УДК 8.81.33

SKETCH ENGINE КАК ИНСТРУМЕНТ ДЛЯ ВЫЯВЛЕНИЯ СПЕЦИАЛЬНЫХ НОМИНАЦИЙ (НА ПРИМЕРЕ АНГЛОЯЗЫЧНЫХ ТЕКСТОВ ПРЕДМЕТНОЙ ОБЛАСТИ «ЯЗЫКОВАЯ ПОЛИТИКА»)

Е. Ю. Гацук
Гродненский государственный университет имени Янки Купалы
Республика Беларусь
E-mail: gacuk_ej@grsu.by

Проблема инвентаризации терминологии при исследовании системы понятий предметной области является актуальной в современных лингвистических исследованиях. В данной статье предлагается алгоритм работы с онлайн-инструментом Sketch Engine, который позволяет извлекать не только однословные, но и много-

словные специальные номинации для обозначения понятий, не зафиксированных в отраслевых терминологических словарях.

Ключевые слова: термин; терминологичность; автоматизация выявления терминов в тексте; Sketch Engine

Развитие предметных областей науки «сопровождается непрерывным конструированием идей и обозначающих их слов» [3, с,29], появлением новых терминов, номинирующих понятия, которыми оперирует определенная область знания. Выявление новых терминов, незафиксированных в существующих словарях, «на основе выполняемой в тексте функции называния понятий данной области» [1, с. 48] является актуальной задачей в процессе инвентаризации терминологии исследуемой области знания.

Цель данной статьи – продемонстрировать возможности онлайн-инструмента Sketch Engine для выявления специальных номинаций в текстах на примере предметной области «Языковая политика».

Известно, что все тексты, с точки зрения терминоведения, делятся на терминопорождающие (те, в которых представлены теории, описывающие специальные области знаний и деятельности), терминоиспользующие (те, в которых описываются объекты и процессы, относящиеся к соответствующей специальной области знания) и терминофиксирующие (те, в которых закреплены специальные номинации, использующиеся в определенной области знания) [1, с. 204]. Именно «в терминопорождающих и терминоиспользующих текстах можно наблюдать новые тенденции терминообразования, появляются новые термины» [1, с. 204].

Терминологическая работа состоит из нескольких направлений, одним из которых является инвентаризация терминологии, рассматриваемая как исходный этап терминологического менеджмента. Именно на данном этапе происходит отбор источников терминологии, извлечение специальных номинаций из текстов и определение степени их терминологичности (termhood).

Termhood (терминологичность) «refers to a degree that a candidate term is related to a domain specific concept (относится к степени, в которой термин-кандидат связан с понятием, специфичным для данной области) [6, с. 248] (здесь и далее перевод наш – Е.Г.). Критерий терминологичности, который, по мнению С.В. Гринева-Гриневича, следует заменить термином «*специальная область употребления*» [1, с. 26], базируется на специфичности употребления терминов. С.Д. Шелов, определяя терминологичность, утверждает, что «чем выше в рамках контекстуального определения встречаемость языковой

единицы, тем степень ее терминологичности должна быть больше, поскольку тем больше сведений необходимо учитывать для того, чтобы идентифицировать значение определяемой единицы» [5, с.145], «причем узкоспециальная терминология, по-видимому, относится к словам и словосочетаниям большей терминологичности, к терминологии “в чистом виде”» [4 с.48]. Таким образом, можно определить критерии для определения терминологичности: соотнесенность с определенной предметной областью и частотность употребления в специальных текстах этой области знания.

При обработке больших объемов текстов в современном терминологическом менеджменте используется специализированное программное обеспечение для автоматизации извлечения терминов. Наиболее популярными программными инструментами такого рода являются AntConc и Sketch Engine.

Корпус-менеджер AntConc оптимален для извлечения однословных специальных номинаций. Однако он не позволяет выявлять многословные номинации, т.к. при работе с данным инструментом не может быть использована технология сравнения с референтным корпусом.

В то же время, согласно мнению авторитетных ученых-терминологов, «многословные термины в большинстве европейских языков составляют 60 - 80% от общего количества терминов» [1, с. 121]. Для извлечения многословных специальных номинаций из англоязычных текстов предметной области «Языковая политика» был выбран онлайн-инструмент Sketch Engine.

Sketch Engine «contains 500 ready-to-use corpora in 90+ languages, each having a size of up to 30 billion words to provide a truly representative sample of language» (содержит 500 готовых к использованию корпусов на 90+ языках, каждый из которых имеет размер до 30 миллиардов слов, чтобы обеспечить действительно репрезентативную выборку) [7]. Для английского языка имеется корпус, собранный из опубликованных в Интернете текстов, English Web 2015 (enTenTen15), содержащий 15 млрд. слов (словоформ) и являющийся одним из самых крупных корпусов английского языка [2, с.107].

Для работы с инструментом Sketch Engine требуется регистрация, после которой предоставляется бесплатный доступ на 30 дней. По истечении данного срока пользователю необходимо оформить платную подписку, если необходимо продолжить работу. После регистрации пользователь попадает на главную страницу инструмента (см. рис.1):

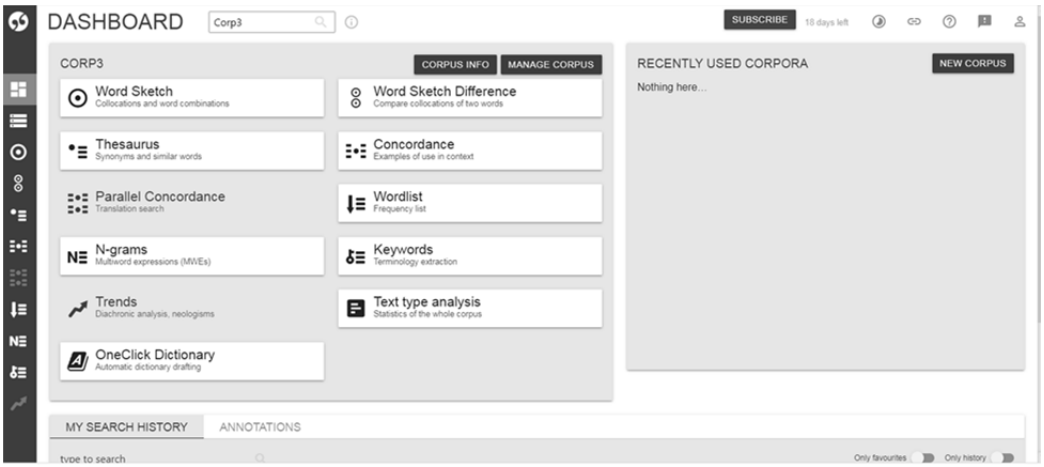


Рис.1. Внешний вид инструмента Sketch Engine

На данной странице пользователь может загрузить свой собственный заранее подготовленный текстовый корпус объемом не более 1000000 слов на одном из языков, которые поддерживает Sketch Engine (см. рис.2). В рамках проводимого исследования был сформирован корпус из текстов предметной области «Языковая политика», содержащий 3432122 слова, поэтому полученный текстовый корпус был поделен на несколько корпусов для удобства работы с инструментом Sketch Engine.

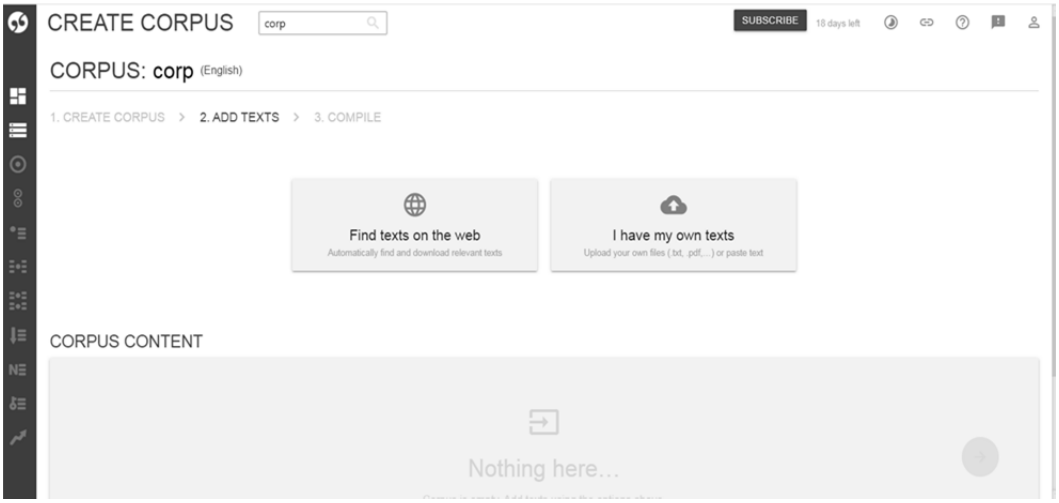


Рис.2. Добавление текстового корпуса для обработки

После добавления текстового корпуса инструмент начинает его обработку для последующего извлечения специальных номинаций (см. рис.3):

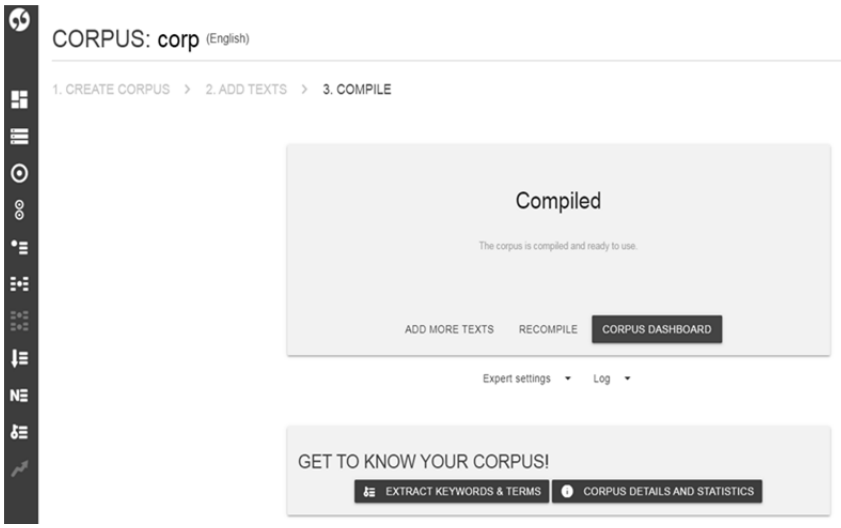
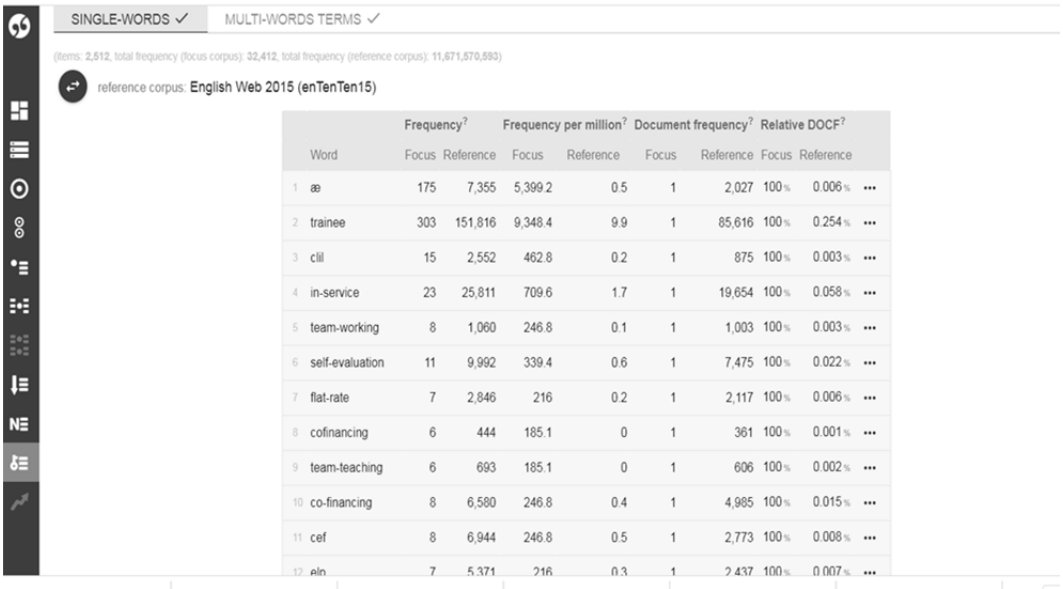


Рис.3. Извлечение специальных номинаций из текстового корпуса

В результате обработки корпуса Sketch Engine предлагает 3 группы специальных номинаций:

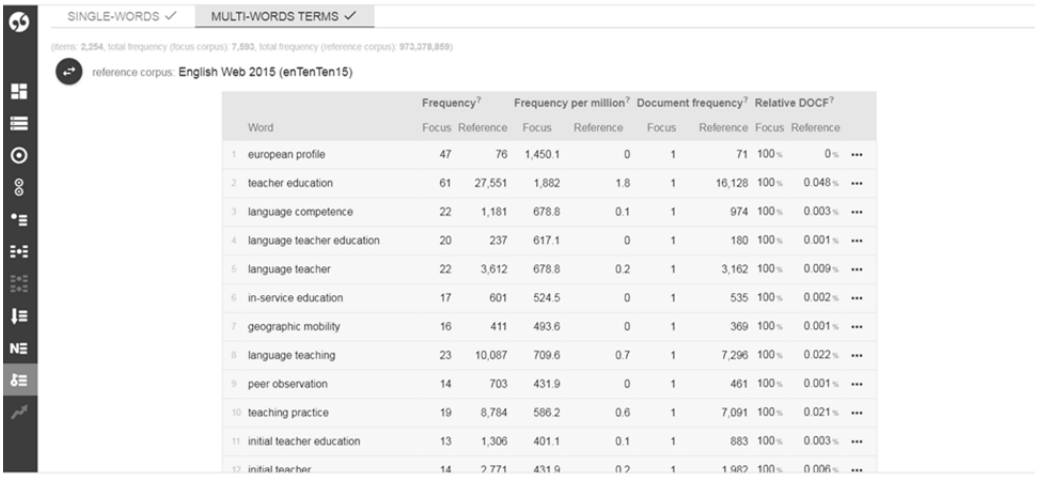
1) «keywords: individual words (any token can be included)» (ключевые слова: отдельные слова (может быть включен любой токен, т.е. словоупотребление) (см. рис.4) [8]:



	Word	Frequency		Frequency per million		Document frequency		Relative DOCF	
		Focus	Reference	Focus	Reference	Focus	Reference	Focus	Reference
1	æ	175	7,355	5,399.2	0.5	1	2,027	100%	0.006%
2	trainee	303	151,816	9,348.4	9.9	1	85,616	100%	0.254%
3	clil	15	2,552	462.8	0.2	1	875	100%	0.003%
4	in-service	23	25,811	709.6	1.7	1	19,654	100%	0.058%
5	team-working	8	1,060	246.8	0.1	1	1,003	100%	0.003%
6	self-evaluation	11	9,992	339.4	0.6	1	7,475	100%	0.022%
7	flat-rate	7	2,846	216	0.2	1	2,117	100%	0.006%
8	cofinancing	6	444	185.1	0	1	361	100%	0.001%
9	team-teaching	6	693	185.1	0	1	606	100%	0.002%
10	co-financing	8	6,580	246.8	0.4	1	4,985	100%	0.015%
11	cef	8	6,944	246.8	0.5	1	2,773	100%	0.008%
12	elp	7	5,371	216	0.3	1	2,437	100%	0.007%

Рис.4. Пример списка ключевых слов

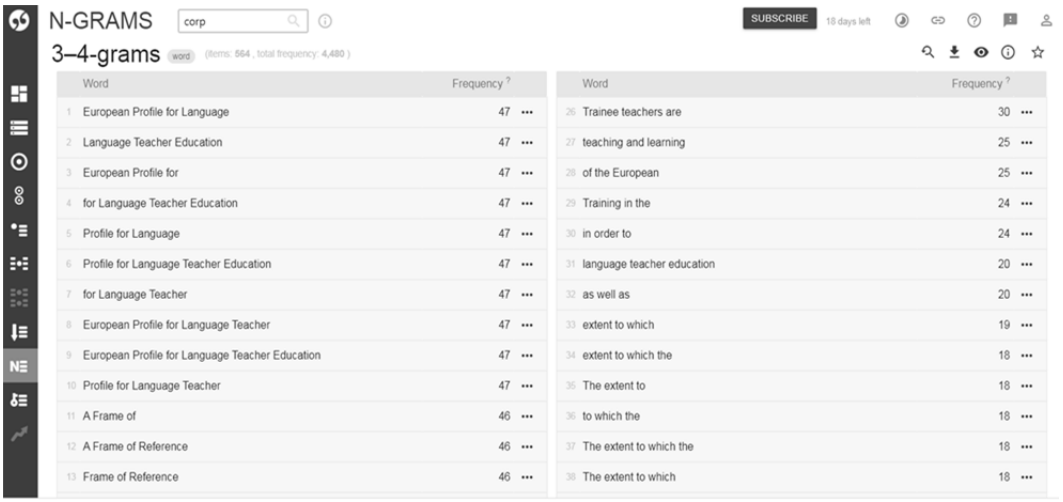
2) «terms: key multi-word expressions in a format typical of terminology in the language of the corpus» (термины: ключевые многословные выражения в формате, характерном для терминологии языка корпуса) (рис.5) [8]:



	Word	Frequency Focus	Frequency Reference	Frequency per million Focus	Frequency per million Reference	Document frequency Focus	Document frequency Reference	Relative DOCF Focus	Relative DOCF Reference
1	european profile	47	76	1,450.1	0	1	71	100%	0%
2	teacher education	61	27,551	1,882	1.8	1	16,128	100%	0.048%
3	language competence	22	1,181	678.8	0.1	1	974	100%	0.003%
4	language teacher education	20	237	617.1	0	1	180	100%	0.001%
5	language teacher	22	3,612	678.8	0.2	1	3,162	100%	0.009%
6	in-service education	17	601	524.5	0	1	535	100%	0.002%
7	geographic mobility	16	411	493.6	0	1	369	100%	0.001%
8	language teaching	23	10,087	709.6	0.7	1	7,296	100%	0.022%
9	peer observation	14	703	431.9	0	1	461	100%	0.001%
10	teaching practice	19	8,784	586.2	0.6	1	7,091	100%	0.021%
11	initial teacher education	13	1,306	401.1	0.1	1	883	100%	0.003%
12	initial teacher	14	2,771	431.9	0.2	1	1,982	100%	0.006%

Рис.5. Пример списка терминов

3) «N-grams: key multi-word expressions (any sequences of tokens). Only items which appear more frequently in the selected corpus than in the reference corpus are included. The results indicate what is typical of the selected corpus compared to the reference corpus» (N-граммы: ключевые многословные выражения (любая последовательность токенов). Включаются только те элементы, которые чаще встречаются в выбранном корпусе, чем в референтном корпусе. Результаты показывают, что типично для выбранного корпуса по сравнению с референтным корпусом) (рис. 6) [8]:



	Word	Frequency		Word	Frequency
1	European Profile for Language	47	26	Trainee teachers are	30
2	Language Teacher Education	47	27	teaching and learning	25
3	European Profile for	47	28	of the European	25
4	for Language Teacher Education	47	29	Training in the	24
5	Profile for Language	47	30	in order to	24
6	Profile for Language Teacher Education	47	31	language teacher education	20
7	for Language Teacher	47	32	as well as	20
8	European Profile for Language Teacher	47	33	extent to which	19
9	European Profile for Language Teacher Education	47	34	extent to which the	18
10	Profile for Language Teacher	47	35	The extent to	18
11	A Frame of	46	36	to which the	18
12	A Frame of Reference	46	37	The extent to which the	18
13	Frame of Reference	46	38	The extent to which	18

Рис.6. Пример списка N-грамм

Как следует из таблиц, представленных на рисунках 4 – 6, Sketch Engine позволяет не только выделить потенциальные термины из корпуса, но и увидеть частоту употребления в выбранном и референтном корпусах, частоту употребления потенциального термина на 1000000

словоупотреблений, а также дает возможность увидеть контексты употребления выделенных номинаций (рис.7).

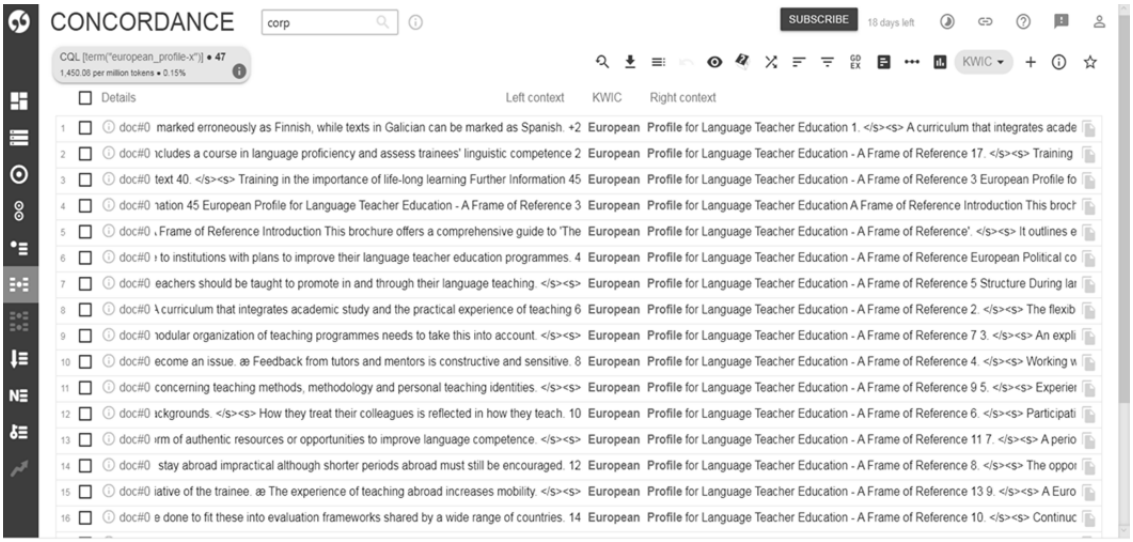


Рис.7. Контексты употребления выделенной номинации

Таким образом, в результате обработки корпуса текстов предметной области «Языковая политика» средствами онлайн-инструмента Sketch Engine было получено 4840 многословных специальных номинаций разной степени терминологичности. Анализ контекстов и частотность употребления полученных номинаций, которые предоставляет Sketch Engine, позволит в дальнейшем выявить, какие из данных номинаций могут выступать в качестве терминов для исследуемой области знания.

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. Гринев-Гриневич, С.В. Терминоведение : учеб. пособие для студ. высш. учеб. заведений / С. В. Гринев-Гриневич. – М. : Издательский центр «Академия», 2008. – 304 с.

2. Кротова, Е.Б. Sketch Engine для лингвистических исследований / Е.Б. Кротова // Германистика сегодня : материалы Международной научно-практической конференции, Казань, 16–17 октября 2018 г. / Казанский (Приволжский) федеральный университет ; ред.: М.А. Кулькова [и др.]. – Казань, 2019. –С. 107-112.

3. Суперанская, А.В. Общая терминология: Вопросы теории / А.В. Суперанская, Н.В. Подольская, Н.В. Васильева – Изд. 6-е. — М. : Книжный дом «ЛИБРОКОМ», 2012. — 248 с.

4. Шелов, С.Д. Очерк теории терминологии: состав, понятийная организация, практические приложения / С.Д. Шелов. – М. : ПринтПро, 2018. – 472 с.

5. Шелов, С.Д. Термин. Терминологичность. Терминологические определения / С.Д. Шелов. – Санкт-Петербург : Филологический факультет Санкт-Петербургского государственного университета, 2003. – 277 с.

6. Korkontzelos, I. Reviewing and Evaluating Term Recognition Techniques / I. Korkontzelos, I.P. Klapaftis, S. Manandhar // Advances in Natural Language

Processing: 6th International Conference, GoTAL 2008, Proceedings, 25 – 27August, 2008 / Gothenburg, Sweden. – Gothenburg, 2008. – P. 248-259.
7. Sketch Engine [Electronic resource]. – Mode of access: https://www.sketchengine.eu/#blue. – Date of access: 20.12.2020.
8. Sketch Engine [Electronic resource]. – Mode of access: https://app.sketchengine.eu/#keywords?corpname=user%2FEkaterin1%2Fcorp3. – Date of access: 20.12.2020.

SKETCH ENGINE AS A TOOL FOR IDENTIFYING SPECIAL NOMINATIONS (ON THE EXAMPLE OF THE ENGLISH TEXTS IN THE SUBJECT FIELD 'LANGUAGE POLICY')

E. Y. Hatsuk
Yanka Kupala State University of Grodno, Republic of Belarus
E-mail: gacuk_ej@grsu.by

The problem of inventory of terminology of the subject field is relevant in modern linguistic research. This article offers a guidance to work with the online tool Sketch Engine, which allows to extract not only one-word, but also multi-word special nominations of the concepts that are not recorded in the subject field dictionaries.

Keywords: term; terminology; term extraction; Sketch Engine