



Мухин М. Ю. Универсальный идеографический словарь-тезаурус русской лексики : организация данных и формирование словника / М. Ю. Мухин // Научный диалог. — 2022. — Т. 11. — № 10. — С. 88—103. — DOI: 10.24224/2227-1295-2022-11-10-88-103.

Mukhin, M. Yu. (2022). Universal Ideographic Dictionary-Thesaurus of Russian Vocabulary: Data Organization and Glossary Formation. *Nauchnyi dialog*, 11 (10): 88-103. DOI: 10.24224/2227-1295-2022-11-10-88-103. (In Russ.).



Журнал включен в Перечень ВАК

DOI: 10.24224/2227-1295-2022-11-10-88-103

Универсальный идеографический словарь-тезаурус русской лексики: организация данных и формирование словника

Мухин Михаил Юрьевич

orcid.org/0000-0001-8716-9260

доктор филологических наук, доцент,
заведующий кафедрой

фундаментальной и прикладной
лингвистики и текстоведения

mikhail.mukhin@urfu.ru

Уральский федеральный университет
имени первого Президента России

Б. Н. Ельцина

(Екатеринбург, Россия)

Благодарности:

Исследование выполнено
за счет гранта

Российского научного фонда

№ 22-18-00352,

<https://rscf.ru/project/22-18-00352/>

Universal Ideographic Dictionary-Thesaurus of Russian Vocabulary: Data Organization and Glossary Formation

Mikhail Yu. Mukhin

orcid.org/0000-0001-8716-9260

Doctor of Philology, Associate Professor,
Head of the Department of Fundamental
and Applied Linguistics and Text Science

mikhail.mukhin@urfu.ru

Ural Federal University
named after the First President of Russia

B. N. Yeltsin

(Yekaterinburg, Russia)

Acknowledgments:

The research was supported

by RSF (project № 22-18-00352),

<https://rscf.ru/project/22-18-00352/>

ОРИГИНАЛЬНЫЕ СТАТЬИ

Аннотация:

Рассматриваются принципы сбора и организации данных для нового Универсального идеографического словаря-тезауруса русской лексики, создаваемого в Уральском федеральном университете (Екатеринбург, Россия) под руководством профессора Л. Г. Бабенко. Обозревается специфика предшествующих лексикографических источников. Охарактеризованы идеографические базы данных, с которыми коллектив Уральской семантической школы работал ранее. Приводятся примеры выделения денотативных сфер и денотативно-идеографических групп, структура которых составляет синопсис тезауруса. Обсуждается количество этих классов и набор записей в исходной базе данных, которая является основой будущего словаря. Кроме того, показаны примеры форм для работы со словарями и структура данных (набор полей) Универсального идеографического словаря-тезауруса. Основной акцент в статье делается на соотношении исходной словарной базы данных и словаря создаваемого словаря, формируемого на основании принципов частотности и логико-семантической целостности денотативно-идеографических групп. Рассматриваются причины количественной асимметрии нового тезауруса и исходной базы данных: статистические, логико-семантические и лексико-грамматические факторы. Делаются выводы о перспективах нового проекта как объединяющего и развивающего идеи Уральской семантической школы.

Ключевые слова:

лексикография; идеографический словарь; тезаурус; словарь; база данных; Уральская семантическая школа.

ORIGINAL ARTICLES

Abstract:

The principles of data collection and organization for the new Universal Ideographic Dictionary-Thesaurus of Russian Vocabulary, created at the Ural Federal University (Yekaterinburg, Russia) under the guidance of Professor L. G. Babenko is discussed in the article. The specificity of previous lexicographic sources is reviewed. Ideographic databases, with which the staff of the Ural Semantic School worked earlier, are characterized. Examples of identifying denotative spheres and denotative-ideographic groups, the structure of which makes up the synopsis of the thesaurus are given. The number of these classes and the set of entries in the original database, which is the basis of the future dictionary, are discussed. In addition, examples of forms for working with dictionaries and the data structure (a set of fields) of the Universal Ideographic Dictionary-Thesaurus are shown. In addition, examples of forms for working with dictionaries and the data structure (a set of fields) of the Universal Ideographic Dictionary-Thesaurus are shown. The reasons for the quantitative asymmetry of the new thesaurus and the original database are considered: statistical, logical-semantic and lexical-grammatical factors. Conclusions are drawn about the prospects of the new project as uniting and developing the ideas of the Ural Semantic School.

Key words:

lexicography; ideographic dictionary; thesaurus; glossary; database; Ural semantic school.

Универсальный идеографический словарь-тезаурус русской лексики: организация данных и формирование словника

© Мухин М. Ю., 2022

1. Введение = Introduction

Цель данной статьи — обсудить научно-практические проблемы создания нового Универсального словаря-тезауруса русской лексики, который будет представлять собой объемный межчастеречный лексический свод, упорядоченный по денотативно-идеографическому принципу. С одной стороны, любой словарь должен быть оригинальным ресурсом с точки зрения макро- и микроструктуры. С другой стороны, никакой лексикографический источник не может появиться на «пустом поле», без опоры на словари-предшественники. Особенно это касается организации данных идеографических словарей, словарей-тезаурусов, в основе которых лежат классификационные структуры. Их составители решают ряд проблем, связанных с противоречивостью семантических классификаций, пересечением семантических парадигм, существованием лексики с комплексной семантикой и т. д. Чисто лингвистические проблемы влияют, во-первых, на разработку архитектуры лексикографических баз данных, которые должны отражать семантическую иерархию. Во-вторых, требуется решать вопросы формирования нового словника и наполнения идеографических баз словарным контентом. При этом любая практическая реализация семантической классификации всегда обусловлена конечными задачами создаваемого ресурса, который может быть классическим «бумажным» словарем, электронной словарной базой данных, семантически размеченным корпусом или программой автоматического семантического анализа или синтеза речи.

Универсальный словарь-тезаурус русской лексики будет сочетать черты классического словаря и электронного тезауруса. На сегодняшний день он формируется как электронная база данных, но по решению коллектива может быть издан и в бумажной форме. Отличие электронной версии всегда заключается в удобстве поиска как по лексемам, так и по классификационной сетке. С другой стороны, бумажные версии всех ранее созданных словарей содержали идеографическую часть и алфавитный словник, по которому можно было легко найти любое значение в классификационной структуре. Однако пользовательский интерфейс электронного тезауруса и

структура его «бумажного» аналога — это пока дело будущего, а основные вопросы данной статьи связаны с соотношением создаваемого лексикографического ресурса и уже существующих словарей коллектива Уральской семантической школы и вопросами преемственности и развития данных.

2. Материал, методы, обзор = Material, Methods, Review

2.1. Проблемное поле идеографической лексикографии

Исследовательский и пользовательский интерес к словарям тезаурусного типа объясняется широтой их сегодняшнего применения — от системного изучения и преподавания языка, лингводидактики до компьютерной лингвистики и информационных технологий. Несмотря на разницу в гуманитарных и инженерных трактовках понятий «семантическая сеть», «онтология», «тезаурус», «концептуальная система» и др., современные задачи поиска и интерпретации информации требуют создания классификационных структур, системно представляющих лексику естественного языка и отражающих стандартные семантические отношения гипонимии, когипонимии (между словами одной семантической группы), синонимии, антонимии и т. д.

Не вдаваясь в подробности, можно констатировать, что проблемное поле современного идеографического представления лексики объединяет ряд направлений, имеющих определенные сходства и различия в подходах к словарным классификациям и практическом применении создаваемых ресурсов. Во-первых, это уже ставшие классическими идеографические словари, ориентированные на их использование человеком, а не программой машинной обработки текста. На материале русского языка осуществлены основные проекты Уральской семантической школы, о которых пойдет речь в следующем разделе, а также [Баранов, 2002; Васильев, 2003; Русский семантический словарь, 1998; СТСРИ, 2007; Шушков, 2008] и мн. др. Такие ресурсы ориентированы в первую очередь на изучение специфики репрезентации знаний о мире в языке, особенностей соотношения лексических классов, а также на преподавание языка (в том числе как иностранного). Другое направление связано с созданием электронных машиночитаемых тезаурусов [Русский Wordnet; Braslavsky et al., 2014; RussNet...; Fellbaum, 1998; WordNet... и др.]. Электронные тезаурусы можно использовать классическим способом, а также встраивать их в алгоритмы информационного поиска и решения задач контекстного снятия омонимии. Третье касается семантической разметки лингвистических корпусов [Апресян и др., 2004; НКРЯ; Рахилина и др., 2009], которая пока еще является весьма редкой и не очень качественной в силу трудоемкости и сложности аннотирования.

Естественно, что специфика каждого направления определяет и особенности синопсисов, то есть классификационных сеток. В корпусах, в силу множественности материала, они вряд ли могут быть слишком гранулярными. Например, в Национальном корпусе русского языка все глаголы движения фактически приписаны к одному классу. И наоборот, в толковых идеографических словарях мы видим дробность деления больших классов на семантические или денотативные группы. В глагольном словаре [ТСРГ, 1999] отдельно представлены категории «Движение» и «Перемещение» и каждая имеет разветвленную подструктуру.

В словарях, изданных на бумаге, количество классификационных уровней обычно значительно меньше, чем в электронных тезаурусах, интерфейс которых позволяет воспринимать и длинные гипо-гиперонимические цепочки. Так, в классическом Принстонском ворднете к слову *table* строится в направлении абстрактность / конкретность ряд синсетов, содержащий 9 уровней: *table* ← *furniture*, *piece of furniture*, *article of furniture* ← *furnishing* ← *instrumentality*, *instrumentation* ← *artifact*, *artefact* ← *whole*, *unit* ← *object*, *physical object* ← *physical entity* ← *entity*. А дальше слова *table* существует еще 4 гипонимических уровня, на которых находятся слова и выражения с более конкретной семантикой — *tea table* и др. [WordNet...].

В отличие от ворднетов, в словарях Уральской семантической школы, изданных классическим способом, выделяется от четырех до шести классификационных уровней.

2.2. Словари Уральской семантической школы и формирование общей лексикографической базы данных

Уральская семантическая школа, объединяющая в первую очередь лингвистов-лексикографов Уральского федерального университета и возглавляемая профессором Л. Г. Бабенко [Бабенко, 2022; Уральская семантическая школа..., 2011], известна серией разных по типу и задачам идеографических словарей [ТСРГ, 1999; БТСРС, 2005; БТСРГ, 2007; СТССР, 2007; БТССР, 2008; СТРП, 2011 и др.]. За тридцать лет работы коллективом создано 25 лексикографических ресурсов, в рамках которых менялась и развивалась основная идеографическая концепция. Первый словарь — словарь русских глаголов — был построен исходя из строго семантического критерия. Ранние основы словаря были заложены в 1980-е годы профессором Э. В. Кузнецовой [Лексико-семантические группы..., 1988], а основная синоптическая структура, принцип интерпретации лексических значений и другие параметры определены Л. Г. Бабенко [ТИСРГ, 1997; ТСРГ, 1999; БТСРГ, 2007]. В дальнейшем принцип категоризации лексики изменился в сторону денотативно-идеографического основания — подробнее об этом см.: [Бабенко, 2020]. На сегодняшний день «сам перечень

категорий составлен нами на основе синопсиса (свода) идеографической классификации лексики знаменательных частей речи русского языка. Один из основных принципов его организации — иерархичность категорий, формирующих лексические множества разного ранга: идеографические сферы, классы, группы, подгруппы <...> Всего в синопсисе нами выделены 15 базовых денотативных сфер, отображающих основные фрагменты действительности» [Бабенко, 2017, с. 12]. Перечислим эти сферы, составляющие верхний уровень классификационной структуры:

Неживая природа.
Вещества и материалы.
Живая природа.
Человек как живое существо.
Населенный пункт.
Нации.
Родственные и семейные отношения.
Человек и его внутренний мир.
Язык и речь.
Сверхъестественное.
Общественно-государственная сфера.
Конкретная физическая деятельность.
Социальная сфера жизни человека.
Восприятие окружающего мира.
Универсальные представления, смыслы и отношения.

Указанные денотативные сферы далее членятся на денотативные классы, что образует на нижнем уровне классификации (от второго до шестого, в соответствии с дробностью конкретной сферы) денотативно-идеографические группы (ДИГ), объединяющие конкретные слова. Так, с определенными различиями, построены словари существительных [БТСРС, 2005], прилагательных [СТРП, 2011], синонимов [СТСРР, 2007; БТССРР, 2008; СТСРЯ, 2017], «Концептосфера русского языка» [Концептосфера..., 2017] и др. Например, в текущем варианте синопсиса в сфере «9. Язык и речь» выделяются подклассы «9.1. Язык» и «9.2. Речь», а последний имеет следующую структуру:

9.2.1. Характеристики речи человека.

9.2.1.1. Характерные особенности произношения, выговора, манеры речи.

9.2.1.2. Речевые стили и характерные особенности изложения, текста, стиля.

9.2.2. Речевое сообщение.

9.2.3. Речевое общение.

9.2.4. Речевое воздействие.

9.2.5. Характеристика человека по особенностям речи.

Очевидно, что из перечисленных рубрик класс «9.2.1. Характеристики речи человека», в отличие от прочих, конкретной денотативно-идеографической группой не является, так как распадается на группы четвертого уровня. Различия в количестве классификационных уровней определяются большей или меньшей дробностью понятийной организации денотативных сфер, а самой разветвленной из них (6 уровней) является сфера «3. Живая природа». В ней, например, среди прочих выделяется ДИГ «3.1.2.2.1. Декоративные травянистые растения». В итоге из 974 номинаций денотативных классов в классификационной структуре 783 являются ДИГ, содержащими реальные слова, а 191 — названиями денотативных сфер и субсфер, как, например, «3. Живая природа» или «3.1. Растительный мир». Вследствие развития и уточнения словарных данных не исключается дальнейшая частная правка синопсиса.

Исторически для каждого словарного проекта разрабатывалась своя база данных, в том числе архитектура хранилища и пользовательский интерфейс. Впоследствии эти данные были объединены и дополнены в соответствии с частеречными, частотными и другими характеристиками лексем, не представленных в созданных ранее словарях (см. об этом подробнее в [Мухин, 2014]). Полученная единая база данных отличалась, с одной стороны, полнотой охвата лексики и содержала 97 тысяч (в сегодняшней версии — 101 тысячу) записей, соответствующих значениям слов, или лексико-семантическим вариантам. С другой стороны, это лексическое собрание было в определенном смысле эклектично, что связано с особенностями объединения данных из разных по концепциям и словникам словарей. Для удобства пользовательской работы была организована объединяющая данные форма, позволяющая добавлять, редактировать и удалять записи (рис. 1).

2.3. Работа по созданию Универсального идеографического словаря-тезауруса русской лексики

Новый этап обращения к словарным данным озаглавлен проектом создания Универсального идеографического словаря-тезауруса русской лексики [Универсальный идеографический словарь..., 2015; Свод лексики...]. Лексикографический ресурс будет отражать «свод русской лексики, систематизированной по группам и расположенной в словаре в определенном порядке, в схематическом отвлеченном виде представляемом в виде

Тезаурус

Поиск по индексу: 1.13. Ветер и его признаки

Поиск по слову:

Лемма	Вокабула	Сл.	ЛСВ	Найти в	Стар.	Нов. №	Изм.
баргузин	БАРГУЗИН	сущ.	1	статья	3.12.	1.13.	
бешеный	БЕШЕНЫЙ	при	9	статья	1.13.	1.13.	18
бриз	БРИЗ	тез	1	1.8.	1.13.	1.13.	18
буйный	БУЙНЫЙ	сущ.	1	статья	3.12.	1.13.	
буйный	БУЙНЫЙ	при	7	статья	1.13.	1.13.	7
бурный	БУРНЫЙ	тез	1	1.8.	1.13.	1.13.	7
буранный	БУРАННЫЙ	при	1	статья	3.12.	1.13.	4
бурьян	БУРЯН	тез	1	1.7.	1.13.	1.13.	4
бурьянный	БУРАННЫЙ	при	1	статья	1.13.	1.13.	
бурьян	БУРЬНЫЙ	при	9	статья	1.13.	1.13.	27
бурьян	БУРЯН	тез	1	1.8.	1.13.	1.13.	27
бурьян	БУРЯ	сущ.	1	статья	3.12.	1.13.	20
ветер	ВЕТЕР	сущ.	4	статья	3.12.	1.13.	140
ветер	ВЕТЕР	тез	1	1.8.	1.13.	1.13.	140
ветр	ВЕТР	тез	1	1.8.	1.13.	1.13.	
ветровой	ВЕТРОВОЙ	при	1	статья	1.13.	1.13.	2
ветровый	ВЕТРОВЫЙ	тез	1	1.8.	1.13.	1.13.	2
ветровый	ВЕТРОВЫЙ	при	1	статья	1.13.	1.13.	
ветряной	ВЕТРЯНОЙ	при	1	статья	1.13.	1.13.	1
солнечный	СОЛНЕЧНЫЙ	тез	1	1.8.	1.13.	1.13.	
ветер	ВЕТЕР	тез	1	1.8.	1.13.	1.13.	6
вихревой	ВИХРЕВОЙ	при	1	статья	1.13.	1.13.	
вихревой	ВИХРЕВОЙ	тез	1	1.8.	1.13.	1.13.	
вихрь	ВИХРЬ	сущ.	1	статья	3.12.	1.13.	9
вихрь	ВИХРЬ	тез	1	1.8.	1.13.	1.13.	9
встречный	ВСТРЕЧНЫЙ	при	2	статья	1.13.	1.13.	22
встречный	ВСТРЕЧНЫЙ	тез	1	1.8.	1.13.	1.13.	22
вьюга	ВЬЮГА	сущ.	1	статья	3.12.	1.13.	2

Затем: 1. 27 из 974

Рис. 1. Базовая форма для работы со словарями

идеографической карты, построенной на основе структуры денотативно-идеографической сети. Она обусловлена также обращенностью к изучению нового типа глобального лексического множества — поличастеречного свода русской лексики» [Свод лексики...]. Такой подход, частично реализованный уже в указанных выше словарях синонимов, позволяет уйти от жестко заданной частеречной принадлежности слов и, соответственно, категориально-лексической семантики, образывавшей ранее лексико-семантические группы слов одной части речи, в сторону денотативно-идеографических множеств. Таким образом, денотативный принцип, в отличие от чисто семантического, реализованного во многих классических словарях (например, в [Русский семантический словарь, 1998; БТСРГ, 2007 и др.], в электронных тезаурусах [Fellbaum, 1998] и корпусах [НКРЯ; Рахилина и др., 2009], не предполагает заведомого разделения слов на разные части речи, которые могут входить в одну идеографическую группу (ДИГ).

Для группировки слов, входящих в Универсальный словарь-тезаурус, была создана новая таблица с полями для следующей информации:

Лемма.

Часть речи.

Порядковый номер значения в группе.

Индекс в текущем тезаурусе.

Специфика подачи леммы в одном из созданных ранее словарей (с ударением, как заголовочная или производная единица).

Словарные пометы:

стилистическая,

эмоциональная,

семантическая (*перен.*),

хронологическая,

другая.

Вариант леммы (например, ЗАТМ'ЕНЬЕ к ЗАТМ'ЕНИЕ).

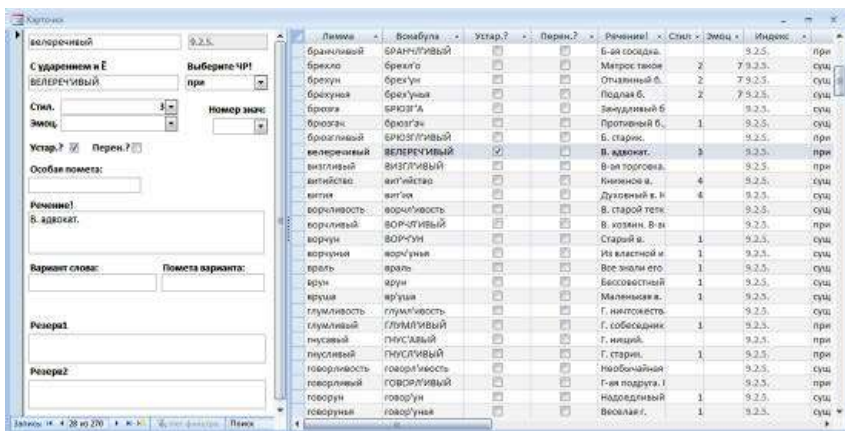
Словарная помета для варианта леммы.

Пример употребления (в виде предложения).

Участник коллектива — разработчик идеографической группы, добавивший запись.

Резервные поля для дополнительной информации, которая может войти в уточненную структуру словаря в ходе работы.

Основная часть этих полей, предназначенная для редактирования вручную, представлена в виде новой пользовательской формы (рис. 2).



Лемма	Варианты	Устар.?	Перен.?	Речение!	Стил.	Эмб.	Индекс
бранчивый	БРАНЧИВЫЙ	☐	☐	Б-я соседка.		9.2.5.	при
брежу	БРЕЖУ	☐	☐	Магрос такое	2	7.9.2.5.	сущ
брежу	БРЕЖУ	☐	☐	Отчаянный б.	2	7.9.2.5.	сущ
бредуха	БРЕДУХА	☐	☐	Подлый б.	2	7.9.2.5.	сущ
бредуха	БРЕДУХА	☐	☐	Завидный б.		9.2.5.	сущ
бродяга	БРОДЯГА	☐	☐	Проточный б.	1	9.2.5.	сущ
бродяга	БРОДЯГА	☐	☐	Б. старик.		9.2.5.	при
верпачивый	ВЕРПЕРЧИВЫЙ	☒	☒	В. адвокат.	3	9.2.5.	при
визгливый	ВИЗГЛИВЫЙ	☐	☐	В-ля торговка.		9.2.5.	при
визгливый	ВИЗГЛИВЫЙ	☐	☐	Князёв в.	4	9.2.5.	сущ
визгливый	ВИЗГЛИВЫЙ	☐	☐	Духовный в. И	4	9.2.5.	сущ
ворчливый	ВОРЧЛИВЫЙ	☐	☐	В. старой тетк.		9.2.5.	сущ
ворчливый	ВОРЧЛИВЫЙ	☐	☐	В. колзин. В-я		9.2.5.	при
ворчун	ВОРЧУН	☐	☐	Старый в.	1	9.2.5.	сущ
ворчун	ВОРЧУН	☐	☐	Их в-лейтн.	1	9.2.5.	сущ
ворчун	ВОРЧУН	☐	☐	Все знают его.	1	9.2.5.	сущ
ворчун	ВОРЧУН	☐	☐	Бессовестный	1	9.2.5.	сущ
ворчун	ВОРЧУН	☐	☐	Маленькая в.	1	9.2.5.	сущ
глумливый	ГЛУМЛИВЫЙ	☐	☐	Г. неложность.		9.2.5.	сущ
глумливый	ГЛУМЛИВЫЙ	☐	☐	Г. собеседник.	1	9.2.5.	при
глумливый	ГЛУМЛИВЫЙ	☐	☐	Г. нидий.		9.2.5.	при
глумливый	ГЛУМЛИВЫЙ	☐	☐	Г. старик.	1	9.2.5.	при
глумливый	ГЛУМЛИВЫЙ	☐	☐	Необычайная		9.2.5.	сущ
глумливый	ГЛУМЛИВЫЙ	☐	☐	Г-я подруга. Г		9.2.5.	при
говорливый	ГОВОРЛИВЫЙ	☐	☐	Находливый	1	9.2.5.	сущ
говорливый	ГОВОРЛИВЫЙ	☐	☐	Веселая г.	1	9.2.5.	сущ

Рис. 2. Форма для работы с Универсальным словарем-тезаурусом

Поскольку основная идеографическая иерархия уже представлена в ряде материнских таблиц основной базы, для размещения данных нового словаря достаточно одной реляционной таблицы, содержащей конечные слова и словарную микроструктуру. В то же время новая форма позволяет эффективно работать с одной денотативно-идеографической группой (в примере на рис. 2 — «9.2.5. Характеристика человека по особенностям

речи»). Таким образом, Универсальный словарь-тезаурус, в целом являясь автономным источником, наследует общую идеографическую структуру, а модель взаимодействия данных позволяет реализовать в системе управления базами данных структуры для любых новых идеографических словарей.

3. Результаты и обсуждение = Results and Discussion

На сегодняшний день авторами-составителями добавлено в таблицу Универсального словаря-тезауруса около 30 тысяч записей, и эта работа в данный момент продолжается. Конечный тезаурус по составу словника будет отличаться от исходной редактируемой базы и в качественном, в количественном отношении. Во-первых, в нем будут преодолены проблемы асимметричности слияния в один ресурс нескольких сделанных ранее лексикографических источников, а также представлены слова, ранее отсутствовавшие в словарях глаголов, существительных, прилагательных и в тезаурусе синонимов. Впервые в тезаурусе будет полностью, а не фрагментами реализована измененная макроструктура словаря, то есть денотативно-идеографический синопсис.

С точки зрения объема — количества представленных значений — Универсальный словарь-тезаурус должен отличаться от исходной базы: в нем будут удалены повторы ЛСВ и применен частотный фильтр для входа лексемы в новый словарь. Например, субсфера «11.11. Военная служба» исходно включает 1420 ЛСВ, из них в Универсальный тезаурус вошли 902 единицы описания. Выше, на рис. 1, виден столбец с показателем частоты *ipm* (instances per million words) — согласно Новому частотному словарю русского языка [Новый частотный словарь...]. Ориентировочно в новый тезаурус должны войти лексемы с индексом не менее 3.0. Однако каждый автор с учетом фактора многозначности и здравого смысла решает вопрос о добавлении в тезаурус менее частотных лексем и, наоборот, отказа от лексем с частотой больше 3.0. Величина *ipm* в нашей работе — это чисто справочный показатель, так как он связан с леммой, а не с конкретным значением. Например, частота слова *праздник* высока и равна 115 вхождений на миллион слов. Из этого, однако, не следует, что данный показатель равномерно формируется всеми его лексико-семантическими вариантами. Например, в значении «испытываемое от чего-либо наслаждение, приятное, радостное чувство» слово *праздник* используется, на наш взгляд, достаточно редко, а столь важной статистики по употреблению значений, а не лексем, как известно, не будет существовать до появления корпусов с качественной семантической разметкой.

Кроме того, в парадигме слов одной денотативно-идеографической группы отсутствие слова, менее частотного, но существенного для полноты этой парадигмы может стать заметным для пользователя и досадным недостатком словарных данных. Например, в группе «11.11.3.2. Род войск» слова *пехота* и *пехотный* с частотой *ipm* 14,5 и 5,3 оба соответствуют количественному критерию и без проблем входят в новый тезаурус. Другая пара — *флот* и *флотский* (41 и 0,4) — имеет другие количественные показатели, но отсутствие относительного прилагательного в группе на фоне наличия других производных слов нарушает пропорциональность представления данных. Поэтому прилагательное *флотский*, несмотря на относительно небольшую частоту, также включено в словарь-тезаурус. В итоге при формировании конечного словника сочетаются статистический и логико-семантический критерии.

Объем нового словаря может быть увеличен за счет добавления значений, не описанных ранее, и разведения в отдельные словарные статьи видовых пар глаголов. Ранее в словаре [ТСРГ, 1999] глаголы совершенного вида (*рассмотреть*, *развести*, *найти* и т. д.) при существовании видовых пар были представлены в статьях глаголов несовершенного вида (*рассматривать*, *разводить*, *находить*). Отдельные статьи будут посвящены в тезаурусе и некоторым лексемам, считавшимся до этого вариантами основных вокабул. В первую очередь это феминитивы (*учительница* ← *учитель*, *школьница* ← *школьник*).

Вследствие таких решений, например, в сфере «9. Язык и речь» количество в 3974 ЛСВ выросло до 4425 значений в таблице Универсального словаря-тезауруса. Рассматриваемое количественное соотношение зависит от подробности разработки значений конкретной группы слов в словарях-предшественниках и представленности в разных классах глагольной лексики. К примеру, среди «военной» лексики глаголы составляют 11,5 %, и после применения частотного фильтра в тезаурусе эта сфера несколько уменьшилась в объеме. В речевой сфере глаголов значительно больше (почти 40 %), что в значительной степени объясняет ее увеличение. Только в группе «9.2.2. Речевое сообщение» из 985 ЛСВ 668 составляют глагольные, например: *анонсировать*, *аттестовать*, *балабонить*, *благовестить*, *болтать*, *бредить*, *брехать*, *брякать*, *вернуть*, *вдалбливать*, *вещать* <...> *утверждать*, *уточнять*, *характеризовать*, *хватить*, *чернить*, *шептать*, *шепаться*, *шушукаться*, *ябедничать*.

Разведение видовых пар глаголов, которые ранее рассматривались в одной словарной статье, производных и производящих существительных и некоторые другие моменты обусловлены лексико-грамматическим критерием формирования нового словника.

4. Заключение = Conclusions

Поскольку в статье речь идет о только создаваемом в данный момент словаре, его различные параметры и лингвистические характеристики могут корректироваться по ходу работы. Формирование словника является универсальной лексикографической проблемой, особенно актуальной для идеографических словарей, а также для справочников специального назначения (школьных, узкопрофессиональных и т. п.). Критерий частотности слова срабатывает при ее решении далеко не всегда, тем более что надежной информации о частоте значения (ЛСВ), а не лексемы пока нет. Семантическая разметка в Национальном корпусе русского языка [НКРЯ] проведена без снятия омонимии — самой трудной задачи аннотирования корпуса.

Как мы видим, асимметрия исходной базы данных словарей, созданных лексикографическим коллективом УрФУ, и Универсального идеографического словаря-тезауруса русской лексики, а также критерии формирования нового словника определяются рядом причин, в том числе статистических, логико-семантических и лексико-грамматических:

- общая эклектичность исходной базы, в которую вошли данные нескольких словарей, отличающихся по внутреннему устройству;
- включение в словник лексем, ранее не представленных в словарях коллектива;
- разделение пар и вариантов лексем, ранее разрабатывавшихся в одной словарной статье;
- объединение слов разных частей речи в одной денотативно-идеографической группе;
- частотные характеристики слов;
- стремление к гармоничности распределения слов в новой группе, пропорциональному представлению лексики с точки зрения семантических и грамматических признаков.

Таким образом, Универсальный идеографический словарь-тезаурус русской лексики — это объединяющий проект, который корректирует и развивает идеи Уральской семантической школы. Кроме того, новый словарь имеет хорошие перспективы использования как в традиционной форме, так и в формате электронного (в том числе — машиночитаемого) ресурса.

Источники

1. БТСРГ — *Большой толковый словарь русских глаголов* : свыше 10 000 глаголов. Идеографическое описание. Синонимы. Антонимы. Английские эквиваленты / [авт.-сост.: Л. Г. Бабенко и др.] ; под общ. ред. Л. Г. Бабенко. — Москва : АСТ-ПРЕСС, 2007. — 573 с. — ISBN 978-5-462-00615-9.

2. БТСРС — *Большой толковый словарь русских существительных : Идеографическое описание. Синонимы. Антонимы* / Под ред. Л. Г. Бабенко. — Москва : АСТ-ПРЕСС, 2005. — 864 с. — ISBN 978-5-462-00738-5.
3. БТСРП — *Большой толковый словарь синонимов русской речи : идеографическое описание, антонимы, фразеологизмы* / Под ред. Л. Г. Бабенко. — Москва : АСТ-ПРЕСС, 2008. — 784 с. — ISBN 978-5-462-00893-1.
4. НКРЯ — *Национальный корпус русского языка* [Электронный ресурс]. — Режим доступа : <http://www.ruscorpora.ru> (дата обращения 24.10.2022).
5. *Русский Wordnet* [Электронный ресурс]. — Режим доступа : <https://wordnet.ru/> (дата обращения 24.10.2022).
6. *Свод лексики как идеографическая карта мира : Универсальный словарь-тезаурус русского языка. Карточка проекта, поддержанного Российским научным фондом* [Электронный ресурс]. — Режим доступа : <https://rscf.ru/project/22-18-00352/> (дата обращения 24.10.2022).
7. СТРП — *Словарь-тезаурус русских прилагательных, распределенных по тематическим группам* / Под ред. Л. Г. Бабенко. — Москва : Проспект, 2011. — 232 с. — ISBN 978-5-392-12243-1.
8. СТСРИ — *Словарь-тезаурус современной русской идиоматики : около 8000 идиом современного русского языка* / под ред. : А. Н. Баранова, Д. О. Добровольского. — Москва : Мир энциклопедий Аванта+, 2007. — 1134 с. — ISBN 978-5-98986-143-9.
9. СТСПР — *Словарь-тезаурус синонимов русской речи* / Под ред. Л. Г. Бабенко. — Москва : АСТ-Пресс, печ., 2007. — 512 с. — ISBN 978-5-462-00729-3.
10. СТСПРЯ — *Словарь-тезаурус синонимов русского языка* / Под ред. Л. Г. Бабенко. — Москва : АСТ-Пресс Школа, 2017. — 448 с. — ISBN 978-5-9909262-0-2.
11. ТИСПРГ — *Толковый идеографический словарь русских глаголов : с указанием англ. эквивалентов : проспект / под общ. ред. Л. Г. Бабенко.* — Екатеринбург : Издательство Уральского университета, 1997. — 72 с. — ISBN 5-7525-0569-0.
12. ТСПРГ — *Толковый словарь русских глаголов : Идеографическое описание ; Английские эквиваленты ; Синонимы ; Антонимы* / под ред. Л. Г. Бабенко. — Москва : [б. и.], 1999. — 704 с.

Литература

1. *Апресян Ю. Д.* Семантическая разметка в глубоко аннотированном корпусе русского языка / Ю. Д. Апресян, Л. Л. Иомдин, А. В. Санников, В. Г. Сизов // Труды международной конференции «Корпусная лингвистика-2004». — Санкт-Петербург : [б. и.], 2004. — С. 41—54.
2. *Бабенко Л. Г.* Концепция, структура и основные лексикографические параметры словаря / Л. Г. Бабенко // Концептосфера русского языка : ключевые концепты и их репрезентации в языке и речи (на материале лексики, фразеологии и паремиологии) : словарь / под общ. ред. Л. Г. Бабенко. — Москва : Азбуковник, 2017. — С. 11—21. — ISBN 978-5-91172-128-2.
3. *Бабенко Л. Г.* От полипарадигмальности — к интеграции парадигм : Уральская семантическая школа в контексте лингвистики конца XX — начала XXI в. / Л. Г. Бабенко // Динамика языковых и культурных процессов в современной России. — Материалы VII Конгресса РОПРЯЛ (г. Екатеринбург, 6—9 октября 2021 года). — Санкт-Петербург : РОПРЯЛ, 2022. — Выпуск 7. — С. 23—28.

4. *Бабенко Л. Г.* Типы лексических множеств в структурно-семантическом, когнитивно-дискурсивном и лексикографическом освещении : динамика интерпретаций / Л. Г. Бабенко // Научный диалог. — 2020. — № 9. — С. 9—47. — DOI: 10.24224/2227-1295-2020-9-9-47.
5. *Баранов О. С.* Идеографический словарь русского языка / О. С. Баранов. — Москва : [б. и.], 2002. — 1200 с. — ISBN 5-7042-1080-5.
6. *Васильев Л. М.* Системный семантический словарь русского языка. Предикатная лексика. Ментальные предикаты. Модальные предикаты. Предикаты восприятия / Л. М. Васильев. — Уфа : Издательство БашГУ, 2003. — 125 с.
7. *Концептосфера* русского языка : ключевые концепты и их репрезентации в языке и речи (на материале лексики, фразеологии и паремиологии) : словарь / под общ. ред. Л. Г. Бабенко. — Москва : Азбуковник, 2017. — 1020 с. — ISBN 978-5-91172-128-2.
8. *Лексико-семантические группы русских глаголов* : Учеб. словарь-справочник / Авт.-сост. Э. В. Кузнецова и др. — Свердловск : Издательство Уральского университета, 1988. — 151 с.
9. *Мухин М. Ю.* Базы данных лексикографической группы «Русский глагол» : статистические параметры и возможности их лингвистической интерпретации / М. Ю. Мухин // Известия Уральского федерального университета. — Серия 2. — Гуманитарные науки. — 2014. — № 4 (133). — С. 256—262.
10. *Новый частотный словарь русской лексики* / Под ред. О. Н. Ляшевской и С. А. Шарова [Электронный ресурс]. — Режим доступа : <http://dict.ruslang.ru/freq.php> (дата обращения 24.10.2022).
11. *Рахилина Е. В.* Задачи и принципы семантической разметки лексики в НКРЯ / Е. В. Рахилина, Г. И. Кустова, О. Н. Ляшевская, Т. И. Резникова, О. Ю. Шеманаева // Национальный корпус русского языка : 2006—2008. Новые результаты и перспективы. — Санкт-Петербург : Нестор-История, 2009. — С. 215—239.
12. *Русский семантический словарь* : Толковый словарь, систематизированный по классам слов и значений / РАН. Ин-т рус. яз. им. В. В. Виноградова; [Авт.-сост.: Н. Ю. Шведова и др.] ; Под общ. ред. Н. Ю. Шведовой. — Москва : Азбуковник : ИРЯ РАН, 1998. — ISBN 5-88744-008-2.
13. *Универсальный идеографический словарь русского языка* : Проспект / под общ. ред. Л. Г. Бабенко. — Москва — Екатеринбург : Кабинетный ученый, 2015. — 208 с. — ISBN 978-5-7525-3067-8.
14. *Уральская семантическая школа* : история, люди, события. — Екатеринбург : Издательство Уральского университета, 2011. — 345 с. — ISBN 978-5-7996-0634-3.
15. *Шушков А. А.* Толково-понятийный словарь русского языка 600 семантических групп : около 16500 слов и устойчивых выражений / А. А. Шушков. — Москва : АСТ, 2008. — 988 с. — ISBN 978-5-17-018959-5.
16. *Braslavsky P. A* Spinning Wheel for YARN : User Interface for a Crowdsourced Thesaurus / P. Braslavsky, M. Mukhin, D. Ustalov // Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics. — Gothenburg, Sweden : [b. i.], 2014. — Pp. 101—104.
17. *Fellbaum C.* WordNet : An electronic database. — Cambridge, MA : MIT Press, 1998. — 422 p.
18. *RussNet* : тезаурус русского языка [Электронный ресурс]. — Режим доступа : http://project.phil.spbu.ru/RussNet/index_ru.shtml (дата обращения 26.10.2022).
19. *WordNet* : A Lexical Database for English [Electronic resource]. — Access mode : <https://wordnet.princeton.edu/> (accessed 25.10.2022).

Material resources

- Babenko, L. G., Volchkova, I. M., Kazarin, Yu. V. (1997). Explanatory ideographic dictionary of Russian verbs: With an indication of English. Equivalents: Prospect. Yekaterinburg: Ural University Publishing House. 72 p. ISBN 5-7525-0569-0. (In Russ.).
- Babenko, L. G. (ed.). (1999). *Explanatory Dictionary of Russian verbs: Ideographic description; English equivalents; Synonyms; Antonyms*. Moscow: [b. i.]. 704 p. (In Russ.).
- Babenko, L. G. (ed.). (2005). *A large explanatory dictionary of Russian nouns: Ideographic description. Synonyms. Antonyms*. Moscow: AST-PRESS. 864 p. ISBN 978-5-462-00738-5. (In Russ.).
- Babenko, L. G. (ed.). (2007). *A large explanatory dictionary of Russian verbs: over 10,000 verbs. Ideographic description. Synonyms. Antonyms. English equivalents*. Moscow: AST-PRESS. 573 p. ISBN 978-5-462-00615-9. (In Russ.).
- Baranov, A. N., Dobrovolsky, D. O. (2007). *Thesaurus of modern Russian idiomatics: about 8000 idioms of the modern Russian language*. Moscow: World of Avanta+ Encyclopedias. 1134 p. ISBN 978-5-98986-143-9. (In Russ.).
- Babenko, L. G. (ed.). (2007). *Dictionary-thesaurus of synonyms of Russian speech*. Moscow: AST-Press, pech. 512 p. ISBN 978-5-462-00729-3. (In Russ.).
- Babenko, L. G. (ed.). (2008). *A large explanatory dictionary of synonyms of Russian speech: ideographic description, antonyms, phraseological units*. Moscow: AST-PRESS. 784 p. ISBN 978-5-462-00893-1. (In Russ.).
- Babenko, L. G. (ed.). (2011). *Dictionary-thesaurus of Russian adjectives distributed by thematic groups*. Moscow: Prospekt. 232 p. ISBN 978-5-392-12243-1. (In Russ.).
- Babenko, L. G. (ed.). (2017). *Dictionary-thesaurus of synonyms of the Russian language*. Moscow: AST-Press School. 448 p. ISBN 978-5-9909262-0-2. (In Russ.).
- Babenko, L. G. (ed.). *The vocabulary set as an ideographic map of the world: A universal thesaurus of the Russian language. The project card supported by the Russian Scientific Foundation-house*. Available at: <https://rscf.ru/project/22-18-00352/> (accessed 24.10.2022). (In Russ.).
- Russian National Corpus*. Available at: <http://www.ruscorpora.ru> (accessed 24.10.2022). (In Russ.).
- Russian Wordnet*. Available at: <https://wordnet.ru/> (accessed 24.10.2022). (In Russ.).

References

- Apresyan, Yu. D., Iomdin, L. L., Sannikov, A. V., Sizov, V. G. (2004). Semantic markup in a deeply annotated corpus of the Russian language. In: *Proceedings of the international conference "Corpus Linguistics-2004"*. St. Petersburg: [b. i.]. 41—54. (In Russ.).
- Baranov, O. S. (2002). *Ideographic dictionary of the Russian language*. Moscow: [b. i.]. 1200 p. ISBN 5-7042-1080-5. (In Russ.).
- Babenko, L. G. (2017). Concept, structure and basic lexicographic parameters of the dictionary. In: *Conceptosphere of the Russian language: key concepts and their representations in language and speech (based on the material of vocabulary, phraseology and paremiology): dictionary*. Moscow: Azbukovnik. 11—21. ISBN 978-5-91172-128-2. (In Russ.).
- Babenko, L. G. (2020). Types of Lexical Sets in Structural-Semantic, Cognitive-Discursive and Lexicographic Coverage: Dynamics of Interpretations. *Nauchnyi dialog*, 9: 9—47. DOI: 10.24224/2227-1295-2020-9-9-47. (In Russ.).

- Babenko, L. G. (2022). From polyparadigmality to integration of paradigms: the Ural semantic school in the context of linguistics of the late XX — early XXI century. In: *Dynamics of linguistic and cultural processes in modern Russia*, 7. St. Petersburg: ROPRYAL. 23—28. (In Russ.).
- Babenko, L. G. (ed.). *The conceptsphere of the Russian language: key concepts and their representations in language and speech (based on vocabulary, phraseology and paremiology): dictionary*. Moscow: Azbukovnik. 1020 p. ISBN 978-5-91172-128-2. (In Russ.).
- Babenko, L. G. (ed.). *Universal ideographic dictionary of the Russian language: Prospect*. Moscow — Yekaterinburg: Cabinet Scientist. 208 p. ISBN 978-5-7525-3067-8. (In Russ.).
- Braslavsky, P., Mukhin, M., Ustalov, D. (2014). A Spinning Wheel for YARN: User Interface for a Crowdsourced Thesaurus. In: *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden: [b. i.]. 101—104. (In Russ.).
- Fellbaum, C. (1998). *WordNet: An electronic database*. Cambridge, MA: MIT Press. 422 p.
- Lexical and semantic groups of Russian verbs: Studies. dictionary-reference*. (1988). Sverdlovsk: Ural University Press. 151 p. (In Russ.).
- Lyashevskaya, O. N., Sharov, S. A. (ed.). *New frequency dictionary of Russian vocabulary*. Available at: <http://dict.ruslang.ru/freq.php> (accessed 24.10.2022). (In Russ.).
- Mukhin, M. Yu. (2014). Databases of the lexicographic group “Russian verb”: statistical parameters and possibilities of their linguistic interpretation. *Izvestiya Ural Federal University. Series 2. Humanities*, 4 (133): 256—262. (In Russ.).
- Rakhilina, E. V., Kustova, G. I., Lyashevskaya, O. N., Reznikova, T. I., Shemanaeva, O. Y. (ed.). (2009). Tasks and principles of semantic markup of vocabulary in the RNC. In: *Russian National Corpus: 2006—2008. New results and prospects*. St. Petersburg: Nestor-Istoriya. 215—239. (In Russ.).
- Russian semantic dictionary: An explanatory dictionary systematized by classes of words and meanings*. (1998). Moscow: Azbukovnik: IRYa RAS. ISBN 5-88744-008-2. (In Russ.).
- RussNet: thesaurus of the Russian*. Available at: http://project.phil.spbu.ru/RussNet/index_ru.shtml (accessed 26.10.2022). (In Russ.).
- Shushkov, A. A. (2008). *Explanatory and conceptual dictionary of the Russian language 600 semantic groups: about 16500 words and stable expressions*. Moscow: AST. 988 p. ISBN 978-5-17-018959-5. (In Russ.).
- Ural semantic school: history, people, events*. (2011). Yekaterinburg: Ural University Publishing House. 345 p. ISBN 978-5-7996-0634-3. (In Russ.).
- Vasiliev, L. M. (2003). *System semantic dictionary of the Russian language. Predicate vocabulary. Mental predicates. Modal predicates. Predicates of perception*. Ufa: Bashgu Publishing House. 125 p. (In Russ.).
- WordNet: A Lexical Database for English*. Available at: <https://wordnet.princeton.edu/> (accessed 25.10.2022).

Статья поступила в редакцию 27.10.2022,
одобрена после рецензирования 24.11.2022,
подготовлена к публикации 25.12.2022.