однословных потенциальных специальных номинаций. Представители второго направления описывают возможности и алгоритм работы с терминологическими экстракторами для извлечения и универбов, и полилексов, или многокомпонентных потенциальных специальных номинаций [10–19]. Представители третьего направления описывают возможности и алгоритм работы с сопоставимыми корпусами [5], [20–22]. Данное направление исследований ориентировано на извлечение информации о стандартизованной и рекомендованной терминологии и на «обеспечение работ по переводу научно-технической литературы и документации» [23, с. 284].

Анализ научных работ также показал, что, как правило, в них описаны алгоритмы работы с отдельными терминологическими экстракторами, в то время как многообразие программного обеспечения, предназначенного для извлечения терминологии из специальных текстов, требует разработки объективных критериев, позволяющих осуществлять выбор терминологических экстракторов для решения конкретных задач инвентаризации терминологии.

Цель данной статьи – осуществить сопоставительный анализ терминологических экстракторов с точки зрения их доступности и результативности извлечения потенциальных терминов из специального текста для решения конкретной исследовательской задачи, а именно инвентаризации номинаций понятий языковой политики из текстов официальных документов Европейского Союза.

Методология и методика исследования

В качестве методологической базы использован комплекс современных научных методов: таксономический, объяснительное описание, обобщение, сопоставительный анализ.

Терминологические экстракторы являются незаменимым ресурсом на этапе инвентаризации терминологии, но отличаются по функционалу. Следовательно, выбор терминологического экстрактора необходимо осуществлять исходя из опций, которые помогут достичь желаемого результата при минимальных временных затратах и не требуют специальной технической подготовки от пользователя.

В свободном доступе имеется достаточное количество программного обеспечения для извлечения терминов, но с целью анализа наиболее апробированных средств они были отобраны из перечня терминологических экстракторов, представленных на официальном сайте отдела по координации терминологии Европейского союза TermCoord³. Одной из основных задач TermCoord является содействие исследованию терминологии и управление терминологией в подразделениях Европейского союза, осуществляющих перевод, а также развитие терминологической базы данных IATE (InterActive Terminology for Europe). В рамках выполнения данных задач сотрудни-

³ Free Term Extractors [Electronic resource]. URL: <https://termcoord.eu/free-term-extractors/#> (дата обращения: 30.10.2021).



ки TermCoord ищут новую терминологию, разрабатывают глоссарии и дефиниции терминов для законодательных процедур, выполняют проекты с участием терминологов и стажеров подразделений перевода, организуют тренинги и семинары по вопросам, связанным с терминологией, поддерживают контакты с университетами, терминологическими организациями и экспертами для обмена знаниями, опытом и материалами. Кроме того, данное структурное подразделение оказывает помощь переводчикам, предоставляя им различные терминологические и документационные инструменты и ресурсы⁴.

На сайте TermCoord представлен список из 11 терминологических экстракторов⁵: AlchemyAPI keyword⁶, AntConc⁷, Apsic Xbench⁸, Araya Bilingual Term Extraction Tool⁹, fivefilters.org¹⁰, KEA¹¹, «maui-indexer»¹², TerMine¹³, Terminology Extraction¹⁴, topia.termextract¹⁵, WebCorp¹⁶, Wordfast¹⁷. Материал исследования, результаты которого представлены в данной статье, составили не все представленные программные инструменты. Так, доступ к официальным страницам таких экстракторов, как AlchemyAPI keyword и «maui-indexer» оказался закрытым; экстракторы Apsic Xbench, Araya Bilingual Term Extraction Tool и Wordfast ориентированы на работу с многоязычными ресурсами, что не соответствует задачам исследования, для которого проводится выбор экстрактора; KEA и topia.termextract были исключены, поскольку их работу не представляется возможным протестировать в силу необходимости установки специализированного программного обеспечения для их работы. Не рассматривалась и поисковая система WebCorp, поскольку она создавалась для «изучения языка в Интернете»¹⁸, и, соответственно, извлечение терминологии не входит в основной перечень задач, выполняемых данным ресурсом.

Таким образом, непосредственным материалом исследования для данной статьи послужили 4 терминологических экстрактора, находящихся в свободном доступе и разработанных для извлечения потенциальных специальных номинаций: AntConc, fivefilters.org, TerMine, Terminology Extraction.

Кроме того, анализ научной литературы, посвященной различным терминологическим экстракторам, позволил установить, что «corpus query tool» ‘инструмент корпусного запроса’ [24, с. 6] системы Sketch Engine¹⁹ предоставляет пользователям опцию извлечения потенциальных специальных номинаций, причем эффективность использования этого программного инструмента доказана проведенными ранее исследованиями [25, 26]. На основе технологии Sketch Engine, которая гарантирует быструю обработку корпуса, работает также «мощный онлайн-экстрактор терминов с возможностями извлечения одноязычных и двуязычных терминов OneClick Terms²⁰. Разработчики данного онлайн-экстрактора заявляют, что результатом его работы является «список терминов, практически не требующий ручной обработки. Извлеченные термины готовы для импорта в инструмент CAT (Computer Assisted Translation ‘Компьютерный перевод’

⁴ Terminology Coordination Unit [Electronic resource]. URL: <https://termcoord.eu/> (дата обращения: 20.03.2020).

⁵ Все названия терминологических экстракторов даны в орфографии, представленной на сайте TermCoord.

⁶ AlchemyAPI keyword [Electronic resource]. URL: <http://www.alchemyapi.com/api/keyword/> (дата обращения: 30.10.2021).

⁷ AntConc Homepage [Electronic resource]. URL: <https://www.laurenceanthony.net/software/antconc/> (дата обращения: 01.06.2019).

⁸ Apsic Xbench [Electronic resource]. URL: <https://www.xbench.net/?rd=y> (дата обращения: 30.10.2021).

⁹ Araya Bilingual Term Extraction Tool [Electronic resource]. URL: <http://www.heartsome.de/en/termextraction.php> (дата обращения: 30.10.2021).

¹⁰ fivefilters.org [Electronic resource]. URL: <https://www.fivefilters.org/term-extraction/> (дата обращения: 30.10.2021)

¹¹ KEA [Electronic resource]. URL: <http://community.nzdl.org/kea/> (дата обращения: 30.10.2021).

¹² maui-indexer [Electronic resource]. URL: <http://maui-indexer.appspot.com/> (дата обращения: 30.10.2021).

¹³ TerMine [Electronic resource]. URL: <http://www.nactem.ac.uk/software/termine/#form> (дата обращения: 30.10.2021).

¹⁴ Terminology Extraction [Electronic resource]. URL: <https://translatedlabs.com/terminology-extraction/> (request date: 30.10.2021).

¹⁵ topia.termextract [Electronic resource]. URL: <https://pypi.org/project/topia.termextract/1.1.0/#:~:text=topia.termextract%201.1.0n> (дата обращения: 30.10.2021).

¹⁶ WebCorp LSE [Electronic resource]. URL: <https://www.webcorp.org.uk/live/> (дата обращения: 30.10.2021).

¹⁷ Wordfast [Electronic resource]. URL: <http://www.wordfast.net/index.php?whichpage=downloadpage> (дата обращения: 30.10.2021).

¹⁸ WebCorp LSE. User Guide [Electronic resource]. URL: <http://wse1.webcorp.org.uk/home/guide.html> (дата обращения: 30.10.2021).

¹⁹ Sketch Engine [Electronic resource]. URL: <https://www.sketchengine.eu/> (дата обращения: 30.10.2021).

²⁰ OneClick Terms [Electronic resource]. URL: <https://terms.sketchengine.eu/#> (дата обращения: 30.10.2021).

или систему управления терминами»²¹. Это дает основания рассмотреть оба инструмента системы Sketch Engine. Таким образом, непосредственный материал исследования составили 5 терминологических экстракторов, находящихся в свободном доступе, и инструмент корпусного запроса Sketch Engine.

Анализ материала исследования осуществлялся в четыре этапа. На первом этапе были проанализированы характеристики терминологических экстракторов, заявленные разработчиками (информация представлена либо на главной странице официального сайта ресурса, либо содержится в руководствах пользователя, предлагаемых для скачивания и дальнейшего ознакомления). На втором этапе на основании установленных характеристик были выделены таксоны и определен набор значений в каждом из таксонов для дальнейшей систематизации информации о терминологических экстракторах. На третьем этапе в рамках каждого из таксонов были установлены терминологические экстракторы, соответствующие характеристики которых имеют наилучшие значения. На последнем, четвертом, этапе была проведена проверка на эффективность программных средств, имеющих лучшие показатели по совокупности рассмотрения таксонов. С этой целью результаты обработки фрагмента корпуса специальных текстов терминологическими экстракторами были сопоставлены со списком терминов, извлеченных из этого фрагмента корпуса вручную.

Корпус специальных текстов, фрагмент которого использовался для определения эффективности терминологических экстракторов, включает тексты 340 англоязычных официальных документов Европейского Союза, отражающих основные направления европейской языковой политики. Документы, составившие целевой корпус, представлены на официальных сайтах органов и институтов данного международного формирования и были отобраны автором данной статьи на основании отражения в них различных аспектов языковой политики ЕС. Общий объем сформированного корпуса составил 3498502 словоупотребления. Во фрагмент корпуса, на котором проводилась апробация терминологических экстракторов на эффективность, вошли терминологически насыщенные тексты, включающие в том числе неоднословные номинации понятий, имеющих отношение к языковой политике.

Известно, что, терминология является ядром языков для специальных целей, или ЯСЦ, а «специальные цели появляются тогда, когда в тех или иных группах индивидов или у отдельных ее членов возникает потребность в особом обозначении отдельных понятий, в выработке целостных понятийных систем, не известных общему языку» [27, с. 2]. Как показывает анализ научной литературы, посвященной терминологическим экстракторам, их эффективность рассматривается применительно к обработке текстов ЯСЦ определенной предметной области.

Специфика и новизна данного исследования заключается не только в таксономическом анализе, использованном для комплексного сопоставления терминологических экстракторов, но и в апробации возможностей лучших из них в обработке нормативных документов, в которых закономерно представлена разнообразная по понятийной отнесенности и функционалу терминология.

Таким образом, выбор эффективного терминологического экстрактора осуществлялся исходя из конкретной задачи исследования — извлечения максимального количества номинаций понятий языковой политики из текстов официальных документов Европейского Союза.

Результаты исследования и обсуждение

Анализ характеристик, заявленных создателями терминологических экстракторов, осуществлялся на основе использования таксономического метода, который, как известно, берет свое начало в биологии²². В современной науке таксономический метод применяется «для структуризации объектов, имеющих социальную природу (для систематизации и выявления внутренней

²¹ About OneClick Terms. User Guide [Electronic resource]. URL: <https://terms.sketchengine.eu/#about> (дата обращения: 30.10.2021).

²² Одной из первых таксономических классификаций принято считать классификацию, предложенную Карлом Линнеем (1707-1778), который установил определенную градацию для представителей живой природы: класс, отряд, род, вид, вариация.



структуры и иерархии взаимосвязей параметров)» [28, с. 60]. Под таксоном понимают совокупность объектов, «выделяемую по некоторому условию или комплексу признаков, которые признаются таксономически значимыми с точки зрения той или иной теории» [29, с. 19–20]. Выделение таксонов опирается на различные свойства и признаки объектов, набор которых должен быть достаточным для того, чтобы данный таксон занимал единственную нишу в системе и не пересекался с другими таксонами.

В данном исследовании выделение таксонов осуществлялось по критериям, заявленным разработчиками терминологических экстракторов. В представленной ниже таксономии терминологических экстракторов таксоны сконструированы по взаимодополняющим признакам, т.е. таксоны содержат «два или более признака без перекрытия значений [30, с. 91].

Таким образом, таксономию, приведенную ниже, можно отнести к политетической²³, поскольку в таксонах сгруппированы объекты, которые обладают наибольшим числом общих признаков, но при этом ни один из признаков не является необходимым и достаточным для включения объектов в таксономию [31, с. 91].

Исходя из характеристик, заявленных разработчиками терминологических экстракторов, было выделено 5 таксонов.

Первый таксон – степень доступности ресурса. Анализ характеристик терминологических экстракторов, заявленных разработчиками, показал, что данный таксон может включать следующие признаки:

1. Автономный доступ (необходима предварительная установка терминологического экстрактора на устройство пользователя для дальнейшей работы);
2. Онлайн-доступ;
3. Ограниченный доступ (разработчики данных терминологических экстракторов предлагают пользователю либо демонстрационную версию программного обеспечения с неполным набором функций, либо предоставляют доступ ко всем функциям только на определенный период, как правило, 30 дней, после чего требуется оформление платной подписки).

Распределение экстракторов по данным критериям представлено в табл. 1 Степень доступности ресурса (программные инструменты, продемонстрировавшие лучший результат в данном таксоне, выделены цветом).

Таблица 1. Степень доступности ресурса
Table 1. The degree of resource availability

Название экстрактора	Признак	Автономный доступ	Онлайн доступ	Ограниченный доступ
AntConc		+		
fivefilters.org			+	
OneClick Terms			+	+
Sketch Engine			+	+
TerMine		+	+	+
Terminology Extraction			+	+

Как следует из признаков, представленных в табл. 1, только терминологический экстрактор **AntConc**, разработанный Лоуренсом Энтони (Laurence Anthony), работает в автономном режиме. Данный экстрактор предоставляется с полным комплектом функций на безвозмездной основе.

²³ Политетическая классификация берет свое начало в биологической систематике, где установлено, что некоторые «несомненно естественные группы не подходят под такое понимание классификации» [31, с. 91], где выделение групп основано на положении о том, что каждый из диагностирующих признаков обязателен для любого члена группы.

Разработчики **fivefilters.org** предоставляют доступ ко всем функциям онлайн-ресурса на безвозмездной основе неограниченный период времени.

Доступ к онлайн-экстрактору **TerMine**, который также предлагается для установки на устройство пользователя, предоставлен на безвозмездной основе. Однако, разработчики данного программного инструмента, Национальный центр анализа текстов (The National Centre for Text Mining (NaCTeM)), заявляют, что поскольку **TerMine** – это свободно доступный ресурс из академической области, то «необходимо ограничивать нагрузку на сервер и отдавать предпочтение отдельным пользователям... Существует ограничение на то, сколько раз в день незарегистрированные пользователи могут воспользоваться этой услугой»²⁴.

Разработчики онлайн-экстрактора **Terminology Extraction**, компания Translated Labs, предоставляют демонстрационную версию работы, в которой данный ресурс выдает только «top19 terms» ‘19 наиболее частотных специальных номинаций’²⁵, что недостаточно для проверки данного терминологического экстрактора на эффективность. Для доступа ко всем функциям необходимо оформить платную подписку.

Разработчики инструмента корпусного запроса **Sketch Engine** Павел Рыхлый и Адам Килгарифф (Pavel Rychly and Adam Kilgariff) и Lexical Computing CZ s.r.o., разработчики онлайн-экстрактора **OneClick Terms**, предоставляют бесплатный доступ ко всем функциям данных программных инструментов только на период в 30 дней, после чего необходима платная подписка, которая может быть как индивидуальной, так и от научной организации.

Лучшим онлайн-экстракторами, исходя из признаков, представленных в таксоне, является **fivefilters.org**, который предоставляется на безвозмездной основе без ограничений по времени использования. Поскольку онлайн-экстрактор **OneClick Terms** и инструмент корпусного запроса **Sketch Engine**, как указано выше, предлагают доступ ко всем функциям даже на время бесплатного использования, их также можно рассматривать в качестве кандидатов при выборе необходимого программного обеспечения для выявления потенциальных специальных номинаций. Поскольку терминологический экстрактор **AntConc** является единственным из анализируемых ресурсов, который работает в автономном режиме, его также следует включить в список кандидатов при выборе необходимого программного обеспечения для выявления потенциальных специальных номинаций.

Второй таксон – удобство пользования ресурсом (*usability*). Под удобством пользования ресурсом понимают простоту и комфорт работы с ресурсом. Анализ характеристик терминологических экстракторов, заявленных разработчиками, показал, что данный таксон может включать следующие признаки:

1. Удобная навигация;
2. Наличие руководства пользователя;
2. Наличие диалоговых элементов.

Распределение ресурсов по данным критериям представлено в табл. 2 Удобство пользования ресурсом (программные инструменты, продемонстрировавшие лучший результат в данном таксоне, выделены цветом).

Разработчики всех анализируемых программных инструментов предоставляют удобную навигацию по ресурсам. Отсутствие признака ‘Наличие диалоговых элементов’ у терминологического экстрактора **AntConc** обусловлено тем, что данный экстрактор работает автономно и не предоставляет возможности перейти на сайт разработчика, на котором можно связаться с разработчиком через опцию ‘Contact’ ‘Контакт’, предоставляющую адрес электронной почты разработчика.

Разработчики онлайн-экстракторов **TerMine** и **Terminology Extraction** предоставляют не руководство пользователя, а демонстрационную версию работы с данными ресурсами, в то время как

²⁴ TerMine [Electronic resource]. URL: <http://www.nactem.ac.uk/software/termine/#form> (дата обращения: 30.10.2021).

²⁵ Terminology Extraction [Electronic resource]. URL: <https://translatedlabs.com/terminology-extraction> (request date: 30.10.2021).

Таблица 2. Удобство пользования ресурсом
Table 2. Ease of use of the resource

Название экстрактора	Признак	Удобная навигация	Наличие руководства пользователя	Наличие диалоговых элементов
AntConc		+	+	
fivefilters.org		+		+
OneClick Terms		+	+	
Sketch Engine		+	+	+
TerMine		+		+
Terminology Extraction		+		+

на официальных страницах программных инструментов **fivefilters.org** и **OneClick Terms** пользователь может сразу приступить к работе с ресурсом, используя всплывающие подсказки.

Разработчики инструмента корпусного запроса **Sketch Engine** предоставляет не только удобную навигацию по ресурсу с подробным объяснением использования каждой опции, но также предоставляют руководство пользователя и видео-инструкцию работы на главной странице ресурса. Кроме того, пользователям предоставлена возможность связаться с разработчиками данного программного инструмента через опцию ‘Contact’. Отсутствие диалоговых окон на официальной странице онлайн-экстрактора **OneClick Terms** можно объяснить тем, что данный ресурс создан на основе технологии Sketch Engine, соответственно всю необходимую информацию пользователь может получить на официальной странице инструмента корпусного запроса **Sketch Engine**.

Таким образом, инструмент корпусного запроса **Sketch Engine** демонстрирует лучший результат в данном таксоне и может быть рассмотрен как потенциальный кандидат при выборе программного обеспечения для извлечения специальных номинаций.

Третий таксон – способность экстракторов выявлять полилексы, или многокомпонентные специальные номинации. Анализ характеристик терминологических экстракторов, заявленных разработчиками, показал, что данный таксон может включать следующие признаки:

1. Выдача только универбов;
2. Выдача только полилексов;
3. Выдача универбов и полилексов.

Распределение ресурсов по данным критериям представлено в табл. 3 Виды выявляемых специальных номинаций (программные инструменты, продемонстрировавшие лучший результат в данном таксоне, выделены цветом).

Согласно мнению авторитетных ученых-терминологов, «многословные термины в большинстве европейских языков составляют 60 – 80% от общего количества терминов» [32, с. 121]. Таким образом, наличие опции «Способность экстракторов выявлять полилексы» – это необходимое условие при выборе программного инструмента.

Как следует из значений, представленных в табл. 3, возможности получения полилексов не предоставляет лишь терминологический экстрактор **AntConc**, а инструмент **TerMine** не ориентирован на поиск терминов-универбов.

Выдача полилексов экстрактором **fivefilters.org** зависит от запроса пользователя: предоставляется возможность задавать количество элементов в потенциальных специальных номинациях от 1 до 10. Следует отметить, что, если для терминологического экстрактора **fivefilters.org** задана опция выдачи потенциальных специальных номинаций с одним элементом, результатом выдачи будут только универбы, а если задана опция выдачи потенциальных специальных номинаций

Таблица 3. Виды выявляемых специальных номинаций
Table 3. Types of identified special nominations

Название экстрактора	Признак	Выдача только универбов	Выдача только полилексов	Выдача универбов и полилексов
AntConc		+		
fivefilters.org				+
OneClick Terms				+
Sketch Engine				+
TerMine			+	
Terminology Extraction				+

с количеством элементов более одного, то экстрактор выдает и универбы, и полилексы в виде онлайн-таблицы, где содержатся потенциальные специальные номинации и информация об их частоте употребления в корпусе. Пользователь может задать формат выдачи результатов данного экстрактора (.html, .json, .xml, .text, .php) для сохранения и дальнейшей обработки. Существенный недостаток работы данного экстрактора – выдача только ограниченного количества потенциальных специальных номинаций (максимальное количество – 100 единиц), что недостаточно для проверки данного терминологического экстрактора на эффективность.

Терминологический экстрактор **OneClick Terms** и инструмент корпусного запроса **Sketch Engine** выдают и универбы, и полилексы. Результаты выдачи обоих программных инструментов можно сохранить для дальнейшей обработки в форматах .csv, .xlsx. Инструмент корпусного запроса **Sketch Engine** позволяет также сохранить результаты выдачи для дальнейшей обработки в формате .xml.

В демонстрационной версии терминологического экстрактора **Terminology Extraction**, представленной на официальном сайте ресурса, можно получить только «top19 terms» ‘19 наиболее частотных специальных номинаций’, что, как упоминалось выше, не позволит проверить на эффективность данный терминологический экстрактор. В списке полученных специальных номинаций представлены и универбы, и полилексы, ранжированные по количеству употреблений. Результаты выдачи предоставляются в виде онлайн-таблицы, в которой потенциальные специальные номинации отсортированы по частоте их употребления в корпусе.

Таким образом, анализ значений, представленных в данном таксоне, позволил установить, что терминологические экстракторы **OneClick Terms** и инструмент корпусного запроса **Sketch Engine**, которые позволяют получить максимально возможное количество как универбов, так и полилексов и предоставляют возможность сохранить результаты выдачи для дальнейшей обработки, могут считаться лучшими в данном таксоне.

Четвертый таксон – наличие опции обращения к контекстам терминопотребления. Анализ характеристик терминологических экстракторов, заявленных разработчиками, показал, что данный таксон может включать следующие признаки:

1. Наличие опции «Предоставление контекстов»;
2. Отсутствие опции «Предоставление контекстов».

Распределение ресурсов по данным критериям представлено в табл. 4 Предоставление контекстов терминологическими экстракторами (программные инструменты, продемонстрировавшие лучший результат в данном таксоне, выделены цветом)

Наличие опции «Предоставление контекстов» представляется важной для современного терминологического менеджмента (*term management*), одним из этапов которого является оптимиза-

Таблица 4. Предоставление контекстов терминологическими экстракторами
Table 4. Provision of contexts by terminological extractors

Название экстрактора \ Значение	Наличие опции «Предоставление контекстов»	Отсутствие опции «Предоставление контекстов»
AntConc	+	
fivefilters.org		+
OneClick Terms	+	
Sketch Engine	+	
TerMine	+	
Terminology Extraction	+	

ция инвентаризированной терминологии. На этапе оптимизации устанавливаются особенности функционирования и использования терминов, выявляются терминологическая омонимия, синонимия, антонимия и устанавливаются разные значения многозначных терминов, что достигается анализом контекста употребления терминов. Кроме того, выделение контекстов употребления терминов необходимо при составлении объяснительных глоссариев, которые «предоставляют информацию о концептуальной структуре специальной предметной области» [33, с. 119].

Таким образом, отсутствие опции «Предоставление контекстов» у терминологического экстрактора **fivefilters.org** понижает его позицию в данном таксоне, в то время как терминологические экстракторы **AntConc**, **OneClick Terms**, **TerMine**, **Terminology Extraction** и инструмент корпусного запроса **Sketch Engine** занимают равные позиции в данном таксоне.

Следует отметить, что наличие опции «Предоставление контекстов» у оставшихся кандидатов имеет свои особенности. Представляется удачной реализация данной опции в терминологическом экстракторе **AntConc**: пользователь самостоятельно может выбрать размер контекста справа и / или слева от выделенного универба.

Терминологический экстрактор **OneClick Terms** и инструмент корпусного запроса **Sketch Engine** предоставляют контексты употребления специальных номинаций либо в виде законченных смысловых фрагментов (**OneClick Terms**), либо в виде лево- и правосторонних контекстов (**Sketch Engine**).

Несмотря на то, что у терминологического экстрактора **TerMine** отсутствует отдельная опция «Предоставление контекстов», полилексы выделяются цветом в корпусе, загруженном для обработки данным экстрактором, при этом отсутствует возможность поиска необходимой специальной номинации, что можно отнести к недостаткам работы данного экстрактора.

Специальные номинации выделяются цветом и в тексте подготовленного корпуса, обработанном терминологическим экстрактором **Terminology Extraction**, и в итоговой сводной таблице, при этом цвет выделенной специальной номинации в таблице соответствует цвету, которым выделена специальная номинация в тексте подготовленного корпуса, что облегчает поиск специальной номинации в контексте.

Таким образом, терминологические экстракторы **AntConc**, **OneClick Terms**, **Terminology Extraction** и инструмент корпусного запроса **Sketch Engine** демонстрируют лучшие результаты в данном таксоне.

Пятый таксон – количество, объем и форматы обрабатываемых файлов. Анализ характеристик терминологических экстракторов, заявленных разработчиками, показал, что данный таксон может включать следующие признаки:

1. Обработка одного файла;
2. Обработка двух файлов с сопоставимыми корпусами;

3. Ограничение объема обрабатываемых файлов;
4. Отсутствие ограничений объема обрабатываемых файлов;
5. Ограничение форматов обрабатываемых файлов;
6. Отсутствие ограничений форматов обрабатываемых файлов.

Распределение ресурсов по данным критериям представлено в табл. 5 Количество, объем и форматы обрабатываемых файлов, табл. 6 Объем обрабатываемых файлов и табл. 7 Форматы обрабатываемых файлов (программные инструменты, продемонстрировавшие лучший результат в данном таксоне, выделены цветом).

Таблица 5. Количество обрабатываемых файлов
Table 5. Number of processed files

Название экстрактора	Значение	Обработка одного файла	Обработка двух файлов с сопоставимыми корпусами
AntConc		+	
fivefilters.org		+	
OneClick Terms		+	+
Sketch Engine		+	+
TerMine		+	
Terminology Extraction		+	

Таблица 6. Объем обрабатываемых файлов
Table 6. Volume of processed files

Название экстрактора	Значение	Ограничение объема обрабатываемых файлов	Отсутствие ограничений объема обрабатываемых файлов
AntConc			+
fivefilters.org			+
OneClick Terms			+
Sketch Engine		+	
TerMine		+	
Terminology Extraction			+

Анализ признаков, представленных в табл. 5, позволил выявить 2 группы терминологических экстракторов в данном таксоне: к первой группе принадлежат экстракторы, которые обрабатывают корпус на одном языке (**AntConc**, **fivefilters.org**, **TerMine**, **Terminology Extraction**), ко второй относятся универсальные программные инструменты, которые могут работать как с корпусами на одном языке, так и с сопоставимыми корпусами (**OneClick Terms** и инструмент корпусного запроса **Sketch Engine**). Очевидно, что программные инструменты, относящиеся ко второй группе, демонстрируют лучшее значение в данном таксоне.

Разработчики терминологических экстракторов **AntConc**, **fivefilters.org** и **Terminology Extraction** не выставляют ограничений на объем файла(-ов). Для работы с терминологическим экстрактором **AntConc** необходимо сохранить подготовленный корпус в файле с расширением .txt.

Экстрактор **Terminology Extraction** обрабатывает скопированный из подготовленного корпуса текст, который необходимо вставить в поле, предназначенное для данной цели, соответственно, ограничений по объему обрабатываемых файлов нет.

Таблица 7. Форматы обрабатываемых файлов
Table 7. Formats of processed file

Название экстрактора	Значение	Ограничение формата обрабатываемых файлов	Отсутствие ограничений формата обрабатываемых файлов
AntConc		+	
fivefilters.org			+
OneClick Terms		+	
Sketch Engine		+	
TerMine		+	
Terminology Extraction			+

По подобному принципу работает и терминологический экстрактор **fivefilters.org**, который, помимо обработки скопированного текста из специально подготовленного корпуса, может обработать и онлайн-тексты, для чего необходимо вставить URL-ссылку в соответствующее поле интерфейса данного экстрактора.

Терминологический экстрактор **TerMine** также обрабатывает тексты, которые представлены в онлайн-режиме, но разработчики отмечают, что содержание страницы должно быть представлено либо в формате с расширением `.html`, либо в формате с расширением `.pdf`, а объем обрабатываемой информации не должен превышать 2 МБ. Особые требования предъявляются также и к файлам, содержащим подготовленный корпус (файлы должны быть представлены в формате `.txt` в кодировке ASCII или в формате `.pdf`, размер файла не должен превышать 2 МБ), и к тексту подготовленного корпуса, который можно вставить в соответствующее поле терминологического экстрактора (допускаются только символы ASCII).

Инструмент корпусного запроса **Sketch Engine** обрабатывает подготовленные корпуса в форматах с расширением `.csv`, `.doc`, `.docx`, `.htm`, `.html`, `.ods`, `.pdf`, `.tar.bz2`, `.tar.gz`, `.tei`, `.tgz`, `.tmx` (для сопоставимых корпусов), `.txt`, `.vert`, `.xlf`, `.xliff`, `.xml`, `.zip.tmx`, а также онлайн-тексты, которые предоставляются по URL-ссылке. Онлайн-экстрактор **OneClick Terms**, как упоминалось выше, создан на основе технологии **Sketch Engine**, однако может обработать подготовленные корпуса только в форматах с расширением `.doc`, `.docx`, `.htm`, `.html`, `.pdf`, `.txt` (для моноязычных корпусов) и `tmx`, `.xlf2.0+` `.xliff2.0+` (для сопоставимых корпусов). Терминологический экстрактор **OneClick Terms** не ограничивает объем обрабатываемых файлов, в отличие от инструмента корпусного запроса **Sketch Engine**, который позволяет обработать подготовленный текстовый корпус объемом не более 100 файлов, при этом максимальный размер файла не должен превышать 500 МБ.

Следовательно, лучший результат в пятом таксоне, несмотря на ограничение объема обрабатываемых файлов, демонстрирует инструмент корпусного запроса **Sketch Engine**, который обрабатывает как моноязычные, так и сопоставимые корпуса в наиболее популярных пользовательских форматах. Терминологические экстракторы **fivefilters.org** и **Terminology Extraction**, которые не выставляют ограничений на объем файла(-ов) и позволяют обработать корпус независимо от формата, в котором он сохранен, занимают равные позиции в данном таксоне. Терминологический экстрактор **OneClick Terms** также может быть рассмотрен как потенциальный кандидат при выборе программного обеспечения для извлечения специальных номинаций, поскольку обрабатывает как моноязычные, так и сопоставимые корпуса, не выставляя ограничений на объем файла(-ов).

Таким образом, анализ характеристик терминологических экстракторов, заявленных разработчиками, с применением таксономического метода позволил выявить лучшие ресурсы в каждом таксоне, исходя из заявленных их создателями характеристик, степени доступности, простоты работы. В табл. 8 Сравнительный анализ терминологических экстракторов отмечены

Таблица 8. Сравнительный анализ терминологических экстракторов
Table 8. Comparative analysis of terminological extractors

Название экстрактора \ № таксона	Таксон 1	Таксон 2	Таксон 3	Таксон 4	Таксон 5
AntConc	+			+	
fivefilters.org	+				+
OneClick Terms	+		+	+	+
Sketch Engine	+	+	+	+	+
TerMine				+	
Terminology Extraction				+	+

программные инструменты, продемонстрировавшие лучшие результаты в каждом из таксонов. Исходя из сравнительного анализа таксонов лучший результат показали инструмент корпусного запроса **Sketch Engine** и онлайн-экстрактор **OneClick Terms** (выделены цветом), которые, следовательно, были отобраны для проверки на эффективность.

Поскольку эффективность – это «комплексная характеристика системы, отражающая степень ее соответствия потребностям и интересам ее заказчиков, пользователей, других заинтересованных лиц» [34], то эффективность выбранных программных инструментов оценивалась в зависимости от того, насколько полученные результаты соответствуют опциям, заявленным разработчиками.

Выбранное программное обеспечение было протестировано на фрагменте корпуса текстов официальных документов из сферы языковой политики ЕС, объемом 1207 словоупотреблений, который прошел предобработку путем удаления фрагментов терминофиксации, чисел и ссылок на интернет-источники и последующего удаления в автоматическом режиме разрывов строк. Для проверки инструмента корпусного запроса **Sketch Engine** и онлайн-экстрактора **OneClick Terms** на эффективность результаты обработки ими фрагмента корпуса специальных текстов были сопоставлены со списком терминов, извлеченных из этого фрагмента корпуса вручную.

Для того, чтобы определить кандидаты в термины при ручной обработке текста, были использованы следующие критерии терминологичности:

- ◆ логический, согласно которому, по мнению В.М. Лейчика, терминологичной будет считаться такая специальная номинация, которая «обозначает понятие в системе понятий по его отличительному признаку (признакам)» [35, с. 71];

- ◆ семантический, согласно которому терминологичностью обладает такая специальная номинация, которая называет класс объектов при условии, что в «терминосистеме нет других единиц, частично совпадающих с ней по значению» [35, с. 71];

- ◆ критерий частотности, согласно которому специальную номинацию, встречающуюся более 2 раз, можно рассматривать как потенциальный термин;

- ◆ критерий использования регулярных терминоэлементов. Под терминоэлементом понимают «широкое понятие, включающее в себя на равных основаниях производящую основу, словообразующую морфему, слово в составе терминологического словосочетания, символы, цифры, графические знаки, включаемые в особый тип символа-слов» [36, с. 79]. Согласно данному критерию, терминологичностью будет обладать такая специальная номинация, в состав которой входит терминоэлемент, встречающийся более 2 раз в выделенных специальных номинациях;

- ◆ формальный, согласно которому многокомпонентные специальные номинации должны быть представлены моделями, для которых характерны виды связи элементов, соответствующие



ющие грамматическим особенностям языка целевой области, «что может быть использовано, например, для создания алгоритма автоматического распознавания составных терминов в тексте» [32, с. 136].

Кроме того, отбор полилексов в ручном режиме основывался на определении степени их синтагматической целостности (*unithood*).

Понятие ‘синтагматическая целостность’ было введено в 1996 К. Кагеура и У. Умино для обозначения степени устойчивости синтагматических сочетаний и словосочетаний [37, с. 272]. Таким образом, синтагматическая целостность определяется исключительно на уровне синтаксиса, соответственно данный критерий важен при отборе полилексов, поскольку, если наблюдается тенденция употребления компонентов специальных единиц в одной и той же позиции, данные единицы могут считаться кандидатами в многокомпонентные специальные номинации различной степени терминологичности.

В результате ручной обработки фрагмента корпуса было получено 22 термина, номинирующих понятия языковой политики в текстах официальных документов Европейского Союза: 2 универба (*language(s)*, *multilingualism*) и 20 полилексов, представленных 14 двухкомпонентными номинациями (*communications technologies*, *digital content*, *human resources*, *impact of multilinguality*, *language industries*, *language issues*, *language resources*, *language services*, *language technologies*, *language tools*, *linguistic customization*, *linguistic diversity*, *multilingual services*, *promotion of multilingualism*), 4 трехкомпонентными номинациями (*human language technologies*, *language industry players*, *language policy framework*, *multilingual digital content*) и 2 четырехкомпонентными номинациями (*adoption of multilanguage product development*, *language-related EU projects*). Для уточнения терминологичности выделенных вручную номинаций были использованы глоссарии²⁶, сопровождающие ряд документов.

Термины *human language technologies*, *language industries*, *language resources*, *language services*, *language technologies*, *linguistic diversity*, *multilingualism*, *multilingual services* зафиксированы в глоссариях. Основываясь на понятии синтагматической целостности и критериях терминологичности, представленных выше, номинации *language industry players* и *language policy framework* можно рассматривать как производные термины от представленных в глоссариях терминов *language industry* и *language policy*.

Номинация *language issues* входит в дефиницию термина *language awareness*²⁷, что позволяет отнести ее к терминам, несмотря на то, что она обладает меньшей степенью терминологичности по отношению к дефинируемому термину: «любое слово или словосочетание обладает большей терминологичностью, чем все участвующие в его дефиниции слова и словосочетания» [38, с. 22].

Применив исчисленные выше критерии терминологичности и опираясь на знание о том, что для полилексов «наряду с устойчивостью (цельностью номинации), обусловленной их функцией наименования одного понятия, указывается их номинативный характер и атрибутивный (определятельный) вид связи составляющих их элементов» [32, с. 136], к специальным номинациям различной степени терминологичности можно отнести такие полилексы, как *communications technologies*, *digital content*, *human resources*, *language tools*, *linguistic customization*, *promotion of multilingualism*, *multilingual digital content*.

В результате обработки фрагмента корпуса инструментом корпусного запроса **Sketch Engine** было получено 448 однословных потенциальных терминов и 257 полилексов, которые, соответственно, представлены во вкладках: «SINGLE WORDS» ‘Отдельные слова’ и «MULTI-WORD TERMS» ‘Термины, состоящие из нескольких слов’. В результате обработки фрагмента корпуса онлайн-экстрактором **OneClick Terms** было получено 104 однословных специальных номинации во вкладке «SINGLE WORDS» ‘Отдельные слова’, что почти в 4,3 раза меньше результатов выдачи

²⁶ В глоссариях, как известно, содержатся термины, сопровождающиеся определениями, а «наличие определения значения той или иной языковой единицы является условием достаточным для того, чтобы признать ее термином, а точнее – обладающей некоторой, ненулевой степенью терминологичности» [38, с. 20].

²⁷ «the extent to which language issues are embedded into the strategies and policies of the company». [39]

инструмента корпусного запроса **Sketch Engine**, и 104 полилекса во вкладке «MULTI-WORDS» ‘Многокомпонентные слова’, что почти в 2,5 раза меньше результатов выдачи опцией извлечения потенциальных специальных номинаций инструментом корпусного запроса **Sketch Engine**.

Существенная разница в выдаче потенциальных номинаций может свидетельствовать о том, что результаты выдачи онлайн-экстрактора **OneClick Terms** потенциально точнее результатов выдачи инструмента корпусного запроса **Sketch Engine** и подтверждает заявление разработчиков данного онлайн-экстрактора о том, что результатом его работы является «список терминов, практически не требующий ручной обработки»²⁸.

Очевидно, что официальные документы содержат большое количество терминов, отражающих их деловой и / или правовой характер, но поскольку для нашей исследовательской задачи важен отбор терминов, имеющих отношение к языковой политике, то результаты выдач экстракторов были сопоставлены с перечнем номинаций, выделенных в ручном режиме, обладающими различной степенью терминологичности (см. табл. 9 Результаты выдачи терминологических экстракторов).

Таблица 9. Результаты выдачи терминологических экстракторов
Table 9. Results of issuing terminological extractors

Кандидаты в термины (ручной режим)	Специальные номинации, выявленные Sketch Engine	Специальные номинации, выявленные OneClick Terms
<i>adoption of multilanguage product development</i>	+	+
<i>communications technologies</i>	<i>communications technology</i>	
<i>digital content</i>	+	+
<i>human language technologies</i>	<i>human language technology</i>	
<i>impact of multilinguality</i>		
<i>human resources</i>	<i>human resource</i>	
<i>language</i>	+	+
<i>language industries</i>	<i>language industry</i>	<i>language industry</i>
<i>language industry players</i>	<i>language industry player</i>	<i>language industry player</i>
<i>language issues</i>	<i>language issue</i>	
<i>language policy framework</i>	+	
<i>language resources</i>	<i>language resource</i>	
<i>language services</i>	<i>language service</i>	<i>language service</i>
<i>language technologies</i>	<i>language technology</i>	<i>language technology</i>
<i>language tools</i>	<i>language tool</i>	
<i>language-related EU projects</i>		
<i>linguistic customisation</i>	+	+
<i>linguistic diversity</i>	+	
<i>multilingualism</i>	+	+
<i>multilingual digital content</i>		
<i>multilingual services</i>	<i>multilingual service</i>	
<i>promotion of multilingualism</i>		

Термины *language* и *multilingualism* представлены в выдачах обоих экстракторов. Наличие терминов *language* и *multilingualism* в глоссариях, сопровождающих официальные документы Европейского Союза, и в результатах выдач обоих программных инструментов позволяет считать термины *language* и *multilingualism* «терминами-доминантами» [40, с. 126], номинирующими

основные понятия европейской языковой политики. Это подтверждается также частотностью этих единиц как терминоэлементов в терминах-полилексах.

Как уже отмечалось выше, инструментом корпусного запроса **Sketch Engine** было выдано 257 многокомпонентных потенциальных специальных номинаций, из них 16 соответствуют 20 терминам, выделенным вручную, что составляет 80%. Онлайн-экстрактором **OneClick Terms** выдано 104 многокомпонентные номинации, из них лишь 7 из 20 (35%) соответствуют выделенным вручную полилексам²⁹.

Для сопоставления результативности экстракторов относительно нашей исследовательской задачи были использованы такие показатели эффективности информационного поиска, как полнота и точность.

Для определения полноты поиска необходимо найти отношение выделенных в автоматическом режиме целевых специальных номинаций (AP) к целевым специальным номинациям, полученным в результате обработки фрагмента корпуса в ручном режиме (PP).

Таблица 10. Полнота поиска специальных номинаций
Table 10. Completeness of the search for special nominations

Получено специальных номинаций в результате обработки программным инструментом		Получено специальных номинаций в ручном режиме	Полнота поиска (AP / PP *100), %
Sketch Engine	18	22	81
OneClick Terms	9		41

Как следует из полученных результатов, показатель полноты для инструмента корпусного запроса **Sketch Engine** почти в 2 раза выше, чем у экстрактора **OneClick Terms**.

Для определения точности поиска необходимо найти отношение выделенных в автоматическом режиме целевых специальных номинаций (ЦН) к общему количеству специальных номинаций (ОК), полученных также в автоматическом режиме.

Таблица 11. Точность поиска специальных номинаций
Table 11. The accuracy of the search for special nominations

Программный инструмент	Получено всего специальных номинаций	Получено всего целевых специальных номинаций	Точность поиска (ЦН / ОК *100), %
Sketch Engine	705	18	2,6
OneClick Terms	208	9	4,3

Как следует из полученных результатов, точность поиска незначительно выше у онлайн-экстрактора **OneClick Terms** (4,3%). Поскольку разница в точности поиска не является статистически значимой, а полнота поиска значительно выше у инструмента корпусного запроса **Sketch Engine**, то для решения поставленной задачи, инвентаризации целевой терминологии, наиболее эффективным следует признать инструмент корпусного запроса **Sketch Engine**.

²⁹ Поскольку терминологические экстракторы выдают специальные номинации в форме единственного числа, то полилексы, полученные в результате обработки фрагмента корпуса инструментом корпусного запроса **Sketch Engine** и онлайн-экстрактором **OneClick Terms**, представлены в табл. 9 в форме единственного числа за исключением тех случаев, когда полилекс содержит неисчисляемое существительное

Заключение

В научных публикациях, посвященных определению эффективности терминологических экстракторов, как правило, рассматриваются результаты их работы применительно к специальным текстам определенной предметной области либо определяются преимущества конкретных экстракторов с точки зрения доступных опций обработки языкового материала. В данной работе предложен комплексный подход к определению результативности экстракторов в решении конкретной исследовательской задачи, основанный на применении таксономического метода для сопоставления экстракторов в разрезе критериев, заявленных разработчиками каждого из инструментов, на сопоставлении результатов выдач экстракторов со списком терминов, извлеченных вручную, и на исчислении показателей полноты и точности, традиционно используемых для оценки эффективности результатов информационного поиска. Реализация такого подхода при определении терминологического экстрактора, оптимального для решения исследовательской задачи, связанной с инвентаризацией номинаций понятий языковой политики, представленных в текстах официальных документов Европейского Союза, позволила обосновать наибольшую эффективность для этой цели инструмента корпусного запроса Sketch Engine.

СПИСОК ИСТОЧНИКОВ

1. **Ananiadou S., Nenadic G., Mima H., Tsujii J.** Mining Biomedical Terminology from literature // *Terminology, Computing and Translation*. / Hacken, P. (ed.) Tübingen, Gunter Narr Verlag, 2006. Pp. 117–140.
2. **Vázquez M., Oliver A.** Improving Term Candidates Selection Using Terminological Tokens // *Terminology*. 2018. No. 24(1). Pp. 122–146. DOI: 10.1075/term.00016.vaz
3. **Bowker L.** Off the Record and On the Fly: Examining the Impact of Corpora on Terminographic Practice in the Context of Translation // *Corpus-Based Translation Studies: Research and Applications* / Kruger, A. [et al.] (ed) New York, Bloomsbury Publishing, 2011. Pp. 115–127.
4. **Ahmad K., Rogers M.** Corpus Linguistics and Terminology Extraction // *Handbook of Terminology Management: Vol. 2: Application-Oriented Terminology Management* / Wright S.E., Budin G. (ed.) Amsterdam, John Benjamins Publishing Company, 2001. Pp. 725–760.
5. **Zaretskaya A., Pastor G.C., Seghiri M.** Translators' Requirements for Translation Technologies: a user survey // *Proceedings of the AIETI7 International Conference «New Horizons in Translation and Interpreting Studies» (Full papers)*. 2015. Pp. 247–254.
6. **Станкевич А.Ю.** Поиск контекстов и оценка их типичности средствами AntConc (Laurence Anthony) // *Материалы V Междунар. науч.-метод. конф. «Теория и практика преподавания русского языка как иностранного: достижения, проблемы и перспективы развития»*. Минск, 2011. С. 210–213.
7. **Anthony L.** Issues in the Design and the Development of Software Tools for Corpus Studies: The Case for Collaboration // *Contemporary Corpus Linguistics* / Baker, P. (ed) London, A&C Black, 2012. Pp. 87–105.
8. **Weisser M.** *Practical Corpus Linguistics: An Introduction to Corpus-Based Language Analysis*. Hoboken: John Wiley & Sons, 2016. 312 p.
9. **Котюрова И.А.** Корпусные исследования с помощью сервиса Antconc в условиях работы в вузе // *Научный журнал «Язык и культура»*. 2020. № 52. С. 36–50. DOI: 10.17223/19996195/52/3
10. **Frantzi K., Ananiadou S., Mima H.** Automatic Recognition of Multi-Word Terms // *International Journal of Digital Libraries*. 2000. No. 3(2). Pp. 117–132.
11. **Ludeling A., Evert S., Baroni M.** Using Web Data for Linguistic Purposes // *Corpus Linguistics and the Web* / Hundt M., Nesselhauf N., Biewer C. (ed.) Amsterdam, Rodopi, 2007. Pp. 7–25.
12. **Renouf A., Kehoe A., Banerjee J.** Weborp: an Integrated system for Web text search // *Corpus Linguistics and the Web* / Hundt M., Nesselhauf N., Biewer C. (ed.) Amsterdam, Rodopi, 2007. Pp. 47–69.

13. **Huang Y.-F., Ciou C.-S.** Constructing Personal Knowledge Base: Automatic Key Phrase Extraction from Multi-Domain Web Page // Proceedings of International Workshops «New Frontiers in Applied Data Mining: PAKDD 2011» (Revised Selected Papers). 2012. Pp. 65–76.
14. **Lybbert T.J., Zolas N.J.** Getting Patents and Economic Data to Speak to Each Other: An “Algorithmic Links with Probabilities” Approach for Joint Analyses of Patenting and Economic Activity. World Intellectual Property Organization, 2012. 31 p.
15. **Yanliang Qi et al.** Combining Supervised Learning Techniques to Key-Phrase Extraction for Bio-medical Full-Text // Organizational Efficiency through Intelligent Information Technologies / Sugumaran, V. (ed.) Hershey, Information. Science. Reference, 2013. Pp. 33–45.
16. **Захаров В.П., Хохлова М.В.** Автоматическое выявление терминологических словосочетаний // Структурная и прикладная лингвистика. 2014. Вып. 10. С. 182–200.
17. **Kehoe A.** Diachronic Linguistic Analysis on the Web with WebCorp // The Changing Face of Corpus Linguistics / Renouf, A., Kehoe, A. (ed.) Amsterdam, Rodopi, 2016. Pp. 297–309.
18. **Kosa V. et al.** Cross-Evaluation of Automated Term Extraction Tools by Measuring Terminological Saturation // Proceedings of 13th International Conference «Information and Communication Technologies in Education, Research, and Industrial Applications» (Revised Selected Papers). 2017. Pp. 135–163.
19. **Савельев С.В.** Место средств автоматизированного извлечения терминологии в работе терминолога переводческой компании // Материалы Междунар. науч.-практ. конф. «Актуальные вопросы лингвистики и лингводидактики: традиции и инновации». М., 2018. С. 66–71.
20. **Pavelec M.** Translation Quality: Concepts, Procedures and Tools // Teaching Translation and Interpreting: Advances and Perspectives / Bogucki, Ł., Deckert, M. (ed.) Newcastle, Cambridge Scholars Publishing, 2012. Pp. 137–146.
21. **Беляева Л.Н.** Сетевые ресурсы в технологии перевода // Вестник СПбГУ. Серия 9. Филология. Востоковедение. Журналистика. 2016. Вып. 4. С. 45–55. DOI: 10.21638/11701
22. **Sin-Wai Ch.** The Future of Translation Technology: Towards a World without Babel. Abingdon-on-Thames: Routledge. 2016. 316 p.
23. **Лейчик В.М.** Прикладное терминоведение и его направления // Прикладное языкознание / под ред. А. С. Герда. СПб., 1996. С. 276–286.
24. **Thomas J.** Discovering English with Sketch Engine. Brno : Versatile, 2016. 228 p.
25. **Ковязина М.А.** Извлечение ключевых терминов на базе корпуса текстов о разработке нефтяных и газовых месторождений // Вестник Тюменского государственного университета. Гуманитарные исследования. Humanitates. 2016. Том 2. № 3. С. 61–69. DOI: 10.21684/2411-197X-2016-2-3-61-69
26. **Новикова А.А.** Сравнение инструментов Sketch Engine и TermoStat для извлечения терминологии // International Journal of Open Information Technologies (INJOIT). 2020. Vol. 8. No. 11. С. 73–79.
27. **Герд А.С.** Введение в изучение языков для специальных целей. Санкт-Петербург : СПбГУ, РИО, Филологический факультет, 2011. 60 с.
28. **Любутов А.С.** Метод структурной таксономии: возможности применения для анализа социальных и духовных процессов // Научный результат. Социология и управление. 2019. Т. 5. № 4. С. 58–79. DOI: 10.18413/2408-9338-2019-5-4-0-6
29. **Шаталкин А.** Таксономия. Основания, принципы и правила. М.: Товарищество научных изданий КМК, 2012. 600 с.
30. **Неизвестный С.И.** О применении таксономии в области информационных технологий // Транспортные системы и технологии. 2016. Т. 2. № 1. С. 89–111.
31. **Шайкевич А., Андриющенко В., Ребецкая Н.** Дистрибутивно-статистический анализ языка русской прозы 1850–1870-х гг., Том 1. М: Языки славянской культуры, 2014. 504 с.
32. **Гринев-Гриневич С.В.** Терминоведение. М.: Издательский центр «Академия», 2008. 304 с.
33. **Sabré M.T.** Terminology Theory, Methods and Applications. Amsterdam: John Benjamins Publishing Company, 1998. 248 p.
34. **Зиндер Е.З.** Что такое «эффективность ИТ» // Intelligent Enterprise/IE («Корпоративные системы») : ИТ-журнал. 2006. № 8 (141). URL: <https://www.iemag.ru/master-class/detail.php?ID=1572> (дата обращения: 01.11.2021).
35. **Лейчик В.М.** Реализация комплексного показателя терминологичности в атрибутивных конструкциях // Деривация в норме и терминосистемах / под. ред. Б.И. Барткова. Владивосток: ДВО АН СССР, 1990. С. 64–74.

36. **Даниленко В.П.** О терминологическом словообразовании // Вопросы языкознания. 1973. № 4. С. 76–85.
37. **Kageura K., Umino B.** Methods of Automatic Term Recognition: a Review // Terminology. 1996. Vol. 3. P. 259–289.
38. **Шелов С.Д.** Термин. Терминологичность. Терминологические определения. СПб.: Филологический факультет СПбГУ, 2003. 280 с.
39. ELAN: Effects on the European Economy of Shortages of Foreign Language Skills in Enterprise, Directorate General for Education and Culture (European Commission) // URL: https://ec.europa.eu/assets/eac/languages/policy/strategic-framework/documents/elan_en.pdf (дата обращения: 20.04.2020).
40. **Лейчик В.М.** Терминоведение: предмет, методы, структура. М.: Издательство ЛКИ, 2007. 256 с.

REFERENCES

- [1] **S. Ananiadou, G. Nenadic, H. Mima, J. Tsujii**, Mining Biomedical Terminology from Literature. In: Terminology, Computing and Translation, Hacken, P. (ed.) Tubingen, Gunter Narr Verlag, 2006, pp. 117–140.
- [2] **M. Vázquez, A. Oliver**, Improving Term Candidates Selection Using Terminological Tokens, Terminology. 2018, 24(1). P. 122–146. DOI: 10.1075/term.00016.vaz
- [3] **L. Bowker**, Off the Record and On the Fly: Examining the Impact of Corpora on Terminographic Practice in the Context of Translation. In: Corpus-Based Translation Studies: Research and Applications, Kruger, A. [et al.] (ed.) New York, Bloomsbury Publishing, 2011, pp. 115–127.
- [4] **K. Ahmad, M. Rogers**, Corpus Linguistics and Terminology Extraction. In: Handbook of Terminology Management: Vol. 2: Application-Oriented Terminology Management, Wright, S.E. and Budin, G. (ed.) Amsterdam, John Benjamins Publishing Company, 2001, pp. 725–760.
- [5] **A. Zaretskaya, G.C. Pastor, M. Seghiri**, Translators' Requirements for Translation Technologies: a user survey, Proceedings of the AIETI7 International Conference «New Horizons in Translation and Interpreting Studies» (Full papers). (2015) 247–254.
- [6] **A.Yu. Stankevich**, Poisk kontekstov i otsenka ikh tipichnosti sredstvami AntConc (Laurence Anthony) [Search for contexts and evaluation of their typicality by means of AntConc (Laurence Anthony)], Materialy V Mezhdunar. nauch.-metod. konf. «Teoriya i praktika prepodavaniya russkogo yazyka kak inostrannogo: dostizheniya, problemy i perspektivy razvitiya» [«Theory and practice of teaching Russian as a foreign language: achievements, problems and prospects of development». Proc. of the Int. Conference], Minsk, 2011, pp. 210–213.
- [7] **L. Anthony**, Issues in the Design and the Development of Software Tools for Corpus Studies: The Case for Collaboration. In: Contemporary Corpus Linguistics, Baker, P. (ed.) London, A&C Black, 2012, pp. 87–105.
- [8] **M. Weisser**, Practical Corpus Linguistics: An Introduction to Corpus-Based Language Analysis, John Wiley & Sons, Hoboken, 2016.
- [9] **I.A. Kotyurova**, Korpusnye issledovaniia s pomoshchiu servisa Antconc v usloviakh raboty v vuze [Corpus Research with Antconc in a Higher Education Setting], Nauchnyy zhurnal «Yazyk i kultura» [Scientific journal “Language and Culture”]. 52 (2020) 36–50. DOI: 10.17223/19996195/52/3
- [10] **K. Frantzi, S. Ananiadou, H. Mima**, Automatic Recognition of Multi-Word Terms, International Journal of Digital Libraries, 3 (2) (2000) 117–132.
- [11] **A. Ludeling, S. Evert, M. Baroni**, Using Web Data for Linguistic Purposes. In: Corpus Linguistics and the Web, Hundt, M., Nesselhauf, N., Biewer, C. (ed.) Amsterdam, Rodopi, 2007, pp. 7–25.
- [12] **A. Renouf, A. Kehoe, J. Banerjee**, Weborp: an Integrated System for Web Text Search. In: Corpus Linguistics and the Web, Hundt, M., Nesselhauf, N., Biewer, C. (ed.) Amsterdam, Rodopi, 2007, pp. 47–69.
- [13] **Y.-F. Huang, C.-S. Ciou**, Constructing Personal Knowledge Base: Automatic Key Phrase Extraction from Multi-Domain Web Page, Proceedings of International Workshops «New Frontiers in Applied Data Mining: PAKDD 2011» (Revised Selected Papers). (2012) 65–76.



- [14] **T.J. Lybbert, N. J. Zolas**, Getting Patents and Economic Data to Speak to Each Other: An “Algorithmic Links with Probabilities” Approach for Joint Analyses of Patenting and Economic Activity. World Intellectual Property Organization, 2012. 31 p.
- [15] **Qi Yanliang et al.**, Combining Supervised Learning Techniques to Key-Phrase Extraction for Bio-medical Full-Text. In: Organizational Efficiency through Intelligent Information Technologies, Sugumaran, V. (ed.) Hershey, Information. Science. Reference, 2013, pp. 33–45.
- [16] **V.P. Zakharov, M.V. Khokhlova**, Avtomaticheskoye vyyavleniye terminologicheskikh slovosochetaniy [Automatic Identification of Terminological Phrases], *Strukturnaya i prikladnaya lingvistika* [Structural and Applied Linguistics]. 10 (2014) 182–200.
- [17] **A. Kehoe**, Diachronic Linguistic Analysis on the Web with WebCorp. In: The Changing Face of Corpus Linguistics, Renouf, A., Kehoe, A. (ed) Amsterdam, Rodopi, 2016, pp. 297–309.
- [18] **V. Kosa et al.**, Cross-Evaluation of Automated Term Extraction Tools by Measuring Terminological Saturation, Proceedings of 13th International Conference “Information and Communication Technologies in Education, Research, and Industrial Applications” (Revised Selected Papers). (2017) 135–163.
- [19] **S.V. Savelyev**, Mesto sredstv avtomatizirovannogo izvlecheniya terminologii v rabote terminologa perevodcheskoy kompanii [Automated terminology extraction tools in the work of a translation company's terminologist], *Materialy Mezhdunar. nauch.-prakt. konf. «Aktualnyye voprosy lingvistiki i lingvodidaktiki: traditsii i innovatsii»* [“Current Issues in Linguistics and Linguodidactics: Traditions and Innovations”. Proc. of the Int. Conference], Moscow, 2018, pp. 66–71.
- [20] **M. Pavelec**, Translation Quality: Concepts, Procedures and Tools. In: Teaching Translation and Interpreting: Advances and Perspectives, Bogucki, Ł., Deckert, M. (ed.) Newcastle, Cambridge Scholars Publishing, 2012, pp. 137–146.
- [21] **L.N. Belyayeva**, Setevyye resursy v tekhnologii perevoda [Web Resources in Translation Technology], *Vestnik SPbGU. Filologiya. Vostokovedeniye. Zhurnalistika. State University Filologiya. Vostokovedeniye. Zhurnalistika* [Bulletin of Saint-Petersburg State University. Philology. Oriental Studies. Journalism]. 4 (2016) 45–55. DOI: 10.21638/11701
- [22] **Ch. Sin-Wai**, The Future of Translation Technology: Towards a World without Babel, Routledge, Abingdon-on-Thames.
- [23] **V.M. Leychik**, Prikladnoye terminovedeniye i yego napravleniya [Applied Terminology and its Prospects], *Prikladnoye yazykoznanie* [Applied Linguistics], 1996, pp. 276–286.
- [24] **J. Thomas**, Discovering English with Sketch Engine, Versatile, Brno, 2016.
- [25] **M.A. Kovyazina**, Izvlecheniye klyuchevykh terminov na baze korpusa tekstov o razrabotke neftnykh i gazovykh mestorozhdeniy [Key Terms Extraction based on the Corpus of Texts on Oil and Gas Field Development], *Vestnik Tyumenskogo Gosudarstvennogo Universiteta. Gumanitarnyye issledovaniya. Humanities* [Bulletin of the Tyumen State University. Humanitarian studies. Humanities]. 2 (3) (2016) 61–69. DOI: 10.21684/2411-197X-2016-2-3-61-69
- [26] **A.A. Novikova**, Sravneniye instrumentov Sketch Engine i TermoStat dlya izvlecheniya terminologii [Comparison of Sketch Engine and Thermostat tools for terminology extraction], *International Journal of Open Information Technologies (INJOIT)*. 8 (11) (2020) 73–79.
- [27] **A.S. Gerd**, Vvedeniye v izucheniye yazykov dlya spetsialnykh tseley [Introduction to the study of languages for special purposes], StPSU, RIO, Faculty of Philology. Saint-Petersburg, 2011.
- [28] **A.S. Lyubutov**, Metod strukturnoy taksonomii: vozmozhnosti primeneniya dlya analiza sotsialnykh i dukhovnykh protsessov [Structural Taxonomy: application possibilities for the analysis of social and spiritual processes], *Nauchnyy rezultat. Sotsiologiya i upravleniye* [Scientific result. Sociology and Management]. 5 (4) (2019) 58–79. DOI: 10.18413/2408-9338-2019-5-4-0-6
- [29] **A. Shatalkin**, Taksonomiya. Osnovaniya, printsipy i pravila [Taxonomy. Grounds, principles and rules], *Tovarishchestvo nauchnykh izdaniy KMK*, Moscow, 2012.
- [30] **S.I. Neizvestnyy**, O primenenii taksonomii v oblasti informatsionnykh tekhnologiy [On the application of taxonomy in the field of information technology], *Transportnyye sistemy i tekhnologii* [Transport systems and technologies]. 2 (1) (2016) 89–111.
- [31] **A. Shaykevich, V. Andryushchenko, N. Rebetskaya**, Distributivno-statisticheskiy analiz yazyka russkoy prozy 1850–1870-kh gg [Distributive and statistical analysis of the Russian prose language of the 1850s–1870s]. Vol. 1, *Yazyki slavyanskoy kultury*, Moscow, 2014.
- [32] **S.V. Grinev-Grinevich**, Terminovedeniye [Terminology], *Izdatelskiy tsentr «Akademiya»*, Moscow, 2008.

[33] **M.T. Cabré**, Terminology Theory, Methods and Applications, John Benjamins Publishing Company, Amsterdam, 1998.

[34] **Ye.Z. Zinder**, What is “IT EfficiencyΦ?”, Intelligent Enterprise, IE («Korporativnyye sistemy»): IT-zhurnal. 8 (141) (2006). Available at: <https://www.iemag.ru/master-class/detail.php?ID=1572> (accessed: 01.11.2021).

[35] **V.M. Leichik**, Realizatsiia kompleksnogo pokazatelya terminologichnosti v atributivnykh konstruktsiiakh [Implementation of a complex indicator of terminology in attributive constructions], Derivatsiia v norme i terminosistemakh [Derivation in normative and term systems], Vladivostok, 1990, pp. 64–74.

[36] **V.P. Danilenko**, O terminologicheskom slovoobrazovanii [On Terminological Word Formation], Voprosy yazykoznaniiya [Topics in the Study of Language]. 4 (1973) 76–85.

[37] **K. Kageura, V. Umino**, Methods of Automatic Term Recognition: a Review, Terminology. 3 (1996) 259–289.

[38] **S.D. Shelov**, Termin. Terminologichnost. Terminologicheskiye opredeleniya [Term. Termhood. Terminological definitions], Faculty of Philology of St. Petersburg State University, St. Petersburg, 2003.

[39] ELAN: Effects on the European Economy of Shortages of Foreign Language Skills in Enterprise, Directorate General for Education and Culture (European Commission). Available at: https://ec.europa.eu/assets/eac/languages/policy/strategic-framework/documents/elan_en.pdf (accessed: 20.04.2020).

[40] **V.M. Leychik**, Terminovedeniye: predmet, metody, struktura [Terminology: subject, methods, structure], Izdatelstvo LKI, Moscow, 2007.

СВЕДЕНИЯ ОБ АВТОРЕ / THE AUTHOR

Гацук Екатерина Юрьевна

Hatsuk Katsyaryna Yu.

E-mail: kadam@tut.by

ORCID: <https://orcid.org/0000-0003-1882-9311>

Статья поступила в редакцию 29.11.2021; одобрена после рецензирования 27.12.2021; принята к публикации 28.12.2021.

The article was submitted 29.11.2021; approved after reviewing 27.12.2021; accepted for publication 28.12.2021.