гласной договоренностью между создателями корпуса и его пользователями. Разрабатываемый в настоящее время Генеральный интернет-корпус русского языка (ГИКРЯ) задумывается как инструмент, позволяющий эксплицировать подобные договоренности и изучать русский язык в его дифференциальной полноте. Исследователи получат ресурс, позволяющий анализировать отдельные сегменты Интернета и создавать подкорпуса на основе метаразметки, извлекаемой автоматически. В настоящее время в ГИКРЯ размечены и доступны для поиска два сегмента русского Интернета: блог-платформа LiveJournal.com и «Журнальный зал». В дальнейшем количество сегментов планируется существенно расширить.

THE GENERAL INTERNET CORPUS OF RUSSIAN AND THE NOTION OF REPRESENTATIVENESS IN CORPUS LINGUISTICS

Piperski A.C.

Russian State University for the Humanities, Institute of Linguistics, Moscow, Russia
(Miusskaya Sq. 6-2, 125993, Moscow),e-mail: apiperski@gmail.com

The present article deals with the notion of representativeness in corpus linguistics. It turns out that there are no exact methods for assessing representativeness, and for this reason the representativeness of a corpus is nothing more than a tacit agreement between the creators of a corpus and its users. The General Internet Corpus of Russian (GICR) which is presently under development tries to make such an agreement explicit. It encourages its uses to study register variation in the Russian language of the Internet. The linguistic community will be able to use a research tool to study different segments of the Web and to create subcorpora using automatically extracted metadata. As for June 2013, GICR contains two segments of the Russian Web, namely the blog platform LiveJournal.com and the "Magazine Reading Room" (http://magazines.russ.ru/). More segments will be added soon.

ЖАНРОВАЯ КЛАССИФИКАЦИЯ В ГЕНЕРАЛЬНОМ ИНТЕРНЕТ-КОРПУСЕ РУССКОГО ЯЗЫКА

Пиперски А.Ч.

Институт лингвистики ФГБОУ ВПО «Российский государственный гуманитарный университет»,
Москва, Россия (125993, Москва, Миусская пл., 6, корп. 2), e-mail: apiperski@gmail.com

Корпуса представляют собой важнейший инструмент современных лингвистических исследований. Для получения достоверных результатов исследователи, пользующиеся корпусами, должны обращать внимание на параметры метатекстовой разметки (информацию о социолингвистической, региональной, жанровой и т. п. принадлежности текста). В большинстве корпусов метатекстовые данные добавляются вручную, однако это невозможно при разработке больших корпусов, создаваемых на основе текстов из Интернета. Одним из таких корпусов является Генеральный интернет-корпус русского языка (ГИКРЯ), в котором применяются автоматические технологии метатекстовой разметки. В частности, предлагается новая схема жанровой разметки, при которой не выделяются априорные категории, а производится кластеризация на основе значений ряда переменных, выполняемая при помощи машинного обучения.

GENRE CLASSIFICATION IN THE GENERAL INTERNET CORPUS OF RUSSIAN

Piperski A.C.

Russian State University for the Humanities, Institute of Linguistics, Moscow, Russia
(Miusskaya Sq. 6-2, 125993, Moscow), e-mail: apiperski@gmail.com

Corpora are indispensable research tool in present-day linguistics. If a scholar wants to achieve reliable results in a corpus-based study, he should take into account metadata, i.e. sociolinguistic, regional and genre-related properties of the texts included into the corpus. In most corpora metadata are added manually, which is not possible when constructing large Web-based corpora. Since the General Internet Corpus of Russian (GICR) is one of such corpora, it has to use automated metadata tagging. The developers of GICR propose a novel approach to genre classification without postulating any a priori categories. Machine learning algorithms are used to cluster texts based on automatically extractable features.

МАКСИМАЛЬНО ПРАВДОПОДОБНАЯ ОЦЕНКА ДИСПЕРСИОННО-КОВАРИАЦИОННОЙ МАТРИЦЫ

Полянский И.С., Патронов Д.Ю.

Академия ФСО России, Орел

В работе сформировано аналитическое выражение, определяющее максимально правдоподобную оценку дисперсионно-ковариационной матрицы наблюдения вектора случайных величин, распределенных по многомерному нормальному закону. Решение основано на получении выражения, определяющего точку экстремума сформированной на основе распределения Уишарта функции правдоподобия. Полученная оценка дисперси-