**Премия СПбГУ за научные труды**
**в номинации «за вклад в науку молодых исследователей (в возрасте до 35 лет)»**

**Хохлова Мария Владимировна,** кандидат филологических наук, доцент кафедры математической лингвистики СПбГУ

**Цикл работ на тему «Исследование сочетаемости и частотных характеристик лексических единиц в корпусах текстов русского языка»**

1. Khokhlova M. Similarity between the Association Measures: a Case Study of Noun Phrases. In Proceedings of the Twelfth Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2018. Edited by Horák A., Rychlý P., Rambousek A. Brno: Tribun EU, 2018. P. 21–27. (**Scopus**)

https://nlp.fi.muni.cz/raslan/2018/paper05-Khokhlova.pdf

2. Khokhlova M. Building a Gold Standard for a Russian Collocations Database. In Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts. Ljubljana, 2018. P. 863–869. (**Scopus**)

https://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202018/118-4-2958-1-10-20180820.pdf

3. Klyshinsky E., Khokhlova M. In Search of Lost Collocations: Combining Measures to Reach the Top Range. In Internet and Modern Society: Proceedings of the International Conference IMS-2017 (St. Petersburg; Russian Federation, 21-24 June 2017). Eds 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. DOI: 10.1145/3143699.3143731 (**Scopus**)

https://dl.acm.org/citation.cfm?id=3143731&dl=ACM&coll=DL

4. Khokhlova M. Comparison of High-Frequency Nouns from the Perspective of Large Corpora. In Proceedings of the Tenth Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2016. Edited by Horák A., Rychlý P., Rambousek A. Brno: Tribun EU, 2016. P. 9-17. (**Scopus**)

https://nlp.fi.muni.cz/raslan/2016/paper05-Khokhlova.pdf

5. Khokhlova M., Rosso P., Patti V. Distinguishing between Irony and Sarcasm in Social Media Texts: Linguistic Observations. In Proceedings of the International FRUCT Conference on Intelligence, Social Media and Web, ISMW FRUCT 2016. P. 17-22. (**Scopus**)

https://ieeexplore.ieee.org/document/7584765

## Аннотация

Работа к. ф. н. Марии Владимировны Хохловой посвящена важной проблеме — исследованию частотных и сочетаемостных характеристик лексических единиц на материале корпусов текстов русского языка. В последнее десятилетие лингвисты стали чаще обращаться к корпусному материалу, который позволяет по-новому взглянуть на явления, многократно описанные в литературе, в том числе на сочетаемость. Несмотря на то, что существуют словари, описывающие сочетаемость, ощущается нехватка в работах и ресурсах, которые бы описывали данное явление на материале корпусных данных в гораздо более полном объеме.

Цель работы заключается в полноценном описании частотных и сочетаемостных характеристик лексических единиц на материале корпусов текстов с привлечением междисциплинарных методов, которые включают в себя дистрибутивный анализ и статистические методы.

Само явление сочетаемости имеет двоякую природу, которая объясняется, с одной стороны, лингвистической составляющей и, с другой стороны, вероятностными (комбинаторными) особенностями. Для статистической оценки связанности элементов словосочетаний (под ней понимается количественный показатель, основанный на частоте совместной встречаемости слов, которые входят в описываемую языковую структуру) применялись меры ассоциации, которые насчитывают более 80 метрик. В статьях М. В. Хохловой был сделан обзор мер ассоциации, которые редко упоминаются в научной литературе (меры Жаккарда, MS, z-score, log-log, истинная взаимная информация и др.). Некоторые из существующих мер были апробированы автором на русскоязычном материале и результаты экспериментов нашли отражение в статьях (работа № 1). Среди найденных биграмм были выделены следующие группы словосочетаний, обладающие разной степенью устойчивости и воспроизводимости: имена собственные («город Эберсвальде»), свободные словосочетания («японский издатель»), термины («контрольная палата»), устойчивые словосочетания («подвергнуться деформации») и фразеологизмы («короткий поводок»). Важно также исследовать, насколько достоверными являются автоматически полученные данные. В связи с этим автором был поставлен вопрос о создании золотого стандарта лексической сочетаемости для русского языка, который должен быть использован для оценки полученных результатов (работа № 2). М. В. Хохловой разрабатывается база данных, в которую войдут словосочетания из лексикографических источников, снабженные информацией, которая была получена в ходе экспериментов по автоматическому выявлению сочетаний. Таким образом, произойдет сопряжение классического подхода, под которым понимается описание сочетаемости в словарях и грамматических справочниках, и современных технологий, включающих корпусный и статистический подходы, а также обработку больших данных. Было произведено сравнение с данными, приведенными в толковых и специализированных словарях русского языка, а также была дана оценка используемым метрикам (MI, t-score, коэффициент логарифмического правдоподобия, MI3, MS, коэффициент Дайса и др.). Было показано, что наилучшие результаты демонстрирует мера logDice, которая выделяет как частотные воспроизводимые словосочетания, так и сочетания, зафиксированные в словарях.

М. В. Хохловой не только исследовались статистические методы, но также были описаны более сложные модели, которые применялись к уже полученным результатам. При автоматическом выделении сочетаний был применен один из методов машинного обучения, заключающийся в автоматическом подборе коэффициентов для функции логистической регрессии. Была предпринята попытка подбора коэффициентов для оптимального поиска коллокаций, зафиксированных в словарях (работа № 3). Была введена метрика качества ранжирования, которая была основана на данных о рангах сочетаний по разным статистическим мерам.

В качестве материала для описания частотных характеристик существительных и выявления данных о сочетаемости были использованы русскоязычные корпусы, которые получили название «больших корпусов». К ним относятся ruWaC, ruTenTen и Aranea Russicum Maximum. Данные корпусы текстов, объемы которых превышают 100 млн слов, появились относительно недавно. Их возникновение связано с ростом технических возможностей и постепенным уходом от «ручного» составления корпусов, к более автоматическому. В настоящий момент можно говорить о корпусах двух типов. К первому типу относятся корпуса, которые создаются лингвистами по заранее описанной технологии, которую можно назвать «классической». Для подобных корпусов тексты отбираются, размечаются и далее загружаются в корпус. Корпусы второго типа создаются автоматически на основе текстов, полученных из Интернета. Наиболее ранним из них является ruWaC (*англ.* Russian Web as Corpus). Корпусы Aranea Russicum Maximum и ruTenTen также относятся ко второму типу, хотя и построены по иному принципу.

М. В. Хохловой были описаны рассмотренные корпусы текстов, были проанализированы частотные распределения высоко- и низкочастотных слов, было проведено их сравнение с данными частотного словаря русского языка (работа № 4). В ходе экспериментов было установлено, что для высокочастотных существительные результаты, полученные на материале 1 млрд выборки из корпуса ruTenTen и его полной версии, не отличаются, что неверно в случае низкочастотных существительных. В целом по сравнению с другими коллекциями корпус ruWaC показал более близкие результаты с частотным словарем русского языка, созданном на основе Национального корпуса русского языка.

Исследуемые статистические методы могут применяться на разнообразном корпусном материале, в том числе при анализе текстов социальных медиа, которые активно изучаются в последнее время. Определению статистических параметров, а также выявлению ограниченной сочетаемости, по которым можно было бы автоматически идентифицировать иронические и саркастические тексты, посвящена работа № 5.

М. В. Хохлова опубликовала 16 работ в изданиях, индексируемых базами данных Web of Science или Scopus, из них 14 за последние 5 лет. Для участия в конкурсе отобраны 5 работ, посвященных исследованию сочетаемости и частотных характеристик лексических единиц в корпусах текстов русского языка.

В настоящее время наблюдается постепенный переход от простых инструментов, позволяющих работать с корпусными данными, к более сложным системам, дающим доступ к «надстройкам» над этими ресурсами и предоставляющим более широкие возможности пользователям. Описываемый в работах М. В. Хохловой инструментарий статистических мер открывает новые перспективы в обработке корпусных данных, являясь своего рода для них фильтром.

Результаты данной работы могут найти применение в курсах по лексикологии, морфологии и синтаксису современного русского языка, при составлении словарей и грамматик, а также в преподавании русского языка. Также они могут быть использованы при машинном обучении программ, связанных с автоматической обработкой языка, например, в системах автоматической кластеризации словосочетаний или при разрешении неоднозначности.