3. Арапова Н. С., Кимягарова Р. С. Словарь иностранных слов. М.: Цитадель, 1999. 784 с.

4. Новейший философский словарь. Минск: Книжный дом А. А. Грицанов, 2003. 1280 с.

5. Cherkasova M. N., Patyukova R. V., Kudinova T. A., Olomskaya N. N. The Language of Insult: The Conflict of Interpretations (Linguo-Juridical Approach to the Problem) // Man in India. 2017. Vol. 97(15). P. 527–538.

6. Гай Саллюстий Крисп. Сочинения / пер., вступ. ст. и коммент. В. О. Горенштейна. М.: Ладомир, АСТ, 1999. 222 с.

7. Marunevich O. V., Kononenko A. P . The "In-Group – Out-Group" Binary Opposition as the Criterion for Identifying the Out-Group Members in the Folk Model of the World (Based on Slavic Languages) // Russian Linguistic Bulletin. 2020. N 2(22). P. 33–37.

8. Юрков С. Е. Под знаком гротеска: антиповедение в русской культуре (XI – начало XX вв.). СПб.: Летний Сад, 2003. 210 с.

9. Marunevich O. V., Kotlyarenko I. Yu., Bessarabova O. N., Shefieva E. Sh., Bobrova T. O. Cognitive-Contrastive Analysis of Somatism-Based Ethnophaulisms in English and Russian [Электронный ресурс] // Modern Global Economic System: Evolutional Development vs. Revolutionary Leap. ISC 2019. Lecture Notes in Networks and Systems, Springer, Cham, 2021. Vol. 198. P. 501–509. URL: https://doi.org/10.1007/978-3-030-69415-9_57

УДК 81

А. А. Новикова (Санкт-Петербург, Россия)
Санкт-Петербургский государственный университет

Методы корпусной лингвистики в терминоведении

В статье рассматривается значение методов корпусной лингвистики для решения лексикографических и терминологических задач. Отмечаются особенности работы с инструментами для автоматического извлечения терминов-кандидатов из текстового корпуса.

Ключевые слова: терминоведение, термин, терминосистема, корпус текстов, язык для специальных целей, терминология

Корпусная лингвистика является активно развивающимся направлением в языкознании. Методы корпусной лингвистики используются для решения задач машинного перевода, терминоведения, лексикографии и терминографии. Специальная, терминологическая лексика фиксируется в специальных словарях, поэтому возникло новое направление – терминография, связанное с лексикографией и терминоведением. «Терминография связана с терминоведением, поскольку предметом анализа в обоих случаях является термин» [2, с. 8].

Специальные корпусы позволяют терминологам разрабатывать терминологические словари на основе актуального, современного текстового материала. Более того, на основе корпусов создаются и пополняются тезаурусы и терминологические базы данных. С помощью корпусов возможно проанализировать все контексты употребления терминов или терминологических сочетаний, то есть изучить термин в «естественных условиях его существования» [1, с. 180]. Корпус текстов представляет собой языковую модель, в которой зафиксированы фонетический, морфологический, синтаксический и семантический уровни. Языковой материал, полученный из корпуса, как правило, точен, потому что отражает современное состояние языка. Отметим, что в этом случае важно учитывать хронологические рамки текстового материала для корпуса.

Под корпусом текстов понимается большой, представленный в электронном виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач [3, с. 7]. Считается, что корпус является таковым при наличии разметки (аннотации) – морфологической или синтаксической. Для решения различных лингвистических задач возможно самостоятельно создать специальный корпус. С помощью программ – корпусных менеджеров возможно работать с корпусом, извлекать информацию о частоте, строить конкордансные списки (выявлять контексты употребления слова). Такие корпусные менеджеры, как Sketch Engine [8], позволяют автоматически извлекать односложные и многосложные термины-кандидаты, получать статистическую информацию, формировать ассоциативный тезаурус на основе данных о частоте. Помимо Sketch Engine довольно известным является корпусный менеджер AntConc [6] со схожим функционалом. Извлечение терминов-кандидатов из корпусов – важная задача при изучении и описании терминологии любой предметной области. Более того, использование подобных программ упрощает терминологическую работу, связанную с отбором терминов для отраслевых словарей. Автоматическое извлечение терминов-кандидатов выполняется с помощью специальных программ, которые подсчитывают частоту совместной встречаемости слов на основе различных мер ассоциации: MI (mutual information), t-score, z-score, log-likelihood, chi-square ($\chi 2$) и др. Программа для автоматического извлечения терминов TermoStat [9] позволяет получать статистическую информацию об извлеченных терминах, список биграмм (многосложных терминов), список семантически связанных слов. В целом подобные программы извлекают термины довольно точно, о чем свидетельствует эксперимент по автоматическому извлечению терминов, описанный автором в статье [5].

Для сравнения работы функции извлечения терминов-кандидатов программ Sketch Engine и TermoStat использован созданный автором специальный корпус технических текстов по тематике «Водоснабжение» на английском языке. При сравнении функционала данных программ не выявлено существенных различий в полученных результатах. Обе программы являются хорошими инструментами для решения задач извлечения терминов-кандидатов и последующего отбора терминов для отраслевых словарей или глоссариев. Использование статистических и лингвистических методов в архитектуре этих программ позволяет получать точные результаты. Справедливо отметить, что при формировании словника для отраслевого словаря недостаточно пользоваться только корпусными методами или программами для извлечения терминологии. Необходима также консультация эксперта – специалиста предметной области.

Отметим, что для лексикографической и терминологической работы с корпусом характерны следующие этапы:

- формирование корпуса текстов, которое предполагает сбор текстового материала одной тематики;
- выбор корпусного менеджера;
- извлечение терминов-кандидатов из корпуса (вручную или автоматическими способами);
- отбор релевантных терминов (при необходимости с привлечением эксперта).

Интересным корпусным инструментом является программа Google Books N-gram Viewer [7], которая позволяет проанализировать языковой материал за определенный пользователем исторический период, просмотреть контексты употребления слов, построить графики встречаемости слов или терминов в текстах. Подобная информация представляет интерес не только с лингвистической точки зрения, но и с точки зрения культурологии и истории. В работе А. Ц. Масевича и В. П. Захарова отмечена связь корпусной лингвистики с понятием культурометрии, под которым подразумевается «исследование культуры человечества, направлений её развития во времени посредством количественного анализа слов и словосочетаний в очень больших корпусах оцифрованных текстов» [4, с. 25]. Авторами проведено исследование социально-политических явлений на основе частотного поведения лексических единиц в большом корпусе текстов с использованием программы Google Books N-gram Viewer, включающей 10 млн книг [4].

При работе с терминами Google Books N-gram Viewer позволяет получить интересные результаты, на основе которых можно сделать выводы о значимости тех или иных понятий в обществе в определенный промежуток времени.

Например, при анализе терминов предметной области «Водоснабжение» частотное поведение терминов (высокая частота встречаемости терминов в текстах за период с XIX по XX вв.) *питьевая вода, очистка воды, система водоснабжения, качество воды* указало на тесную взаимосвязь с историческими событиями в России, а также на значимость этих понятий в обществе начиная с XIX века.

В заключение отметим, что полученная из корпусов информация представляет интерес для составления тезаурусов и словарей различных типов. Корпусные менеджеры и программные инструменты демонстрируют широкие возможности для исследования и описания терминологии любых предметных областей.

Литература

1. Амитрова М. В., Гусарова Ю. В., Крюкова С. В., Павлова О. А. Специальная терминология и различные подходы к её классификации (на примере английского и французского языка) // XXI век: итоги прошлого и проблемы настоящего плюс: период. науч. издание. Пенза: Изд-во Пенз. гос. технол. ун-та, 2015. № 01(23). Т. 2. С. 178–182.
2. Гринев-Гриневич С. В. Терминоведение: учеб. пособие. М.: Академия, 2008. 303 с.
3. Захаров В. П., Богданова С. Ю. Корпусная лингвистика. Иркутск, ИГЛУ, 2011. 161 с.
4. Масевич А. Ц., Захаров В. П. Методы корпусной лингвистики в исторических и культурологических исследованиях // Компьютерная лингвистика и вычислительные онтологии: труды XIX междунар. объед. науч. конф. 2016. С. 24–43.
5. Новикова А. А. Сравнение инструментов Sketch Engine и TermoStat для извлечения терминологии // International Journal of Open Information Technologies. 2020. Т. 8. № 11. С. 73–79.
6. AntConc [Электронный ресурс]. URL: https://www.laurenceanthony.net/software/antconc/
7. Google Books N-gram Viewer [Электронный ресурс]. URL: https://books.google.com/ngrams
8. Sketch Engine [Электронный ресурс]. URL: https://www.sketchengine.eu/
9. TermoStat [Электронный ресурс]. URL: http://termostat.ling.umontreal.ca/index.php