УДК 004.912

А.А. Крижановский, А.В. Смирнов, В.М. Круглов,
Н.Б. Крижановская, И.С. Кипяткова

АВТОМАТИЧЕСКОЕ ИЗВЛЕЧЕНИЕ СЛОВАРНЫХ ПОМЕТ ИЗ РУССКОГО ВИКИСЛОВАРЯ

Крижановский А.А., Смирнов А.В., Круглов В.М., Крижановская Н.Б., Кипяткова И.С. **Автоматическое извлечение словарных помет из Русского Викисловаря.**

Аннотация. Разработана методология извлечения словарных помет из интернет-словарей. В соответствие с этой методологией экспертами построено отображение (соответствие один к одному) системы словарных помет Русского Викисловаря (385 помет) и системы словарных помет Английского Викисловаря (1001 помета). Таким образом, построена интегральная система словарных помет (1096 помет), включающая пометы обоих словарей. Разработан синтаксический анализатор (парсер), который распознаёт и извлекает известные и новые словарные пометы, сокращения и пояснения, указанные в начале текста значений слов в словарных статьях Викисловаря. Следует отметить наличие в парсере большого количества словарных помет известных заранее (385 словарных помет для Русского Викисловаря). С помощью парсера на основе данных Русского Викисловаря была построена база данных машиночитаемого Викисловаря, включающая информацию о словарных пометах. В работе приводятся численные параметры словарных помет в Русском Викисловаре, а именно: с помощью разработанной программы было подсчитано, что в базе данных машиночитаемого Викисловаря к 133 тыс. значений слов приписаны пометы и пояснения; для полутора тысяч значений слов был указан регион употребления слова, подсчитано число словарных помет для разных предметных областей. Вкладом данной работы в компьютерную лексикографию является оценка численных параметров словарных помет в больших словарях (пятьсот тысяч словарных статей).

Ключевые слова: вычислительная лингвистика, компьютерная лексикография, русский язык.

Krizhanovsky A.A., Smirnov A.V., Kruglov V. M., Krizhanovskaya N.B., Kipyatkova I.S. **Automatic extraction of context labels from the Russian Wiktionary.**

Abstract. The methodology of extracting context labels from internet dictionaries was developed. In accordance with this methodology experts constructed a mapping table that establishes a correspondence between Russian Wiktionary context labels (385 labels) and English Wiktionary context labels (1001 labels). As a result the composite system of context labels (1096 labels), which includes both dictionary labels, was constructed. The parser extracting context labels from the Russian Wiktionary was developed. The parser can recognize and extract new context labels, abbreviations and comments placed before the definition in Wiktionary articles. One outstanding feature of this parser is a large number of context labels which are known in advance (385 context labels for Russian Wiktionary). The parser can recognize and extract new context labels, abbreviations and comments placed before the definition in Wiktionary articles. The database of machine-readable Russian Wiktionary including context labels was generated by the parser. An evaluation of numerical parameters of context labels in the Russian Wiktionary was performed. With the help of the developed computer program it was found in the Russian Wiktionary that (1) there are 133 000 definitions with context labels and comments, (2) one and a half thousand definitions were supplied with regional labels, (3) it was calculated a number of definitions with labels for each domain knowledge. This paper is an original contribution to computational lexicography, setting out for the first time an analysis of numerical parameters of context labels in the large dictionary (500 000 entries).

Keywords: computational linguistics, computational lexicology, Russian language.

1. Введение. Викисловарь (http://ru.wiktionary.org/) — это свободно пополняемый многофункциональный многоязычный онлайн-

словарь и тезаурус. В Викисловаре содержатся толкования и переводы слов, описание фонетических и морфологических свойств, семантические (парадигматические) отношения. В словарных статьях приводится произношение слов (указана транскрипция и даны ссылки на аудиофайлы с произношением), правила разбиения слов на слоги, ударения в словах, информация об этимологии слов.

Для использования данных викисловарей при решении задач автоматической обработки текста и речи необходимо преобразовывать тексты словарных статей — слабоструктурированные данные — в машиночитаемый формат. Построение машиночитаемого словаря на основе данных викисловарей является комплексной задачей, рассчитанной не на один десяток лет. В настоящее время созданный синтаксический анализатор (парсер) позволяет извлекать из Английского и Русского Викисловарей следующую информацию: языковую и частеречную принадлежность лексических единиц, дефиниции значений и отдельные семантические компоненты, иллюстративный материал (цитаты и речения — пока только из Русского Викисловаря), а также иноязычные соответствия (переводы значений слов). Здесь и далее название конкретного проекта (Английский Викисловарь, Русский Викисловарь) пишется с заглавной буквы, название вообще словарей данного типа, т.е. викисловарей, пишется с маленькой буквы.

В этой публикации представлены результаты очередного этапа работ по проектированию и разработке синтаксического анализатора. Речь идет об извлечении из Русского Викисловаря лексикографических помет разного типа.

Словарные пометы — это используемые в тексте словарной статьи краткие указания на характеристики описываемой лексемы — грамматические и стилистические. Грамматические пометы указывают на частеречную принадлежность слова, дают информацию о грамматической форме и грамматическом значении. Стилистические пометы заслуживают особого внимания, так как включают в себя несколько подтипов [1, 2, 3, 4].

Во-первых, это пометы, указывающие на стилистическую ограниченность употребления слов, относящихся к литературному языку (*офиц., офиц.-дел., разг., книжн., высок., трад.-поэт., народно-поэт.*). Так, например, помета *офиц.* (официальное) указывает на то, что слово (или значение), имеющее эту помету, характерно для официальных текстов разного характера, а помета *разг.* (разговорное) указывает на то, что слово (или значение) употребляется в живой, непринужденной, преимущественно устной, речи.

Во-вторых, это пометы, указывающие специальную область применения слова — пометы предметных областей (*гидрол., горн., минер.* и т. д.).

В-третьих, это пометы, указывающие на принадлежность слова, находящегося на границе или за пределами литературного языка, к различным пластам лексики (*обл., прост., груб.-прост.*). Так, например, помета *прост.* указывает на то, что слово (или значение) из-за грубости содержания или резкости выражаемой оценки стоит на границе литературного языка и употребляется в сниженном стиле, в обиходной, бытовой речи.

Кроме названных трёх типов помет, характеризующих прагматическую часть лексического значения слова, выделяют еще несколько:

- экспрессивные пометы, указывающие на эмоциональную окраску слова (*пренебр., неодобр.*); экспрессивные пометы далее будут описаны более подробно;
- хронологические пометы, указывающие на временные рамки слов (например, слова, выходящие из употребления, и слова, появляющиеся в языке, обозначаются пометами *устар.* и *неол.*);
- региональные (областные) пометы, указывающие на территориальные рамки слов.

Успешное решение задачи извлечения словарных помет и создания машиночитаемого Викисловаря, включающего пометы, востребовано в различных направлениях вычислительной лингвистики. Например, для определения тематики текста необходим словарь с пометами предметных областей, для анализа тональности текста нужен словарь с экспрессивными пометами (разрабатываемый машиночитаемый Викисловарь может служить основой для создания тональных словарей [5]).

Статья имеет следующую структуру. Во второй главе рассматриваются машиночитаемые словари со словарными пометами. Особенности словарных помет викисловарей указаны в третьей главе. В четвёртой главе описана методология извлечения из викисловарей словарных помет. В пятой главе обсуждаются экспрессивные словарные пометы, и приводится «выровненный» список экспрессивных помет двух викисловарей. В шестой главе представлены численные результаты автоматического извлечения словарных помет из Русского Викисловаря.

2. Состояние дел в проблемной области. Построенный машиночитаемый словарь (http://code.google.com/p/wikokit) представляет интерес благодаря большому объёму данных и наличию в нём распо-

знанных словарных помет. Наиболее близким аналогом среди электронных словарей, включающих словарные пометы и находящихся в открытом доступе, является WordNet Domains. Различие состоит в том, что тезаурус WordNet Domains [6] содержит всего 170 помет, и эти пометы проставлены автоматически в отличие от помет в Викисловаре, где они указываются редакторами словаря вручную.

При решении задачи автоматического построения словарей специального типа, в которых к значениям слов приписаны словарные пометы, не обойтись без лексикографических онлайн-ресурсов, в первую очередь – викисловарей. Кроме словарных входов с толкованиями, семантическими отношениями и переводами, викисловарь содержит словарные пометы. Викисловарь превосходит WordNet Domains по следующим параметрам:

- большой объем и быстрое обновление материала (последнее особенно актуально для неологизмов);
- большое количество редакторов (сотни редакторов в Английском Викисловаре, десятки — в Русском Викисловаре), по всей видимости, обеспечивает более адекватное отражение языковой реальности (противоположностью являются словари, основанные на языковых представлениях небольшой группы лексикографов);
- как и другие проекты Web 2.0, викисловарь не может гарантировать качества данных (как замечают Мейер и Гуревич в работе [7]: “Wiktionary has as yet no reviewing or releasing workflow”), тем не менее внедряются механизмы, направленные на повышение качества словарных статей (см. http://ru.wiktionary.org/Викисловарь:Проверка_страниц).

Существуют свободно распространяемые инструменты [8, 9], которые позволяют напрямую работать с викисловарем как с базой данных. В последнее время появляются проекты, которые ставят целью сравнить [7, 10] и интегрировать [11, 12, 13, 14] ворднеты и викисловари.

Обзор доступной литературы подсказывает, что данная статья является первой, где выполнен комплексный анализ словарных помет в машиночитаемом словаре большого объёма (пятьсот тысяч словарных статей).

3. Особенности словарных помет викисловарей. В среде редакторов Русского Викисловаря для обозначения словарных помет используется термин «условные сокращения» (в Английском Викисловаре – “Context labels”). Полный список условных сокращений Рус-

ского Викисловаря доступен онлайн (см. https://ru.wiktionary.org/wiki/Викисловарь:Условные_сокращения). Задача разрабатываемого синтаксического анализатора состоит в распознавании и извлечении условных сокращений и различных пояснений и комментариев при значениях слов.

Чтобы более чётко очертить решаемую задачу извлечения словарных помет, укажем на особенности помет Русского Викисловаря:

1. *Не только словарные пометы*. При значениях слов кроме словарных помет могут быть указаны различные *пояснения и комментарии* с помощью специального шаблона «помета» (см. фрагмент словарной статьи «улыбнуться» в следующем разделе). Эта информация относится к полю «словарные пометы и пояснения» и тоже должна извлекаться наравне с обычными словарными пометами.

2. *Открытый список*. В традиционных бумажных словарях список словарных помет представляет собой некоторую данность, которая уже не меняется после публикации словаря. В этом отношении викисловари можно отнести к регулярно издаваемым словарям, изменение или расширение списка помет которого обсуждается и принимается редакторами викисловаря в рабочем порядке.

3. *Большой объём и глубокая детализация*. Викисловари создаются большими коллективами редакторов, каждый из которых является специалистом в какой-либо области знания. Для отнесения слов и терминов к этим областям вводятся всё новые пометы предметных областей. Например, в Английском Викисловаре на 2014 год около 370 помет предметных областей.

4. Существуют следующие зоны словарной статьи в Русском Викисловаре, где могут быть указаны пометы: (1) толкование, (2) семантические отношения, (3) переводы. В этой работе исследуются словарные пометы только в зоне толкования.

5. Редакторами викисловарей разработана система категорий словарных помет, призванная упорядочить и систематизировать словарные пометы, каждая помета относится к одной из категорий (рис. 1).

Наиболее детально на данный момент в Английском Викисловаре проработаны категории помет предметных областей (topical на рис. 1). Именно система категорий Английского Викисловаря была взята за основу при создании категорий помет машиночитаемого Викисловаря.

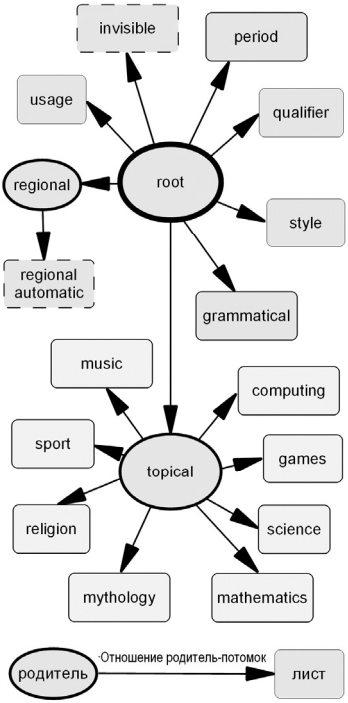

Рис. 1. Система категорий словарных помет в наиболее полном Английском Викисловаре

Пунктиром на рисунке 1 выделены две специальные группы помет машиночитаемого Викисловаря:

• *invisible* — к этой категории отнесены специальные шаблоны Русского Викисловаря, которые (на основе параметров) формируют толкование (например, шаблоны {{as ru}}, {{=}}, {{свойство}}, {{состояние}}, {{соотн.}}, {{действие}}, {{совершить}}); конструкция в двойных фигурных скобках называется шаблоном. Чтобы ознакомиться с работой шаблона, нужно набрать в адресной строке https://ru.wiktionary.org/wiki/Шаблон:название_шаблона , где в качестве параметра «название_шаблона» можно указать «свойство», «состояние» и т.д.

• *regional automatic* — региональные пометы задаются в виде параметра у шаблона {{рег.|}}, например, {{рег.|пск.}}, {{рег.|твер.}}, {{рег.|смол.}}. Интересная особенность этой группы помет в том, что хотя в этом шаблоне пометы задаются свободным текстом, но распознаются как пометы, принадлежащие конкретной категории — *регио-*

нальной. Все прочие пометы, задаваемые произвольным текстом с помощью шаблона {{помета|}}, имеют в машиночитаемом Викисловаре пустое значение в поле «категория пометы».

4. Методология извлечения помет. Рассмотрим словарную статью «улыбнуться» из Русского Викисловаря (см. https://ru.wiktionary.org/wiki/улыбнуться), а именно фрагмент статьи с толкованиями, содержащими словарные пометы и пояснения:

4. *перен.*, *разг.*, *часто с частицей «не»* повезти, достаться, стать предметом обладания ◆ Но и второе место бразильцу не **улыбнулось** — за шестнадцать кругов до финиша его постигла та же участь, что и Райкконена. *Борис Мурадов, «Гран При Малайзии. Время разбрасывать крылья» // «Формула», 2002 г.*

5. *перен.*, *разг.*, *устар.* не достаться кому-либо; утратиться, исчезнуть, пропасть ◆ — Боюсь я, как бы урока мне не лишиться… ученица моя поговаривает, что отец её совсем из Петербурга хочет уехать. Пожалуй, двадцать-то пять рублей в месяц и **улыбнутся**. *М. Е. Салтыков-Щедрин, «Мелочи жизни», 1886—1887 г.*

В приведённом выше фрагменте видно, что условные сокращения предшествуют тексту, описывающему значение слова. В четвёртом значении слова указаны две пометы (*перен., разг.*) и одно пояснение (*о частом употреблении с частицей «не»*). В пятом значении указаны три пометы (*перен., разг., устар.*). Повторим, что в задачу парсера входит извлечение из этого текста всех перечисленных помет и пояснений.

Сформулируем предпосылки, определившие методологию извлечения словарных помет:

- разрабатываемый синтаксический анализатор в перспективе должен уметь извлекать данные не только из Русского Викисловаря, но и Английского Викисловаря;
- база данных машиночитаемого Викисловаря должна автоматически наполняться данными, извлекаемыми парсером из Викисловаря;
- структура базы данных машиночитаемого Викисловаря должна быть одинаковой для разных исходных викисловарей (это обеспечит унификацию полученных данных);
- несколько сотен словарных помет Русского Викисловаря и около тысячи словарных помет Английского Викисловаря известны заранее (например, пометы *перен., разг., устар.* из примера выше), этот список составлен редакторами Викисловаря и доступен онлайн;
- существует открытый список помет и пояснений; объём этого списка (до построения машиночитаемого Викисловаря) неизвес-

тен (например, пояснение «часто с частицей “не”» из примера выше относится к этому списку).

С учётом этих предпосылок и с учётом особенностей словарных помет викисловарей (см. предыдущий раздел) методология извлечения словарных помет включает следующие этапы:

1. соотнесение помет Английского Викисловаря и Русского Викисловаря;
2. добавление информации о пометах (и их соответствия в разных викисловарях) и их категориях в базу знаний парсера;
3. разработка синтаксического анализатора;
4. автоматическое создание базы данных машиночитаемого Викисловаря с помощью синтаксического анализатора;
5. публикация списка извлечённых помет, обсуждение результатов с редакторами викисловаря;
6. исправление найденных ошибок в статьях викисловаря;
7. расширение/изменение списка словарных помет в базе знаний синтаксического анализатора, исправление ошибок в самом анализаторе;
8. переход к шагу 4.

На первом этапе экспертами построено отображение (соответствие один к одному) системы словарных помет Русского Викисловаря (385 помет) и системы словарных помет Английского Викисловаря (1001 помета). В таблице 1 представлен пример такого «выравнивания»: для четырёх помет указаны пары соответствий, для трёх помет Английского Викисловаря не оказалось пары в Русском Викисловаре. Интегральная система словарных помет, включающая пометы обоих словарей, содержит 1096 словарных помет.

При соотнесении помет и при указании категорий (на первом и втором этапе) были выявлены неточности и ошибки в самом словаре, для решения которых привлекались редакторы Английского Викисловаря. В случае трудностей при «выравнивании» помет разных словарей вопросы также выносились на обсуждение с редакторами.

Синонимами словарных помет будем называть такие пары помет викисловаря, которые обозначают одно и то же, например, «ед. ч.» и «ед.», «мн. ч.» и «мн», «собир.» и «собират.» и др. Полный список словарных помет и их синонимов доступен онлайн для Русского Викисловаря (см. http://code.google.com/p/wikokit/source/browse/trunk/common_wiki/src/wikokit/base/wikt/multi/ru/name/LabelRu.java) и Английского Викисловаря (см. http://code.google.com/p/wikokit/source/browse/trunk/common_wiki/src/wikokit/base/wikt/multi/en/name/LabelEn.java). Разным синонимичным словарным пометам, извлечённым из

словаря, в машиночитаемом Викисловаре ставится в соответствие ровно одна помета — та, которая официально указана на странице «Условных сокращений» Русского Викисловаря. Таким образом, разнобой и разнообразие исходных слабоструктурированных данных викисловарей приводятся к единому формализованному виду, представленному в машиночитаемом Викисловаре.

Таблица 1. Фрагмент сопоставления словарных помет Русского Викисловаря и Английского Викисловаря

Английский Викисловарь	Русский Викисловарь		
	краткая	полная форма	примеры словарных статей
beaurocratic	канц.	канцелярское	быть на замечании, взимать
formal	офиц.	официальное	адресант, доложить
high-register	высок.	высокое	борение, изрекать
literary	книжн.	книжное	дефиниция, малодушие
hypercorrect	—	—	—
hyperbolic	—	—	—
loosely	—	—	—

На пятом шаге извлечённые словарные пометы и различная дополнительная информация (например, число помет разных категорий) опубликованы на странице Русского Викисловаря (см. https://ru.wiktionary.org/wiki/User:AKA_MBG/Статистика:Пометы). Такая публикация результатов, во-первых, позволяет редакторам найти ошибки в словарных статьях (например, опечатки в написании словарных помет) и исправить их (шестой шаг). Во-вторых, позволяет увидеть высокочастотные пометы, которые ещё не внесены в базу знаний парсера (см. таблицу "*Labels found by parser*" на той же странице, для сортировки помет по частоте употреблений следует щёлкнуть мышкой по колонке "Counter"). Добавление этих помет в базу парсера будет седьмым шагом.

Таким образом, данная методология описывает итеративный процесс, который приближает нас к цели — точному и максимально полному представлению данных викисловаря в машиночитаемой форме.

5. Экспрессивные пометы. В связи с актуальностью задачи «анализ тональности текста» (англ. sentiment analysis) особое значение приобретает наличие в словаре экспрессивных помет. Пометы, указывающие на эмоциональную окраску слова, являются важным дополнительным источником информации при автоматическом анализе текстов в современных рекомендующих системах, учитывающих мнения (групп) независимых пользователей — участников социальных се-

тей [15]. В таблице 2 представлено 26 экспрессивных помет Русского и Английского Викисловарей, выровненных относительно друг друга.

Таблица 2. Сопоставление помет Русского Викисловаря и Английского Викисловаря, указывающих на эмоциональную окраску слова (экспрессивные пометы)

Английский Викисловарь	Русский Викисловарь	
	краткая	полная форма
abusive	бранн.	бранное
acerbity	груб.	грубое
augmentative	увелич.	увеличительное
contemptuous	презр.	презрительное
corroborative	усилит.	усилительное
derogatory	унич.	уничижительное
diminutive	уменьш.	уменьшительное
diminutive hypocoristic	умласк.	уменьшительно-ласкательное
dysphemism	дисфм.	дисфемизм
elevated	высок.	высокое
endearing	ласк.	ласкательное
euphemistic	эвф.	эвфемизм
expressive	экспр.	экспрессивное
familiar	фам.	фамильярное
humorously	шутл.	шутливое
low style	сниж.	сниженное
obscene language	мат	матерное
offensive	—	—
pejorative	неодобр.	неодобрительное
reproach	укор.	укорительное
rhetoric	ритор.	риторическое
sarcastic	ирон.	ироничное
scornful	пренебр.	пренебрежительное
solemn	торж.	торжественное
tabooed	табу	табуированное
vulgar	вульг.	вульгарное

Для экспрессивной пометы “offensive” Английского Викисловаря не было найдено соответствия в Русском Викисловаре, поэтому в таблице стоит прочерк. Для всех словарных помет Русского Викисловаря соответствия в английской редакции Викисловаря были найдены.

6. Результаты автоматического извлечения словарных помет из Русского Викисловаря. Синтаксический анализатор расширен модулем для автоматического извлечения словарных помет из Русского Викисловаря. В ходе работы данного модуля по данным на август

2013 г. было обработано 482 тыс. статей Русского Викисловаря и была извлечена следующая информация (см. http://ru.wiktionary.org/wiki/User:AKA_MBG/Статистика:Пометы):

- всего уникальных сокращений, помет и пояснений 3316, из них в синтаксический анализатор добавлено вручную 385 помет, а 2931 сокращений, помет и пояснений найдено парсером автоматически;
- к 133 тыс. значений слов в словаре приписаны пометы и пояснения;
- в десятку наиболее употребимых условных сокращений вошли 18662 употреблений «прич.» (причастие), 11129 «перен.» (переносное значение), 10995 «разг.» (разговорное), 7085 «зоол.» (зоологическое), 4727 «=» (обозначение полной синонимии), 4687 «устар.» (устаревшее), 4413 «анат.» (анатомическое), 3991 «ботан.» (ботаническое), 3172 «хим.» (химическое), 2715 «мед.» (медицинское).
- автоматически выделены региональные (областные) пометы у значений слов. Всего найдено 109 уникальных региональных помет. Самые частотные пометы: «амер.» — американское (321 употр.), «брит.» — британское (112 употр.). Всего в Русском Викисловаре выявлено полторы тысячи значений слов (1661), для которых указан регион употребления слова.

Был выполнен подсчёт числа значений в словаре, помеченных словарными пометами разных категорий (рис. 1). Полученное число словарных помет разных категорий представлено на рисунке 2. Большую часть словарных помет (41%) составляют пометы, которые указывают на какую-либо предметную область (в словаре около 54 тыс. значений слов с такими пометами).

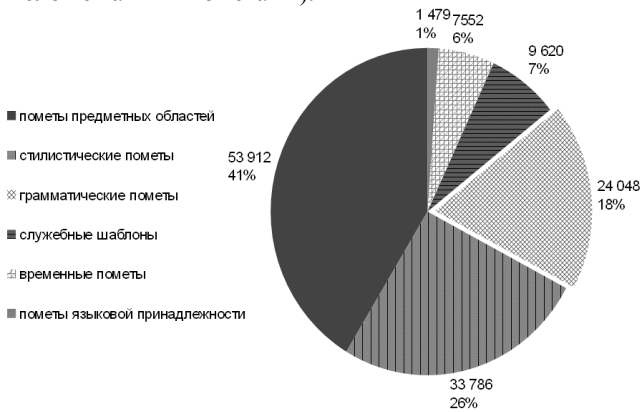

Рис. 2. Количество разных категорий словарных помет в Русском Викисловаре

Среди помет предметных областей больше всего используются научные пометы (18 тыс. употреблений, 33%), указывающие на какую-либо научную дисциплину (рис. 3). Наиболее полно представлена научная терминология по следующим предметам (больше тысячи терминов): зоология (7086 терминов), ботаника (3983), химия (3119), биология (1322), физика (1201). Значений слов с математическими пометами всего 1637 (рис. 3), сюда входят две пометы: «матем.» (математическое) и «геометр.» (геометрическое).

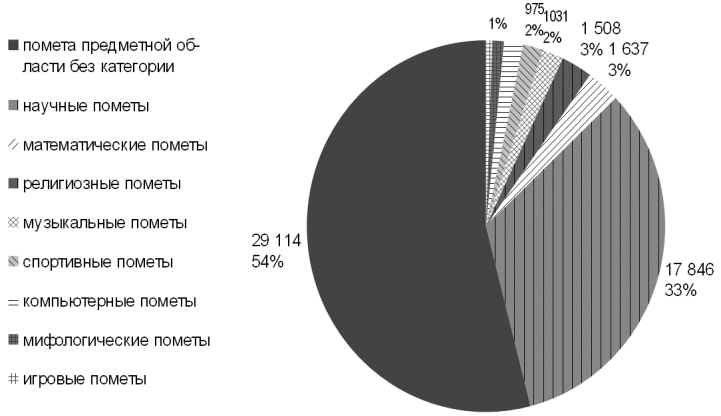

Рис. 3. Количество словарных помет разных предметных областей в Русском Викисловаре

На основе данных машиночитаемого Викисловаря можно получить число словарных статей с пометами и без них для любого из сотен языков Русского Викисловаря. В таблице 3 представлены эти данные для русских слов.

Таблица 3. Число словарных статей, содержащих словарные пометы, о русских словах в Русском Викисловаре (для разных частей речи)

Часть речи	Число статей			
	всего статей с непустыми значениями	без помет	с пометами	
существительное	46780	25999	20781	44,4%
глагол	11353	3913	7440	65,5%
прилагательное	7991	3552	4439	55,6%
наречие	2036	431	1605	78,8%
Всего	68160	33895	34265	50,3%

Из этой таблицы видно, что примерно половина (50.3%) всех словарных статей о русских словах содержит словарные пометы. Несколько большая доля статей с пометами у наречий (78.8%) и глаголов (65.5%). Меньше всего словарных помет у существительных (44.4%), которые тем не менее по количеству (21 тыс.) превышают все остальные вместе взятые части речи. Более детальный анализ числа статей со словарными пометами представлен на следующем рисунке 4.

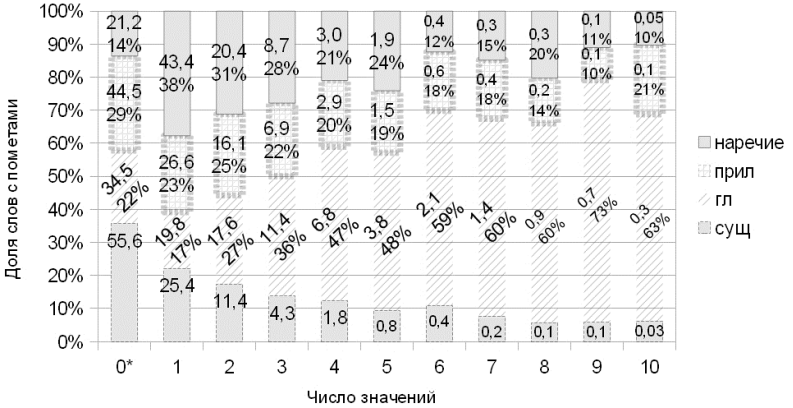

Рис. 4. Распределение числа словарных статей о русских словах с пометами по частям речи (существительное, глагол, прилагательное, наречие) и по числу значений на основе данных Русского Викисловаря

По оси абсцисс отложено число значений в словарной статье — от единицы до десяти. Встречаются статьи, у которых существенно больше значений (например, у глагола «тянуть» — 30 значений), но таких статей мало относительно их общего числа (всего 386 слов Русского Викисловаря (или 0.14%) имеют больше 10 значений). В нулевой столбец *0** попали статьи, в которых может быть разное ненулевое число значений (от 1 до 10), но нет ни одной словарной пометы. В первом столбце подсчитаны словарные статьи, состоящие из одного значения, помеченного словарной пометой. Во втором столбце подсчитаны словарные статьи, содержащие два значения, при этом хотя бы одно значение (или оба сразу) содержат словарные пометы. И так далее.

Каждой части речи соответствует своя горизонтальная полоска прямоугольников, самая нижняя — существительные, вторая снизу (диагонально заштрихованная область, числа наклонные) — глаго-

лы, третья снизу (область обведена пунктиром) — прилагательные, самая верхняя (обведена тонкой линией) — наречия. Первые числа внутри столбцов указывают на пропорцию словарных статей с указанных числом значений для данной части речи. Например, из рисунка 4 видно, что существительных с одним значением, содержащим словарную помету, 25.4% от числа всех русских существительных с непустыми значениями (к сожалению, приходится акцентировать внимание на «непустых значениях», т.к. из 159 тыс. русских словарных статей в Русском Викисловаре 77 тыс. являются «заготовками», т.е. раздел с толкованиями пуст). На рисунке прекрасно видна ожидаемая закономерность, а именно: уменьшение пропорции всех четырёх частей речи при увеличении числа значений в словарных статьях.

Второе число со знаком процента внутри столбцов (назовём эти числа «доли долей») соответствует высоте столбцов и указывает долю слов данной части речи относительно долей других частей речи для словарных статей с заданным числом значений. Это второе число не указано для существительных (нижний ряд), т.к. его легко видеть с помощью значений оси ординат.

Анализ этой «доли долей» на рис**.** показывает интересные закономерности при росте числа значений (и, соответственно, уменьшении абсолютного числа слов):

- доля существительных и наречий постепенно снижается;
- доля прилагательных (второй ряд сверху) относительно долей других частей речи сохраняет стабильную пропорцию (14-25%);
- доля глаголов растёт от 20% до 60-70%.

Таким образом, число словарных статей о глаголах с большим количеством значений, содержащих словарные пометы, преобладает над числом многозначных статей (с пометами) о других частях речи. Эти результаты согласуются с предыдущими (см. рисунки 5 и 6 в работе [10]), где было подсчитано среднее число значений у многозначных слов для разных частей речи.

7. Заключение. Получены следующие основные результаты. Была разработана методология извлечения словарных помет из интернет-словарей. Построено соответствие между словарными пометами Русского Викисловаря (385 помет) и пометами Английского Викисловаря (более тысячи помет), что является необходимым этапом в решении задачи автоматической интеграции сверхбольших словарей (в обоих словарях суммарно более четырёх миллионов словарных статей).

Структура базы данных машиночитаемого словаря расширена таблицами для хранения данных о словарных пометах. Синтаксиче-

ский анализатор расширен модулем для автоматического извлечения словарных помет из Русского Викисловаря. С помощью синтаксического анализатора на основе данных Русского Викисловаря автоматически построена база данных машиночитаемого Викисловаря, включающая словарные пометы.

Спроектирована и разработана компьютерная система автоматического извлечения и учёта словарных помет онлайн-словарей на примере Русского Викисловаря. С помощью разработанной компьютерной программы было подсчитано, что в базе данных машиночитаемого Викисловаря к 133 тыс. значений слов в словаре приписаны пометы и пояснения. В Русском Викисловаре выявлено более полутора тысяч значений слов (1661), для которых указан регион употребления слова.

Наличие словарных помет в построенном машиночитаемом Викисловаре позволит автоматически определять тематику текста. Например, пометы «зоол.» (зоологическое), «анат.» (анатомическое), «ботан.» (ботаническое), «хим.» (химическое) и т.д. позволяют определить, к какой области знаний принадлежат научные термины в тексте. При этом необходимо будет особое внимание обращать на те многозначные слова, значения которых принадлежат разным предметным областям, например, «валентность» (*лингв.*, *хим.*), «ветвление» (*ботан.*, *прогр.*), «парадигма» (*филос.*, *лингв.*).

С одной стороны, и по числу словарных статей, и по числу редакторов, и по другим параметрам Английский Викисловарь превосходит Русский Викисловарь, исследуемый в данной работе, примерно в 6-7 раз. С другой стороны, и по структуре словарной статьи, и по организации системы словарных помет между этими словарями есть много общего. Поэтому Русский Викисловарь является некоторым тестовым полигоном для последующей оценки различных численных параметров системы словарных помет многоязычного Английского Викисловаря и сравнения между собой этих двух викисловарей. Следующим этапом работ будет расширение синтаксического анализатора модулем для автоматического извлечения словарных помет из Английского Викисловаря.

Исходный код парсера и база данных машиночитаемого Викисловаря доступны с открытой лицензией на сайте проекта (http://code.google.com/p/wikokit/). Все, кто заинтересован в решении лингвистических задач и автоматической обработке текста с помощью машиночитаемого Викисловаря, могут рассчитать на помощь авторов в работе с ним.

Благодарности. Мы благодарим сотрудников Института прикладных математических исследований КарНЦ РАН Огийко А. А., Чиркову Ю. В., Чиркова А. В., Головина А. С., Румянцева А. С. за работу по наполнению базы данных словарными пометами (обеспечившую возможность данного исследования) и за трудоёмкое «выравнивание» словарных помет Русского Викисловаря и Английского Викисловаря.

Литература

1. *Скляревская Г.Н.* Еще раз о проблемах лексикографической стилистики // Вопросы языкознания. 1988. № 3. С. 84-97.
2. Словарь русского языка. Том I. А-Й. // М., 1981. С. 9.
3. *Сорокин Ю.С.* О нормативно-стилистическом словаре современного русского языка // Вопросы языкознания. 1967. № 5. С. 22-32.
4. *Токарчук И.Н.* Стилистические параметры в лексикографическом описании служебного слова (на примере частиц) // Вестник ТГПУ. 2012. Т. 1. №. 116. С. 187-191. URL: http://vestnik.tspu.edu.ru/files/PDF/articles/tokarchuk_i._n._187_191_1_116_2012.pdf (дата обращения 2.04.14).
5. *Пазельская А.Г., Соловьев А.Н.* Метод определения эмоций в текстах на русском языке // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 25 – 29 мая 2011 г.). М.: Изд-во РГГУ, 2011. Вып. 10 (17). С. 510-522. URL: http://www.dialog-21.ru/digests/dialog2011/materials/en/pdf/50.pdf (дата обращения 2.04.14).
6. *Gonzalez-Agirre A., Castillo M., Rigau G.* A graph-based method to improve WordNet Domains // In Proceedings of 13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING'12). New Delhi, India. 2012. URL: http://adimen.si.ehu.es/~rigau/publications/cicling12-gcr.pdf (дата обращения 2.04.14).
7. *Meyer Ch.M., Gurevych I.* Wiktionary: A new rival for expert-built lexicons? // там в шаблонеExploring the possibilities of collaborative lexicography. Chapter 13, in Sylviane G., Paquot M. (eds.), *Electronic Lexicography*, Oxford University Press, Oxford, 2012. pp. 259 – 291.
8. *Zesch T., Müller Ch., Gurevych I.* Extracting lexical semantic knowledge from Wikipedia and Wiktionary // In Proceedings of the Conference on Language Resources and Evaluation (LREC). 2008. vol. 15.
9. *Крижановский А. А.* Преобразование структуры словарной статьи Викисловаря в таблицы и отношения реляционной базы данных. Препринт. 2010. URL: http://scipeople.com/publication/100231/ (дата обращения 2.04.14).
10. *Смирнов А. В., Круглов В. М., Крижановский А. А., Луговая Н. Б., Карпов А. А., Кипяткова И. С.* Количественный анализ лексики русского WordNet и викисловарей // Труды СПИИРАН. 2012. Вып. 23. С. 231–253. URL: http://scipeople.com/publication/113406/ (дата обращения 2.04.14).
11. *Henrich V., Hinrichs E., Vodolazova T.* Semi-Automatic extension of GermaNet with sense definitions from Wiktionary // In Proceedings of 5th Language & Technology Conference (LTC 2011). Poznan, Poland, 2011. pp. 126-130, URL: http:// www.sfs.uni-tuebingen.de/lsd/documents/publications/Henrich-et-al-2011_GermaNet-Wiktionary-Mapping.pdf (дата обращения 23.02.14).

12. *McCrae J., Montiel-Ponsoda E., Cimiano Ph*. Integrating WordNet and Wiktionary with lemon // In Conference Proceedings “Linked Data in Linguistics”, 2012. pp. 25 - 34.
13. *Meyer Ch.M., Gurevych I.* What psycholinguists know about chemistry: Aligning Wiktionary and wordnet for increased domain coverage // In Proceedings of the 5th international joint conference on natural language processing (IJCNLP), Chiang Mai, Thailand, 2011. pp. 883–892.
14. *Navigli R, Ponzetto S.P.* BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network // Artificial Intelligence, 2012. vol. 193, pp 217-250.
15. *Adomavicius G., Mobasher B., Ricci F., Tuzhilin A*. Context-aware recommender systems // AI Magazine. 2011. vol. 32(3), pp. 67-80. URL: http://www.ise.bgu.ac.il/faculty/liorr/recsyshb/chcontext.pdf (дата обращения 2.04.14).

References

1. Sklyarevskaya G.N. [Once again about the problems of lexicographic style]. *Voprosyi yazyikoznaniya – Questions of Linguistics*. 1988. no 3. pp. 84-97. (In Russ.).
2. *Slovar' russkogo jazyka* [Russian dictionary]. vol I. A-J. M., 1981. P. 9. (In Russ.).
3. Sorokin Yu.S. [About normative stylistic dictionary of modern Russian]. *Voprosyi yazyikoznaniya – Questions of Linguistics*. 1967. no 5. pp. 22-32. (In Russ.).
4. Tokarchuk I.N. [Stylistic parameters in lexicographic description function word (particles)]. *Vestnik TGPU – TSPU Bulletin*. 2012. vol. 1. no. 116. pp. 187-191. Available at: http://vestnik.tspu.edu.ru/files/PDF/articles/tokarchuk_i._n._187_191_1_116_2012.pdf (In Russ.).
5. Pazelskaya A.G., Solovev A.N. [Sentiment analysis method for Russian texts]. *Kompyuternaya lingvistika i intellektualnyie tehnologii: Po materialam ezhegodnoy Mezhdunarodnoy konferentsii «Dialog»* [Conference Proceedings "Computational Linguistics and Intellectual Technologies"], 2011. pp. 510-522. http://www.dialog-21.ru/digests/dialog2011/materials/en/pdf/50.pdf (In Russ.).
6. Gonzalez-Agirre A., Castillo M. and Rigau G. A graph-based method to improve WordNet Domains. In Proceedings of 13th International Conference on Intelli-gent Text Processing and Computational Linguistics (CICLING'12). New Delhi, India. 2012. Available at: http://adimen.si.ehu.es/~rigau/publications/cicling12-gcr.pdf (accessed: 2.04.14).
7. Meyer Ch.M., Gurevych I. Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. Chapter 13, in Sylviane G., Paquot M. (eds.), Electronic Lexicography, Oxford University Press, Oxford. 2012. pp. 259–291.
8. Zesch T., Muller Ch., Gurevych I. Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In Proceedings of the Conference on Language Re-sources and Evaluation (LREC). 2008. vol. 15.
9. Krizhanovsky A. A. Transformation of Wiktionary entry structure into tables and relations in a relational database schema. Preprint. 2010. Available at: http://arxiv.org/abs/1011.1368 (accessed: 2.04.14).
10. Smirnov A., Kruglov V., Krizhanovsky A., Lugovaya N., Karpov A., Kipyatkova I. [A quantitative analysis of the Russian lexicon in Russian WordNet and Wiktionaries]. *Trudy SPIIRAN – SPIIRAS Proceedings*. 2012. Issue 23. pp. 231–253. Available at: http://scipeople.com/publication/113406/ (accessed: 2.04.14). (In Russ.).

11. Henrich V., Hinrichs E., Vodolazova T. Semi-Automatic extension of GermaNet with sense definitions from Wiktionary. In Proceedings of 5th Language & Technology Conference (LTC 2011). Poznan, Poland, 2011 pp. 126-130. Available at: http://www.sfs.uni-tuebingen.de/lsd/documents/publications/Henrich-et-al-2011_GermaNet-Wiktionary-Mapping.pdf (accessed: 23.02.2014)
12. McCrae J., Montiel-Ponsoda E., Cimiano Ph. Integrating WordNet and Wiktionary with lemon. In Conference Proceedings "Linked Data in Linguistics", 2012. pp. 25-34.
13. Meyer Ch.M., Gurevych, I. What psycholinguists know about chemistry: Aligning Wiktionary and wordnet for increased domain coverage. In Proceedings of the 5th international joint conference on natural language processing (IJCNLP), Chiang Mai, Thailand. 2011. pp. 883–892.
14. Navigli R., Ponzetto S.P. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence, 2012. vol 193. pp 217-250.
15. Adomavicius G., Mobasher B., Ricci F., Tuzhilin A. Context-aware recommender systems. AI Magazine. 2011. vol. 32(3), pp. 67-80. Available at: http://www.ise.bgu.ac.il/faculty/liorr/recsyshb/chcontext.pdf (accessed: 2.04.14).

Крижановский Андрей Анатольевич — к-т техн. наук, старший научный сотрудник лаборатории интегрированных систем автоматизации Федерального государственного бюджетного учреждения науки Санкт-Петербургского института информатики и автоматизации Российской академии наук (СПИИРАН), старший научный сотрудник лаборатории информационных компьютерных технологий Федерального государственного бюджетного учреждения науки Института прикладных математических исследований Карельского научного центра Российской академии наук (ИПМИ КарНЦ РАН). Область научных интересов: автоматическая обработка текста, корпусная лингвистика. Число научных публикаций — 67. andrew.krizhanovsky@gmail.com, code.google.com/p/wikokit; ИПМИ КарНЦ РАН, ул. Пушкинская, д. 11, г. Петрозаводск, 185910, РФ; р.т. +7(8142)76-63-13, факс +7(8142)76-63-13.

Krizhanovsky Andrew Anatoliyevich — Ph.D., senior researcher, Computer Aided Integrated Systems Laboratory at St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), senior researcher, Laboratory for Information Computer Technologies of Institute of Applied Mathematical Research of the Karelian Research Centre of the Russian Academy of Sciences (IAMR). Research Interest: information retrieval, corpus linguistics. The number of scientific publications — 67. andrew.krizhanovsky@gmail.com, code.google.com/p/wikokit; IAMR KRC RAS, 11, Pushkinskaya str., Petrozavodsk, Karelia, 185910, Russia; phone +7(8142)76-63-13, fax +7(8142)76-63-13.

Смирнов Александр Викторович — д-р техн. наук, профессор, заведующий лабораторией интегрированных систем автоматизации Федерального государственного бюджетного учреждения науки Санкт-Петербургского института информатики и автоматизации Российской академии наук (СПИИРАН). Область научных интересов: интеллектуальное управление конфигурациями виртуальных и сетевых организаций, логистика знаний, поддержка принятия решений. Число научных публикаций — 304. smir@iias.spb.su; СПИИРАН, 14-я линия В. О., д. 39, Санкт-Петербург, 199178, РФ; р.т. +7(812)328-2073, факс +7(812)328-4450.

Smirnov Alexander Victorovich — Ph.D., D.Sc., professor, a Head of Computer Aided Integrated Systems Laboratory at St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), a full professor of St.Petersburg State Electrical Engineering University. Research interests: intelligent configuration management of virtual and network organizations, knowledge logistics, decision support. The number of publications — 304. smir@iias.spb.su; SPIIRAS, 39, 14th Line V. O., St. Petersburg, 199178, Russia; tel. +7(812)328-2073, fax: +7(812)328-4450.

Круглов Василий Михайлович — д-р филол. наук, проф., ведущий научный сотрудник Федерального государственного бюджетного учреждения науки Института лингвистических исследований Российской академии наук (ИЛИ РАН), руководитель лаборатории информационных лингвистических технологий. Область научных интересов: русская лексикология и лексикография, компьютерная лингвистика, корпусная лингвистика, электронные картотеки, компьютерная лексикография. Число научных публикаций — 30. vmkruglov@yandex.ru; ИЛИ РАН, Тучков переулок, д. 9, Санкт-Петербург, 199053, РФ; р.т. +7(812)328-1612, факс +7(812)328-4611.

Kruglov Vasil Mikhailovich — Ph.D., D.Sc., Prof., leading senior researcher of Institute for Linguistic Studies of the Russian Academy of Sciences (ILI RAS), a head of Information Linguistics Technologies laboratory. Research interests: Russian lexicology and lexicography, computational linguistics, corpus linguistics, electronic library catalogues, computational lexicology. The number of publications — 30. vmkruglov@yandex.ru; ILI RAS, 39, Tuchkov pereülok 9, St. Petersburg, 199053, Russia; tel. +7(812)328-1612, fax: +7(812)328-4611.

Крижановская Наталья Борисовна — ведущий инженер-программист лаборатории информационных компьютерных технологий Федерального государственного бюджетного учреждения науки Института прикладных математических исследований Карельского научного центра Российской академии наук (ИПМИ КарНЦ РАН). Область научных интересов: разработка информационных систем для поддержки научных исследований и образования с использованием Интернет-технологий. Число научных публикаций — 29. nataly@krc.karelia.ru, http://nataly.krc.karelia.ru; ИПМИ КарНЦ РАН, ул. Пушкинская, д. 11, г. Петрозаводск, 185910, РФ; р.т. +7(8142)76-63-12, факс +7(8142)76-63-13.

Krizhanovskaya Natalia Borisovna — leading programmer, Laboratory for Information Computer Technologies of Institute of Applied Mathematical Research of the Karelian Research Centre of the Russian Academy of Sciences (IAMR). Research Interest: developing of the information system for scientific research and education using internet-technologies. The number of scientific publications — 29. nataly@krc.karelia.ru, http://nataly.krc.karelia.ru; IAMR KRC RAS, 11, Pushkinskaya str., Petrozavodsk, Karelia, 185910, Russia; phone +7(8142)76-63-12, fax +7(8142)76-63-13.

Кипяткова Ирина Сергеевна — к-т техн. наук, старший научный сотрудник лаборатории речевых и многомодальных интерфейсов Федерального государственного бюджетного учреждения науки Санкт-Петербургского института информатики и автоматизации Российской академии наук (СПИИРАН). Область научных интересов: автоматическое распознавание речи, статистические модели языка. Число научных публикаций — 50.

kipyatkova@iias.spb.su; СПИИРАН, 14-я линия В.О., д. 39, Санкт-Петербург, 199178, РФ; р.т. +7(812)328-7081, факс +7(812)328-7081.

Kipyatkova Irina Sergeevna — Ph.D., senior researcher, Laboratory of Speech and Multimodal Interfaces St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS). Research interests: automatic speech recognition statistical language models. The number of publications — 50. kipyatkova@iias.spb.su; SPIIRAS, 39, 14th Line V.O., St. Petersburg, 199178, Russia; office phone +7(812)328-7081, fax +7(812)328-7081.

Поддержка исследований. Работа выполнена при финансовой поддержке РФФИ (проект № 12-01-00481, № 12-07-00070, № 12-08-01265, № 13-07-12095, № 14-07-00345), РГНФ (проект № 12-04-12062, № 13-04-12020), проекта № 213 Программы фундаментальных исследований Президиума РАН «Интеллектуальные информационные технологии, математическое моделирование, системный анализ и автоматизация», проекта № 2.2 Программы ОНИТ РАН «Интеллектуальные информационные технологии, системный анализ и автоматизация», Программы фундаментальных исследований Президиума РАН «Корпусная лингвистика» 2012-2014, направление 3 «Создание и развитие корпусных ресурсов по языкам народов России» (проект «Корпус вепсского языка: пополнение и развитие электронного ресурса», рук. проекта доктор филол. наук, зав. сектором языкознания ИЯЛИ КарНЦ РАН Зайцева Н.Г.).

Acknowledgement. Some parts of the research were carried out under projects funded by grants № 12-01-00481, № 12-07-00070, № 12-08-01265, № 13-07-12095, № 14-07-00345 of the Russian Foundation for Basic Research, grant № 12-04-12062, № 13-04-12020 of the Russian Foundation for Humanities and project of the research program “Intelligent information technologies, mathematical modelling, system analysis and automation” of the Russian Academy of Sciences and project of the research program Corpus Linguistics 2012-2014, direction “Creation and development of body resources on languages of the peoples of Russia”.

РЕФЕРАТ

Крижановский А.А., Смирнов А.В., Круглов В.М., Крижановская Н.Б., Кипяткова И.С. **Автоматическое извлечение словарных помет из Русского Викисловаря.**

В работе рассматривается задача извлечения слабоструктурированных данных из интернет-словарей на примере Русского Викисловаря. Из текстов толкований словарных статей извлекаются словарные пометы — краткие грамматические, стилистические и другие указания на характеристики описываемой лексемы. В статье описываются особенности словарных помет в викисловарях.

Разработана методология извлечения словарных помет из викисловарей. В соответствие с этой методологией экспертами построено отображение (соответствие один к одному) системы словарных помет Русского Викисловаря (385 помет) и системы словарных помет Английского Викисловаря (1001 помета). Таким образом, построена интегральная система словарных помет (1096 помет), включающая пометы обоих словарей. Особое внимание уделено экспрессивным пометам, приводится «выровненный» список экспрессивных помет Русского Викисловаря и Английского Викисловаря.

Разработан синтаксический анализатор (парсер), который распознаёт и извлекает известные и новые словарные пометы, сокращения и пояснения, указанные в начале текста значений слов в словарных статьях Викисловаря. Следует отметить наличие в парсере большого количества словарных помет известных заранее (385 словарных помет для Русского Викисловаря).

С помощью парсера на основе данных Русского Викисловаря была построена база данных машиночитаемого Викисловаря, включающая информацию о словарных пометах. В работе приводятся численные параметры словарных помет в Русском Викисловаре, а именно: с помощью разработанной программы было подсчитано, что в базе данных машиночитаемого Викисловаря к 133 тыс. значений слов приписаны пометы и пояснения; для полутора тысяч значений слов был указан регион употребления слова, подсчитано число словарных помет для разных предметных областей. Вкладом данной работы в компьютерную лексикографию является оценка численных параметров словарных помет в больших словарях (пятьсот тысяч словарных статей).

SUMMARY

Krizhanovsky A.A., Smirnov A.V., Kruglov V.M., Lugovaya N.B., Kipyatkova I.S. **Automatic extraction of context labels from the Russian Wiktionary.**

The problem of extracting semistructured data from internet dictionaries was considered in the paper. Context labels (grammatical or restricted-usage information) about definitions were extracted from Russian Wiktionary entries. Features of Wiktionary context labels were described in the paper.

The methodology of extracting context labels from wiktionaries was developed. In accordance with this methodology experts constructed a mapping table that establishes a correspondence between Russian Wiktionary context labels (385 labels) and English Wiktionary context labels (1001 labels). As a result the composite system of context labels (1096 labels), which includes both dictionary labels, was constructed. Special attention was paid to expression labels, the aligned list of expression labels of Russian Wiktionary and English Wiktionary was presented.

The parser extracting context labels from the Russian Wiktionary was developed. The parser can recognize and extract new context labels, abbreviations and comments placed before the definition in Wiktionary articles. One outstanding feature of this parser is a large number of context labels which are known in advance (385 context labels for Russian Wiktionary). The parser can recognize and extract new context labels, abbreviations and comments placed before the definition in Wiktionary articles.

The database of machine-readable Russian Wiktionary including context labels was generated by the parser. An evaluation of numerical parameters of context labels in the Russian Wiktionary was performed. With the help of the developed computer program it was found in the Russian Wiktionary that (1) there are 133 000 definitions with context labels and comments, (2) one and a half thousand definitions were supplied with regional labels, (3) it was calculated a number of definitions with labels for each domain knowledge. This paper is an original contribution to computational lexicography, setting out for the first time an analysis of numerical parameters of context labels in the large dictionary (500 000 entries).