23 апреля 2019 года на очередном заседании семинара с докладом «**Сводный исторический словарь русского языка XVIII–XX веков»** выступил А.Е. Поляков (Научно-педагогическая библиотека им. К.Д. Ушинского).

1. Описание проекта.

Сводный исторический словарь русского языка (http://dic.feb-web.ru/rusdict/) представляет собой интегрированную базу русских словарей XVIII–XX вв., которая включает глубокую разметку и возможность поиска по зонам словарной статьи.

Словарная база создается на основе наиболее авторитетных толковых и многоязычных словарей:

• САР-1 = Словарь Академии Российской (1789–1794).

• САР-2 = Словарь Академии Российской, по азбучному порядку расположенный (1806–1822).

• СЦРЯ = Словарь церковнославянского и русского языка (1847).

• Даль = Толковый словарь живого великорусского языка / 2-е изд. (1880–1882).

• Рейф = Новые параллельные словари... Ч. 1: Русский словарь (1860).

• Ушаков = Толковый словарь русского языка (1935–1940).

• МАС = Словарь современного русского литературного языка (1950–1965).

Некоторые из этих словарей отсканированы и доступны в виде графических файлов (pdf/dvju/tiff) без текста, но в таком варианте они непригодны для поиска и часто неудобны для просмотра (мелкий шрифт, двухколоночная верстка). Некоторые словари существуют в текстовом виде, но не имеют системной разметки, что резко снижает их ценность для научной работы. В лучшем случае их можно читать как электронные книги, но невозможно использовать как полноценный лексикографический источник.

Создающийся сводный исторический словарь представляет собой полнотекстовую лексикографическую базу данных, которая позволяет решать следующие задачи:

• искать статьи по заголовочному слову (с учетом вариантов);

• искать статьи по грамматическим и другим пометам;

• искать текст в зоне толкований, примеров;

• отслеживать словарную фиксацию слова во времени;

• наблюдать филиацию значений слова;

• регистрировать совпадения и различия в дефинициях;

• видеть динамику изменения стилистических характеристик слова.

Словарь отличается от существующих сводных словарей именно наличием полного текста. Например, «Сводный словарь современной

русской лексики» (1991) представляет собой просто указатель заголовков из других словарей.

Словарь принципиально отличается от существующих словарных сервисов (http://slovari.ru, http://dic.academic.ru) тем, что словарная статья имеет разметку по зонам и возможность поиска по ним.

2. Структура словарной базы.

Все словари в словарной базе данных представляются в унифицированном формате. Основной единицей является словарная статья, в которой выделяются следующие зоны:

1) заголовочное слово (лемма, вокабула, включая варианты);
2) грамматические пометы (часть речи, род, вид, переходность и др.);
3) стилистические пометы (устар., простореч., церк.);
4) этимология (иногда);
5) толкование, обычно разделенное на несколько (под)значений;
6) примеры употребления (для каждого значения), иногда с указанием источника.

Идентификатором статьи является заголовочное слово, которое может иметь несколько вариантов. Кроме того, статья может включать **подстатьи**, где фиксируются производные слова и словосочетания с данным словом (фразеологизмы), которые могут иметь свои пометы, толкования и примеры употребления. В целом статья имеет иерархическую структуру, где данные сгруппированы по слову / словосочетанию, а далее по (под)значениям.

Некоторые словари (САР-1, Даль) устроены по гнездовому принципу, где подзначения даются вперемешку с производными и фразеологизмами, поэтому в них иногда довольно трудно выделить границы (под)статей.

Рассмотрим для примера статью из словаря СЦРЯ, как она выглядит в оригинале.

> **ВИДЪ**, а, *с. м.* 1) Наружность. *Принять, показывать видъ угрюмый, гордый, веселый. Придать лучшій видъ строенію.* 2) *Церк.* Сходство, подобіе, образъ. *Роди сына по виду своему.* Быт. V. 3. 3) Разстояніе мѣста, до котораго зрѣніе можетъ простираться. *Непріятельскій корабль былъ въ виду.* 4) Изображеніе мѣстъ. *Снимать, рисовать виды.* 5) Письменное свидѣтельство о чьемъ либо состояніи или качествѣ. *Получить видъ на жительство.* 6) Сходство вещей, заключающихся въ одномъ родѣ. *Разные виды мрамора, яшмы.* 7) Предположеніе. *Многіе поступаютъ по собственнымъ своимъ видамъ.* — *Имѣть въ виду*, зн. наблюдать, чтобы исполненіе чего либо сдѣлано было въ свое время. — *Выпустить, потерять изъ виду*, зн. забыть или пропустить нечаянно кого или что либо. — *Подъ видомъ*, зн. подъ предлогомъ. — *Быть на виду*, зн. часто находиться у кого либо на глазахъ. — *Поставить на видъ*, зн. обратитъ вниманіе, указать. — *Видомъ не видано.* Пог. зн. совсѣмъ не видано.

Помчали, и за столъ роскошный посадили,
Какого и видомъ не видано у насъ. Дмитр.

Вся статья представляет собой один абзац, где толкования и примеры даются подряд без разделителей, и только стихотворная цитата выделена в отдельный абзац. Курсив в тексте статьи используется в трех значениях: 1) грамматические и стилистические пометы (*с. м.*); 2) примеры употребления; 3) заголовки подстатей (*Имѣть въ виду*). Такой текст неудобен даже для чтения, тем более он непригоден для интеллектуального поиска. Для нормальной работы текст статьи нужно разрезать на маленькие фрагменты, из которых он был составлен — толкования, примеры, пометы, заголовки подстатей, разделители и т.п.

После разметки по зонам текст статьи приобретает такой вид:

ВИДЪ, а, *с. м.*

1) Наружность.

Принять, показывать видъ угрюмый, гордый, веселый.

Придать лучшій видъ строенію.

2) *Церк.* Сходство, подобіе, образъ.

Роди сына по виду своему. Быт. V. 3.

3) Разстояніе мѣста, до котораго зрѣніе можетъ простираться.

Непріятельскій корабль былъ въ виду.

4) Изображеніе мѣстъ.

Снимать, рисовать виды.

5) Письменное свидѣтельство о чьемъ либо состояніи или качествѣ.

Получить видъ на жительство.

6) Сходство вещей, заключающихся въ одномъ родѣ.

Разные виды мрамора, яшмы.

7) Предположеніе.

Многіе поступаютъ по собственнымъ своимъ видамъ. —

Имѣть въ виду, зн.

наблюдать, чтобы исполненіе чего либо сдѣлано было въ свое время. —

Выпустить, потерять изъ виду, зн.

забыть или пропустить нечаянно кого или что либо. —

Подъ видомъ, зн.

подъ предлогомъ. —

Быть на виду, зн.

часто находиться у кого либо на глазахъ. —

Поставить на видъ, зн.

обратитъ вниманіе, указать. —

Видомъ не видано. Пог. зн.

совсѣмъ не видано.

Помчали, и за столъ роскошный посадили,
Какого и видомъ не видано у насъ. Дмитр.

В размеченном тексте каждая зона представляет собой строку (абзац) с определенным стилем. Кроме того, статья содержит скрытую разметку для заголовков, подзаголовков и грамматических помет, которая дается в нормализованном виде, пригодном для поиска.

3. Унификация словарной информации.

Словарная база включает словари, созданные в разное время, по разным принципам и в разной орфографии, поэтому возникает необходимость унификации словарной информации для обеспечения глобального поиска. При этом мы не меняем исходный текст словаря, но добавляем к нему дополнительную разметку, которую можем унифицировать для поиска.

Словари могут различаться по следующим параметрам:

• орфография;

• заголовочное слово;

• грамматические пометы;

• структура статьи.

Орфография словарей XVIII–XIX вв. отличается не только от современной, но и от стандартной дореформенной («гротовской»), сложившейся к концу XIX в. Морфологический анализатор (лемматизатор), ориентированный на современную орфографию, не может нормально работать с текстами в старой орфографии. Основные орфографические отличия легко разрешаются программно, например, если заменить старые буквы на современные эквиваленты (*і, ѣ, ѳ, ѵ* → *и, е, ф, и*, конечный *-ъ* → ноль), то современный лемматизатор начинает опознавать многие старые формы. Однако различия, касающиеся морфологии (написание определенных форм или морфем), требуют радикальной переделки грамматического словаря и грамматических таблиц. Вот неполный список таких различий:

1) флексии *-аго, -яго, -ыя, -ія*;

2) формы *ея, онѣ, однѣ, однѣхъ*;

3) приставки *без-, воз-, из-, низ-, раз-, через-* + глухие (*возходь, изкушать, изчезать, разтворить*);

3б) приставка *з-* + звонкие (*збавить, зберечь, згинуть, здавать*);

4) ударное *-ый/-ій* (*больны́й, босы́й, водяны́й, глухі́й, другі́й, слѣпы́й*);

4б) безударное *-ой/-ей* (*волчей корень, бобръ камчацкой*);

5) *ь/и* перед гласными (*вниманье, занятье, в Итальи*, *милостію*);

6) *е/о* после шипящих и ц (*лице, значекъ, чортъ*);

7) *-ся* после гласных (*валюся, валилася*);

8) суффиксы компаратива *-ѣй* (*скорѣй*), *-яе* (*скоряе*);

9) деепричастия сов. вида от основы презенса (*придя, увидя, взгромоздясь*);

10) церковнославянские флексии *-ти* (*благодѣ́яти*), *-ши* (*благодѣ́еши*);

11) написание слитно/раздельно/дефисно (*то-есть, повидимому, кто нибудь*).

Заголовки статей и подстатей сохраняются в тексте как есть, но к ним добавляются скрытые теги разметки, где заголовки записаны в нормализованном виде. При этом мы сохраняем оригинальную орфографию (*i, ѣ, ѳ, ѵ, -ъ,* ударение), но в ряде случаев написание приходится слегка модернизировать или изменять, чтобы работал поиск.

Например, в САР-1 приставка *с-* часто пишется как *з-* перед звонкими согласными (*збавить, зберечь, збирать, згинуть, здавать*); приставки на *-з* сохраняют *з* перед глухими согласными (*возприять, возходь, изкони, изкушать, источникъ, изчезать, изходъ, разкаяться, разтворить*). Прилагательные с ударным *-ой* в старых словарях часто пишутся через *-ый/-ій* (*больный, босый, глухій, дорогій, другій, плохій, слѣпый*). Здесь нормализованное написание было приведено к современной норме.

В САР-1 заглавная форма глагола дается по античному образцу в форме 1 л. ед. ч. (*алчу, бію, бѣгу, веду*), которую по современным правилам необходимо перевести в инф. (*алкать, бить, бѣжать, вести*). В СЦРЯ глаголы с пометой *Церк./Стар.* даются с окончанием *-ти* (*благовѣстити, блистати, вдати, вергнути*), которое необходимо перевести в современное *-ть*.

Грамматические пометы в словарях даются в разном виде и с разной степенью полноты. Для поиска мы перевели все пометы в унифицированный формат, аналогичный стандарту грамматической разметки в Национальном корпусе русского языка (http://ruscorpora.ru/corpora-morph.html).

Словари XVIII–XIX вв. по сравнению с современными словарями дают более подробную залоговую классификацию глаголов, которая отображается на современную нотацию так: действительный=переходный (tr), средний=непереходный (intr), страдательный=med,pass, возвратный=med,refl, взаимный= med,recip и т.д.

С другой стороны, словари САР-1 и САР-2 очень скудно отображают видовые характеристики глаголов, поскольку объединяют в одну статью все видовые варианты (*двигать, двигнуть, двинуть, двигивать*), которые в современных словарях разнесены по разным статьям и снабжены соответствующими пометами.

Словарь САР-1 построен по гнездовому принципу, поэтому в заголовке статьи могут быть собраны не только дублеты (*архиварій/архиваріусъ, безлюдіе/безлюдье, велблюдъ/верблюдъ*), но и близкородственные слова, включая производные, например: *агнецъ, агнчикъ, агница; баловникъ, -ница, баловщикъ, -щица; волкъ, волчокъ, волчище, волчица; избавитель, -ница; самодержавіе, самодержавство,*

самодержство. Слова в заголовке могут иметь разные грамматические характеристики (*волкъ*=N,m vs. *волчица*=N,f), что создает некоторую путаницу при поиске по грамматическим признакам.

4. Поисковый движок (Sphinxsearch).

В качестве поискового движка в системе был выбран Sphinxsearch (http://sphinxsearch.com). Эта программа представляет собой систему полнотекстового поиска в больших коллекциях текстов и обладает широкими возможностями настройки.

Вот основные возможности программы:

1) поиск в текстах с HTML-образной разметкой, которую можно учитывать или игнорировать;

2) задание поисковых зон при помощи произвольных тегов (<sem>, <sample>);

3) таблица преобразования символов для индексатора;

4) морфологический анализ для русского и других языков;

5) пользовательский словарь для морфологического анализа;

6) поиск точных форм (=слово);

7) контекстный поиск («точная фраза», расстояние, порядок слов, операторы или, нет).

Поиск по зонам словарной статьи (толкования и примеры) легко реализуется при помощи встроенных возможностей Sphinxsearch. В противном случае пришлось бы «распиливать» статью на зоны для поиска, а потом собирать полный текст статьи из фрагментов.

Поиск конкретного слова в разных зонах дает совершенно разные результаты. Например, поиск слова «планета» в зоне толкований дает статьи или значения, связанные с астрономией: *аспе́ктъ, астро́логъ, блудя́щія звѣзды, Вене́ра, вечерняя звѣзда, зодіа́къ, квадрату́ра, коме́та, Ма́рсъ, противостоя́ніе, спу́тникъ.* Поиск слова «планета» в зоне примеров дает самые разнообразные статьи: *гада́ніе, дви́гаться, земля́, идти́, круговы́й, но́вый, обраще́ніе, подви́жный, путь, свѣ́тъ*. Если взять более частотное слово, то его наличие в зоне толкований может что-то значить, тогда как попадание в зону примеров становится почти случайным, особенно в длинных цитатах.

Поиск по старой орфографии частично решается при помощи таблицы преобразования символов. Текст статьи сохраняется в оригинальной орфографии, но при индексации старые буквы заменяются на современные эквиваленты: i=и ї=и v=и ѣ=е є=е ю=е o=o w=o оу=у ѫ=у ыж=у ѧ=я ыа=я ѳ=ф s=з. К сожалению, в таблице можно задать только простые замены, а для остального приходится слегка преобразовывать исходный текст, в частности:

1) конечный Ъ,ъ заменяется на символ U+048c, U+048d (semisoft sign), который игнорируется при индексировании, но сохраняется в выдаче;

2) составные буквы (ѯ, ѱ, ѿ) заменяются на диграфы (кс, пс, от).

После этого современный лемматизатор начинает распознавать многие слова в старой орфографии, кроме тех, где есть морфологические различия (см. п. 3).

Чтобы полностью решить проблему старой орфографии, нужно переделать лемматизатор (как?) или подключить пользовательский словарь, где указать все старые формы и их современные эквиваленты (*новаго→нового, новыя→новые, возстание→восстание*). Пока мы ограничимся нормализацией заголовочных слов, а поиск по всему тексту отложим.

5. Результаты.

В настоящее время полностью подготовлены, размечены и загружены в систему следующие словари:

• САР-1

Объем: 45.6 тыс. статей, 3200 подстатей, 63 тыс. заголовочных слов и словосочетаний. Объем текста: 7.3 млн символов, 1 млн слов.

• СЦРЯ

Объем: 113 тыс. статей, 5600 подстатей, 120 тыс. заголовочных слов и словосочетаний. Объем текста: 10.7 млн символов, 1.5 млн слов.

Эти словари вместе с САР-2 хорошо покрывают русскую лексику вплоть до первой половины XIX века.

Разработана технология разметки и унификации словарной информации для разных источников. Написаны скрипты для автоматизации разметки и проверки словарей, хотя окончательная разметка и проверка все равно делается вручную.

Для работы поиска разработана схема и настроена конфигурация базы данных, написаны скрипты для загрузки текстов в БД. Разработан пользовательский интерфейс поиска, который доступен по адресу: http://dic.feb-web.ru/rusdict/search.htm.

6. Направления дальнейшей работы.

1. Пополнение системы за счет новых словарей.

САР-2 — проверены и частично размечены тома 1–4, тома 5–6 требуют проверки. Текстуально САР-1 во многом совпадает с САР-2 и может использоваться для его проверки.

Словари XX века (МАС, Ушаков) частично размечены по зонам и доступны в словарном разделе ФЭБ (http://feb-web.ru/feb/feb/dict.htm). Для интеграции их в систему требуется проверка и дополнительная разметка.

2. Расширение системы разметки словарей по зонам.

Возможно расширение списка зон за счет более четкой дифференциации компонентов словарной статьи: главное слово vs. производное vs. фразеологизм.

В настоящее время только главное слово получает полный набор грамматических помет, поскольку они явно обозначены в словаре. В дальнейшем можно расширить систему помет для производных слов и

фразеологизмов, например, ввести псевдограмматические категории NP (*антонов огонь*) и VP (*бить челом*), которые придется проставлять вручную.

3. Дальнейшая унификация орфографии и грамматических помет в разных словарях.

В настоящее время многие заголовочные слова (особенно в САР-1) не полностью нормализованы и унифицированы в соответствии с современной орфографией, что затрудняет их поиск.

Предполагается использовать разработанный нами парсер для старой орфографии для нормализации и унификации заголовочных слов в соответствии с современным написанием.

4. Доработка инструментария для автоматизированной разметки словарных статей.

Исходные тексты словарей обычно содержат массу ошибок, в том числе неправильную расстановку маркеров (курсив, жирность и т.д.), что затрудняет автоматическую разметку.

Предполагается сделать верификацию и исправление формальных ошибок на начальном этапе разметки, что позволит сократить ручную работу по проверке зон.