Премия СПбГУ за научные труды
в номинации «за фундаментальные достижения в науке»

Захаров Виктор Павлович, канд. филол. наук, доцент, доцент кафедры математической лингвистики СПбГУ

Цикл работ на тему «Корпусы русского языка и корпусные исследования»

1. Zakharov V. Corpora of the Russian Language. In: Habernal I., Matoušek V. (eds) Text, Speech, and Dialogue. TSD 2013. Lecture Notes in Computer Science, vol 8082. Springer, Berlin, Heidelberg. Pp. 1–13. DOI: 10.1007/978-3-642-40585-3_1. **(SCOPUS, WoS)**

https://link.springer.com/chapter/10.1007/978-3-642-40585-3_1

2. Захаров В. П. Сочетаемость через призму корпусов // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Москва, 27–30 мая 2015 г.). Вып. 14 (21). М.: Изд-во РГГУ, 2015. Т. 1. С. 667–682. **(SCOPUS)**

http://www.dialog-21.ru/digests/dialog2015/materials/pdf/ZakharovVP.pdf

3. Tao Y., Zakharov V. The Development and Use of Russian-Chinese Parallel Corpus // Automatic Documentation and Mathematical Linguistics, 2015, Vol. 49, No. 2. P. 65–75. DOI: 10.3103/S0005105515020077. **(WoS)**

https://link.springer.com/article/10.3103/S0005105515020077

4. Benko V., Zakharov V. P. Very Large Russian Corpora: New Opportunities and New Challenges // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (Москва, 1–4 июля 2016 г.). Вып. 15 (22). М.: Изд-во РГГУ, 2016. P. 83-98. **(SCOPUS)**

http://www.dialog-21.ru/media/3383/benkovzakharovvp.pdf

5. Khokhlova M., Zakharov V. Efficiency of the sketch grammar for Russian // 3rd International Multidisciplinary Scientific Conference on Social Sciences and Arts SGEM 2016, www.sgemsocial.org, SGEM2016 Conference Proceedings, ISBN 978-619-7105-72-8 / ISSN 2367-5659, 24–31 August, 2016, Book 1 Vol. 3, pp. 279–286. DOI: 10.5593/SGEMSOCIAL2016/B13/S03.037. **(WoS)**

https://sgemworld.at/ssgemlib/spip.php?article2506&lang=en
[нет полного текста онлайн]

6. Mikhailova V., Mochalova A., Mochalov V., Zakharov V. Uncovering semantic relations conveyed by Russian prepositions. In: Proceedings of the 18th International Conference on Advanced Communication Technology (ICACT), 2016, pp. 459-465. DOI: 10.1109/ICACT.2016.7423432. **(SCOPUS, WoS)**

http://ieeexplore.ieee.org/document/7423432/

Аннотация

Корпусная лингвистика — раздел компьютерной лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических корпусов (корпусов текстов) с применением компьютерных технологий. Можно сказать, что все современные лингвистические

исследования и работы по составлению словарей и грамматик так или иначе ориентированы на использование представительных корпусов текстов. Развитие современных интеллектуальных программных систем, предназначенных для обработки текстов на естественном языке, также требует большой экспериментальной лингвистической базы. Поэтому развитие корпусной лингвистики имеет большое значение для отечественной науки.

Заявитель является одним из ведущих специалистов по данной дисциплине в стране. Помимо преподавательской и научной работы, он является организатором единственной в России конференции по корпусной лингвистике, которую проводит СПбГУ.

Заявитель опубликовал 24 работы в изданиях, индексируемых базами данных Web of Science или Scopus, из них 21 за последние 5 лет (по 2017 год включительно). Для участия в конкурсе отобраны 6 работ, посвященных созданию и исследованию корпусов русского языка и автоматическим методам выявления определенных лексических единиц и семантических отношений в русском языке.

В мире корпусная лингвистика как особое направление сложилась к началу 1990-х годов. За прошедшие годы корпусная методология становится частью лингвистической науки, и все лингвисты, работающие в самых разных направлениях, как правило, проводят свои исследования на базе корпусов. Россия встала на этот «корпусный» путь с некоторым опозданием, но движется по нему очень быстро. И заявитель вносит свой посильный вклад в развитие и исследование корпусов русского языка (работы №№ 1, 3, 4). Особо стоит отметить участие в работе по созданию русско-китайского параллельного корпуса (№ 3) и вклад в развитие новой технологии создания мегакорпусов на базе веба (№ 4). Работа № 1 представляет собой публикацию приглашенного доклада на конференции Text, Speech and Dialogue (Czech Republic).

Вторая группа работ представляет собой исследования по автоматическому выявлению различных типов лексических единиц (ключевые слова, устойчивые сочетания, лексико-семантические поля) на основе корпусов. Одним из старых и известных методов лингвистического исследования является дистрибутивно-статистический анализ, при котором используется информация о дистрибуции элементов текста и их числовых параметрах. Уже на заре компьютерной лингвистики предпринимались попытки на основе частотной информации о встречаемости лексических единиц в контекстах определенной величины получать по некоторой заданной формуле количественную характеристику их связанности, что в современной корпусной лингвистике нашло выражение в методах выявления коллокаций и многословных единиц на основе мер ассоциации (№№ 2, 5). Одновременно выдвигались идеи распространения этого метода и на парадигматический аспект языка, идеи о том, что парадигматические связи могут выводиться из связей синтагматических. Однако эти идеи на практике нашли свое воплощение только с созданием больших корпусов текстов. Этой проблематике посвящены как работа № 6, так и другие, не вошедшие в данный перечень. Также очень продуктивны исследования различной направленности на основе многоязычных параллельных корпусов. В нашем блоке публикаций эта проблематика прорабатывается в работе № 3.