УДК 81´33

МЕТОДЫ ИЗВЛЕЧЕНИЯ ТЕРМИНОЛОГИЧЕСКИХ ЕДИНИЦ ИЗ КОРПУСА СОПОСТАВИМЫХ ТЕКСТОВ

В. Э. Рогачева

Российский государственный педагогический университет имени А. И. Герцена

Поступила в редакцию 17 декабря 2017 г.

Аннотация: *в статье рассматриваются методы извлечения терминов и терминологических словосочетаний из двуязычного сопоставимого корпуса предметно-ориентированных текстов. Из русского и английского подкорпусов извлекается ряд терминов и терминологических словосочетаний, которые составляют терминосистемы. Сопоставление терминосистем позволяет устанавливать и стандартизировать переводные соответствия между их компонентами.*
Ключевые слова: *терминология, извлечение терминологии, корпус текстов, критерии терминологичности.*

Abstract: *the paper discusses methods of term extraction from bilingual comparable text corpus of sphere-specific texts in Russian and English. The terms can be extracted from both Russian and English sub-corpora and considered as term systems. By comparison of the two term systems the translation equivalents can be found and standardized.*
Key words: *terminology, terminology extraction, text corpus, term criteria.*

Потенциал терминологических ресурсов, создаваемых в научно-технической лексикографии и используемых в работе не только переводчиков, но и специалистов в области анализа текстов на естественном языке (*language workers*) [1], делает актуальной разработку методов автоматического извлечения терминологических единиц из источников информации − текстов. Целью настоящей статьи является исследование возможности использования сетевых конкордансеров для извлечения терминологических единиц из текстов конкретной предметной области на русском и английском языках и установления пар типа термин – перевод.

При таком подходе текст рассматривается как результат решения задачи передачи информации и источник ее извлечения, поскольку знания не только фиксируются и представляются в виде текстов, но и порождаются в языковом смысле как текст [2]. Содержание научного текста, которое может быть извлечено при совпадении тезаурусов автора и получателя, определяет информация об объектах, описываемых именами – существительными и именными словосочетаниями. В научном тексте эти объекты обозначаются терминами. Корректный перевод терминов делает возможным извлечение и передачу информации в процессе перевода [3].

Процедуры извлечения информации основываются на идее сопоставимости контекстов слов с одинаковым значением в текстах на разных языках, представленных в виде параллельных корпусов [4]. При выделении таких контекстов кроме сложностей, связанных с отсутствием симметричности терминологических систем разных языков, особую проблему представляет отбор переводов для корпуса параллельных текстов, поскольку их качество часто не удовлетворяет требованиям корректности перевода терминов. Поэтому обращение к сопоставимым корпусам текстов, при организации которых возможна экспертная оценка текстов на сопоставляемых языках, вполне естественно. Сопоставимые корпуса текстов – это корпуса, чьи материалы отражают единую предметную область, но не являются переводами друг друга. У них есть ряд преимуществ над параллельными, так как дают возможность наблюдать реальное использование терминов в тексте, в то время как в переводах отражается влияние исходного текста, что затрудняет наблюдение терминов в языке перевода [5]. Кроме того, наблюдается недостаток доступных для использования параллельных текстов в различных предметных областях [5; 6].

Соответственно, в качестве источника для извлечения терминов в работе используются материалы международных конференций «Терминология и знание» за 2008–2012 гг. и «Terminology and Knowledge engineering» за 2010–2014 гг. Они организованы как сопоставимый корпус текстов* на русском и английском языках, на материале которого проводится исследование.

* Выражаем благодарность С. Д. Шелову и Т. Горностай за предоставленные материалы.

Материалы конференций – важный и доступный материал для составления сопоставимых корпусов, так как они ориентированы на единую предметную область, отражают современные тенденции лингвистической науки и включают в себя новые термины [6, с. 41]. Исследование проводится при помощи сетевой программы – конкордансера, предназначенного для лингвостатистического анализа текста AntConc [7, с. 27].

С учетом инструментария, предлагаемого программой, в рамках исследования для каждого языка необходимо решить следующие задачи:

• составить частотный словарь лексических единиц, зафиксированных в русском и английском подкорпусах;

• выделить из частотного словаря термины-универбы, учитывая существующие подходы к разграничению терминов и общеупотребительных слов;

• разработать процедуру исследования контекстов на основе использования конкордансов и кластеров, в которых встречаются наиболее частотные термины-универбы.

Конечной целью исследования является создание переводного словаря специальной лексики, отражающего терминополе «Терминология и инженерия знаний».

Частотный словарь по каждому подкорпусу текстов строится на основе сервиса World list программы AntConc. В полученном русском словаре 25 795 разных токенов – лексических единиц, в английском – 10 339.

Оба словаря содержат слова открытых классов (существительные, прилагательные, полнозначные глаголы и наречия), состав которых постоянно изменяется, и закрытых классов (союзы, предлоги, детерминаторы, verbs-to-be), остающихся неизменными в языковой системе [8]. Слова закрытых классов не представляют интереса для исследования в области терминологии и из дальнейшего рассмотрения исключаются.

Важным и признанным критерием для включения слова открытого класса в терминологический словник является его ранг в частотном словаре [9]. Слова, имеющие высокую частоту в корпусе специальных текстов, могут появляться с низкой частотой в корпусе текстов общего языка либо не появляться в нем вообще [8]. Следовательно, частота в случае специализированного корпуса является показателем важности лексической единицы для специальной коммуникации.

Для анализа наиболее частотных слов необходимо установить их значения, отражаемые в словарных толкованиях [10]. В английском частотном словаре к самым частым словам открытого класса относятся лексические единицы (ЛЕ) *terminology, term, language, concept*. В русском частотном словаре самыми частыми являются ЛЕ *термин, язык, слово и терминология*. Чтобы рассматривать данные лексические единицы как термины, являющиеся объектом дальнейшего анализа, отраженные в их толкованиях лексические значения должны быть постоянными, не соотноситься семантически с другими единицами в терминосистеме, а отграничиваться от них [11]. Далее в рамках исследования методов извлечения терминов из текстовых корпусов предполагается провести анализ словарных толкований частотных слов с целью определения их терминологического значения.

Кроме того, в верхней части частотных словарей зафиксированы такие слова, как *knowledge,* и его переводной эквивалент *знание*. Они относятся к общенаучным словам, но могут приобрести терминологичность, так как включены в терминополе специального текста [12]. Их терминологическое значение раскрывается в терминологических словосочетаниях.

Значение ЛЕ может быть уточнено в выделенных из корпуса словосочетаниях, в которых ЛЕ является опорным словом. При помощи сетевой программы AntConc можно получить данные об окружении лексических единиц, используя два сервисных инструмента для получения кластеров и конкордансов. Рассмотрим процедуру выделения словосочетаний с опорными ЛЕ *терминология – terminology*.

В соответствии с частеречной характеристикой этих ЛЕ предполагается, что программно выделяемый кластер – это совокупность именных групп (*noun phrases*).

Инструмент Clusters позволяет выделять словосочетания, в которых ЛЕ является ядром, а поясняющие ее лексические компоненты находятся или в препозиции, или в постопозиции к ядру. Частота появления каждого слова в каждой позиции вокруг ядра регистрируется, и самые частотные словосочетания, компоненты которых связаны между собой синтаксически и семантически, могут быть признаны значимой для лексикографа моделью сочетаний слов. Инструмент Concordance позволяет рассматривать окружение ЛЕ с обеих сторон. Это позволяет увидеть появление ЛЕ в двух значимых моделях сочетаний слов одновременно, что невозможно заметить в кластере. Поэтому результаты, выдаваемые кластерами и конкордансом, должны быть сопоставлены и объединены.

Для получения кластеров был задан минимум (1) и максимум (2) слов, окружающих ядерное слово справа и слева. Особенности морфологии русского и английского языков определяют процедуру анализа кластеров.

Развитая система падежных окончаний в русском языке делает необходимым лемматизацию ЛЕ для рассмотрения ее появления в кластере в разных падежных формах. В результате получаем основу терминолог* (символом «*» обозначаются падежные окончания).

Табл. 1 показывает, что не каждый элемент кластера является именной группой, что видно на при-

мере конструкций «*терминология изображается*» и «*терминология имеет*». Кроме того, эта основа соответствует и другим ЛЕ, таким как *терминологический* или *терминологизация* (табл. 1) [4].

Таблица 1

Кластер для основы ***терминолог**** *(фрагмент)*

Ранг	Частота	Длина с/с	Элементы кластера
1	98	2	Терминологический словарь
2	42	2	Терминологическая единица
3	22	2	Терминологическое определение
4	24	2	Терминологическое описание
5	20	2	Терминологическая работа
6	31	2	Терминологический банк
7	26	2	Терминологическая система
8	31	2	Терминологических сочетаний
9	3	3	Терминологизация общего слова
10	3	3	Терминологизация языкового знака
11	2	2	Терминология изображается
12	2	2	Терминология имеет

Следующим шагом является составление списка всех ЛЕ, входящих в именные группы и включающих в себя искомую лемму, и их последующий компонентный анализ. ЛЕ *терминология*, включает в себя семантический компонент *область знаний.* Соответственно элементы кластера, в которых этот компонент отсутствует, дальнейшему анализу не подлежат.

Из общего количества полученных словосочетаний в кластере выбраны те, которые представляют собой именные группы (ИГ) с ядром *терминология*. В одном из кластеров ядерное слово находится слева (табл. 2), в другом – справа (табл. 3) [4].

Таблица 2

Именные группы с ядром ***терминология*** *в правой позиции*

Частота	Именная группа
8	Терминология православия
3	Терминология налогообложения
3	Терминология прикладной науки
3	Терминология фармакогнозии
2	Терминология в системе языка
2	Терминология различных областей
2	Терминология различных областей знаний
2	Терминология разных областей
2	Терминология разных областей знаний
2	Терминология языка
2	Терминология языкознания
2	Терминология обрядов
1	Терминология базовых понятий
1	Терминология международных сообществ
1	Терминология поддержки
1	Терминология поддержки всего диапазона
1	Терминологии в грамматике
1	Терминологии в упорядочении

Таблица 3

Именные группы с ядром ***терминология*** *в левой позиции*

Частота	Именная группа
20	Лингвистическая терминология
10	Народная терминология
10	Научная терминология
7	Археологическая терминология
7	Народная терминология
6	Литовская терминология
6	Научная терминология
5	Лингвистическая терминология
5	Православная терминология
5	Профессиональная терминология
4	Научнонормативная терминология
3	Научная терминология
3	Отраслевая терминология

Из множества результатов употребления ЛЕ в конкордансе также выбираются те, в которых присутствует сема *область знаний,* с последующей лемматизацией. Для слова *терминология* получен конкорданс, состоящий из 501 употребления слова в контексте. В каждой строке конкорданса выделены ИГ, в которые входит слово *терминология*, притом учитываются определения к ядру. Это позволяет рассмотреть исследуемый термин в более широком контексте, так как в кластере упускаются однородные определения к термину, управляющие и управляемые им существительные, которые также могут иметь определения. С помощью инструментов поиска в конкордансе выделены и подсчитаны частоты наиболее часто встречающихся слов, связанных с ядерным термином (табл. 4) [4]. При этом условие, стоят ли эти слова в непосредственной близости к термину и имеют ли определения и дополнения, игнорируется. В списках, полученных с помощью конкорданса и кластеров, ищутся соответствия.

Таблица 4

Частотные словосочетания в конкордансе

Словосочетание	Частота
Научная терминология	37
Народная терминология	32
Лингвистическая терминология	23
Профессиональная терминология	8
Терминология православия	8
Археологическая терминология	8
Отраслевая терминология	7
Православная терминология	5
Медицинская терминология	5
Предметная терминология	4
Научнонормативная терминология	4
Упорядочение терминологии	4
Терминология прикладной науки	4
Обрядовая терминология	3
Географическая терминология	3
Терминология языка для специальных целей	2
Понятийно четко очерченная терминология	2

Для англоязычного подкорпуса была проделана аналогичная работа: в конкордансе выделены словосочетания с ЛЕ *terminology*. С помощью сравнения конкорданса с данными, полученными в результате исследования ЛЕ инструментом Clusters, подсчитана частота словосочетаний.

ЛЕ *terminology* оказалась зависимым элементом словосочетания в большинстве выделенных словосочетаний. В именной группе ЛЕ является либо ядром, либо атрибутом ядерного компонента. В переводных эквивалентах именных групп она выражается либо прилагательным (*terminology resources – терминологические ресурсы)*, либо существительным, выступающим в роли дополнения к ядерному элементу словосочетания (*terminology teaching – обучение терминологии*). Словосочетания, в которых ЛЕ является ядром, отбираются и сравниваются со списком словосочетаний на русском языке, в которых ядро – русскоязычный эквивалент ЛЕ.

В табл. 5 приведены ранжированные по частоте списки словосочетаний, в которых представлены возможные переводные эквиваленты *терминология – terminology* [4].

Таблица 5

Списки выделенных словосочетаний

1. Научная терминология	1. Frame-based terminology
2. Народная терминология	2. Concept-oriented terminology
3. Лингвистическая терминология	3. Ontology-based terminology
4. Профессиональная терминология	4. Discourse-purposed terminology
5. Терминология православия	5. Bibliographic terminology
6. Археологическая терминология	6. Application terminology
7. Отраслевая терминология	7. Specific terminology.
8. Православная терминология	8. Industrial practice terminology
9. Медицинская терминология	9. EU terminology
10. Предметная терминология	10. Theoretical terminology
11. Терминология прикладной науки	11. Applied terminology
12. Научнонормативная терминология	12. Phraseological terminology
13. Обрядовая терминология	13. Legal terminology
14. Терминология языка для специальных целей	14. Acquired terminology
15. Понятийно четко очерченная терминология	15. Well-developed terminology
	16. Mature terminology
	17. Corresponding terminology

Из списков можно выделить 4 пары эквивалентов:

– терминология языка для специальных целей – *specific terminology*;

– терминология прикладной науки – *applied terminology*;

– отраслевая терминология – *industrial practice terminology*;

– понятийно четко очерченная терминология – *well-developed terminology.*

В сопоставимом корпусе текстов соответствия между всеми терминологическими словосочетаниями не обнаруживаются, однако в них всех прослеживается общая семантическая модель. Терминологические словосочетания, к которым эквиваленты не найдены, необходимо перевести, чтобы можно было из всех словосочетаний из сопоставимого списка составить единую семантическую структуру. Принимая, что ЛЕ *терминология (terminology)* – это гипероним и вершина в семантической иерархии понятий, мы признаем выделенные с помощью инструментария AntConc словосочетания потенциальными гипонимами в семантической структуре, которые могут классифицироваться по различным признакам и объединяться между собой как согипонимы. Иерархическая структура, в которую входят терминологические словосочетания, может быть использована для построения терминологических баз данных, которые могут применяться в научно-технической лексикографии, при автоматической обработке текстов и машинном переводе.

ЛИТЕРАТУРА

1. *Беляева Л. Н.* Лингвистические технологии в современном сетевом пространстве : language worker в индустрии локализации / Л. Н. Беляева. – СПб. : Книжный дом, 2016. – 134 с.

2. *Чернявская В. Е.* Лингвистика текста. Лингвистика дискурса / В. Е. Чернявская. – М. : УРСС-Либроком, 2013. – 224 с.

3. *Беляева Л. Н.* Проблемы перевода текстов на языках для специальных целей / Л. Н. Беляева // Акту-

альные проблемы гуманитарных и социальных наук : сб. трудов участников II Междунар. науч.-практ. конф. (3 февраля 2015 г.). – СПб. : Санкт-Петербургский университет управления и экономики, 2015. – С. 14–17.

4. *Рогачева В. Э.* Методы извлечения терминологических эквивалентов из двуязычного корпуса текстов / В. Э. Рогачева // Прикладная лингвистика в науке и образовании ALPAC REPORT – полвека после разгрома : труды VIII Междунар. науч. конф. (Санкт-Петербург, 24–26 ноября 2016 г.). – СПб. : Книжный дом, 2016. – С. 91–96.

5. *Delpech E.* Dealing with lexicon acquired from comparable corpora : validation and exchange / E. Delpech, B. Daille // Proceedings of 9th Conference on Terminology and Knowledge Engineering (TKE). – Fiontar : Dublin City University, 2010. – P. 229–223.

6. *Беляева Л. Н.* Системы и процедуры выделения терминов из текстов / Л. Н. Беляева // Терминология и знание : материалы IV Междунар. симпозиума (Москва, 6–8 июня 2014 г.) / отв. ред. С. Д. Шелов. – Вып. IV. – М. : Вест-Консалтинг, 2014. – С. 212–223.

7. *Беляева Л. Н.* Сетевой инструментарий филолога : учеб. пособие / Л. Н. Беляева, К. Р. Пиотровская. – СПб. : Книжный дом, 2014. – 48 с.

8. *Gillam L.* Terminology and the construction of ontology / L. Gillam, M. Tariq, K. Ahmad. – Mode of access: http://clara.b.uib.no/files/2011/06/Gillam-Tariq-and-Ahmad.pdf

9. *Кудашев И. С.* Проектирование переводческих словарей специальной лексики / И. С. Кудашев. – Helsinki : Helsinki University Print, 2007. – 443 с.

10. *Апресян Ю. Д.* Лексическая семантика / Ю. Д. Апресян. – М. : Наука, 1974. – 346 с.

11. *Шелов С. Д.* Определение терминов и понятийная структура терминологии / С. Д. Шелов. – СПб. : Санкт-Петербургский университет, 1998. – 233 с.

12. *Табанакова В. Д.* Переводчик-лингвист, переводчик-терминолог, переводчик-специалист : стратегия и тактика перевода термина / В. Д. Табанакова // Вестник Тюмен. гос. ун-та. – 2014. – № 1 (Филология). – С. 72–81.

Российский государственный педагогический университет имени А. И. Герцена

Рогачева В. Э., аспирант кафедры образовательных технологий в филологии

E-mail: valery.xxx.zzz@gmail.com

Тел.: 8-967-432-05-37

State Pedagogical University named after A. I. Herzen

Rogacheva V. E., Post-graduate Student of the Education Technologies in Philology Department

E-mail: valery.xxx.zzz@gmail.com

Tel.: 8-967-432-05-37