

сознания в языке и коммуникации // Представление эмоциональной сферы человека на материале разных языков: коллективная монография. М.: ИЯ РАН, 2020. С. 70–123.

2. Большой толковый словарь русского языка / под ред. С. А. Кузнецова. URL: <https://gramota.ru/biblioteka/slovari/bolshoj-tolkovyy-slovar>

3. Словарь русского языка: В 4-х т. / под ред. А. П. Евгеньевой. М.: Рус. яз.; Полиграфресурсы, 1999.

4. *Фасмер М.* Этимологический словарь русского языка. Т. 1–4. М., 1986–1987. <https://lexicography.online/etymology/vasmer/>

5. *Фестингер Л.* Теория когнитивного диссонанса. М.: Эксмо, 2018. 256 с. (Серия: Классика психологии).

6. *Толстая С. М.* Вера и правда: к истории понятий // Труды института русского языка им. В. В. Виноградова. 2016. № 9. С. 375–388.

7. *Фуллер С.* Постправда. Знание как борьба за власть. URL: <https://biblioclub.ru/index.php?page=book&id=615641>

8. Прологомены когнитивной безопасности: коллективная монография / под ред. И. Ф. Кефели. СПб.: ИД «Петрополис», 2023. 488 с.

© **Е. П. Иванова, М. В. Корышев,
М. В. Хохлова**

*Санкт-Петербургский
государственный университет,
г. Санкт-Петербург, Россия*

АВТОМАТИЧЕСКОЕ ИЗВЛЕЧЕНИЕ КОЛЛОКАЦИЙ ИЗ БОЛЬШИХ КОРПУСОВ РУССКОГО, НЕМЕЦКОГО И ФРАНЦУЗСКОГО ЯЗЫКОВ: СРАВНЕНИЕ КОРПУСНЫХ И СЛОВАРНЫХ ДАННЫХ

Применение статистических методов к большим языковым данным дает возможность по-новому взглянуть на явление сочетаемости, определяющееся как грамматическими и лексическими особенностями, так и вероятностными характеристиками. В работах, посвященных автоматическому извлечению коллокаций, обсуждается использование тех или иных статистических мер [1], при этом уделяется мало внимания возможностям их комбинирования (как, например, в работе [2]). Также отдельного изучения заслуживают коллокации, полученные при помощи одних и тех же метрик на основе разных языков.

Целью нашей работы является сопоставление результатов автоматического выделения коллокаций статистическими методами из корпусов трех языков (русский, немецкий, французский), что позволит описать возможности и ограничения статистического аппарата применительно к задаче извлечения устойчивых словосочетаний в разных языках. В качестве материала для исследования нами были

использованы корпуса большого объема (более 15 млрд слов каждый) семейств TenTen [3] и Aranea [4], которые были созданы автоматически на основе текстов Интернета разных жанров. При помощи статистических метрик (MI, t-score, log-likelihood, logDice и др.) были извлечены словосочетания разных типов, а также произведено сравнение выделенных единиц с единицами, представленными в словарях данных трех языков (своего рода золотым стандартом). Результаты проведенного анализа показали, что в случае русского языка метрики t-score и logDice извлекают наибольшее количество примеров, которые зафиксированы в лексикографических источниках, однако для немецкого языка метрика log-likelihood показывает иные результаты, выделяя наиболее частотные единицы.

Результаты работы востребованы в специализированных корпусо-ориентированных ресурсах. Количественные данные о сочетаемости и списки выделенных сочетаний могут быть использованы при составлении словарей и грамматик, при автоматической кластеризации слов и лексических конструкций, при разрешении лексической неоднозначности и др.

Примечания

1. *Evert S. Corpora and collocations.* // A. Lüdeling, & M. Kytö (Eds.), *Corpus Linguistics. An International Handbook.* Berlin, Germany: de Gruyter, 2008. P. 223–233.
2. *Klyshinsky E., Khokhlova M.* In Search of Lost Collocations: Combining Measures to Reach the Top Range // *Internet and Modern Society: Proceedings of the International Conference IMS-2017 (St. Petersburg; Russian Federation, 21–24 June 2017)* / ed. by Radomir V. Bolgov, Nikolai V. Borisov, Leonid V. Smorgunov, Irina I. Tolstikova, Victor P. Zakharov. ACM International Conference Proceeding Series. N.Y.: ACM Press, 2017. P. 160–163.
3. *Benko V.* Aranea: Yet Another Family of (Comparable) Web Corpora // *Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8–12, 2014* / ed. by Petr Sojka, Aleš Horák, Ivan Kopeček and Karel Pala. Proceedings. LNCS 8655. Springer International Publishing Switzerland, 2014. P. 257–264.

© Н. А. Илюхина

*Самарский национальный исследовательский
университет им. академика С. П. Королева,
г. Самара, Россия*

О ЛЕКСИКОГРАФИЧЕСКОМ ОТРАЖЕНИИ СЦЕНАРНОЙ МЕТОНИМИИ (НА МАТЕРИАЛЕ ГЛАГОЛОВ ДВИЖЕНИЯ)

Отражение метонимических значений в толковых словарях не является последовательным и по причине массовости и свободе их образования в речи не может претендовать на более или менее пол-