Конкорданс к текстам Ломоносова — концепция и реализация

Lomonosov concordance — concept and implementation[1]

Поляков А. Е. (pollex@mail.ru)
НТЦ «Информрегистр»

Бергельсон М. Б. (mirabergelson@gmail.com)
МГУ имени М. В. Ломоносова

Пильщиков И. А. (pilshch@yandex.ru)
ИМК МГУ имени М. В. Ломоносова

В докладе уточняются понятия и термины, связанные с разработкой полного электронного Конкорданса к текстам Ломоносова и обсуждаются практические решения, необходимые для реализации этого лексикографического продукта. Конкорданс строится на основе корпуса авторских текстов, снабженных структурной, филологической и грамматической разметкой. Описывается технология построения корпуса и конкорданса, принципы разметки корпуса, структура словарной статьи конкорданса, а также возможности его применения для лингвистических исследований.

0. Введение.

0.1. Цели и принципы

Конкорданс к произведениями и письмам М. В. Ломоносова, над которым работают участники настоящего проекта, строится на основе электронного корпуса текстов М. В. Ломоносова, представляющего собой филологически корректную цифровую версию академического Полного собрания сочинений и писем Ломоносова в 11-ти томах (1950–1983) и ряда дополнительных изданий. Цель конкорданса — представить авторское словоупотребление во всей его широте и во всей полноте его языковой специфики.

Конкорданс к текстам Ломоносова является частью проекта по созданию электронного научного издания «Ломоносов», которое позволит предоставить широкому кругу пользователей программно-информационную среду для изучения литературного и научного наследия, языка и биографии Ломоносова. Проект предусматривает создание открытого интернет-ресурса, который будет включать в себя:

1) корпус текстов Ломоносова, построенный на основе наиболее авторитетных изданий;
2) биографические, литературно-критические и историко-научные работы о Ломоносове;
3) полный алфавитно-частотный конкорданс к текстам Ломоносова.

Ломоносовский конкорданс строится на основе принципиально новой методологии и технологии подготовки, отвечающей современному уровню филологической науки. В основе его лежит корпус филологически выверенных авторских текстов, снабженных богатой структурной, филологической и грамматической разметкой.

0.2. Основные определения

Ключевыми для данного проекта являются понятия **конкорданса**, **корпуса**, а также их уточнения, как-то — электронный (сетевой) конкорданс,

---

[1] В докладе изложены результаты работы трех участников проекта «Электронное научное издание “Ломоносов”: корпус текстов, справочная информация, алфавитно-частотный конкорданс» (грант РГНФ 08-04-12120в; руководитель — чл.-корр. РАН В. А. Виноградов). Первоначально докладчики предполагали выступить с двумя сообщениями на взаимосвязанные темы, но по предложению оргкомитета объединили оба доклада в один.

полный корпус, дифференциальные словари. Корректное определение этих ключевых понятий возможно только на фоне анализа соответствующей концептуальной области.

Анализ современных лексикографических продуктов и аналитических работ по авторской и — шире — общей лексикографии показывает, что термин «словарь» представляет собой родовой обозначение, соответствующее общеязыковому употреблению слова. Фактически, в широком смысле слова, это некоторым образом упорядоченный список символов, выбранный из некоторого множества текстов.

Если говорить более подробно, то **словарь** в широком смысле подразумевает определенный способ выборки словника, множество текстов, из которого делается эта выборка, характеристику вокабул, дополнительную информацию, которая сопровождает каждую вокабулу, возможности филиации значений и их толкований. Таким образом, ключевые различия проявляются как на уровне микроструктуры — того, что представляет собой единица описания плюс структура словарной статьи, — так и на уровне макроструктуры — пространства текстов, являющегося базой для составления словаря. На основе описания этих параметров можно говорить и о принципиальном различии между конкордансами и собственно словарями (в узком смысле слова); оно состоит в принципах отбора, подачи и описания лексических единиц, в постановке различных целей описания и следовании им.

Минималистское определение конкорданса с необходимостью включает представление о корпусе (наборе текстов, отобранных и препарированных с определенной целью). Тогда **конкорданс** к некоторому корпусу — это список словоупотреблений (элементов корпуса) с отсылкой ко всем контекстам. Противопоставление между словарями и конкордансами идет сразу по нескольким шкалам — репрезентативности, ориентации на инвариант, смыслового или грамматического анализа. Можно сформулировать следующие противопоставления:

- Словарный подход к описанию лексики ориентирован на репрезентативность и тем самым на **нормативность**, а корпусный вариант (конкорданс) — на **исчерпывающее** описание.
- Словарь, анализируя различные употребления лексемы в разных значениях, стремится к нахождению **инварианта**, конкорданс — к **вариативности** и ставит своей первейшей задачей отразить все случаи употребления слова. Поэтому представление слова в конкордансе ставит во главу угла примеры (контексты словоупотреблений), а в словаре — словарную статью.
- На различном понимании термина **полнота** базируется еще одно принципиальное противопоставление словаря и конкорданса: полнота словаря определяется стремлением к исчерпывающему описанию значений, полнота конкорданса определяется исчерпывающим характером описания соответствующего корпуса.
- Из этого вытекает принципиальная необходимость грамматической (морфологической) информации в конкордансе, что помогает охарактеризовать и — если нужно — различить формы, и необходимость семантического анализа (описания или толкования значений) в словаре.

Итак, можно сформулировать представление об «идеальном» (прототипическом) словаре и прототипическом конкордансе. Словарный подход стремится выделить некоторую сущность и истолковать ее. Прототипический словарь — это нормативный толковый словарь, содержащий большое количество семантической информации, в частности, семантических и стилистических помет. Прототипический конкорданс обладает полнотой тезаурусного типа (словаря, в котором максимально полно представлены слова языка с примерами их употребления в тексте, что в полном объеме осуществимо, да и то с оговорками, лишь для мертвых языков), он является первой производной корпуса, «нарезанным» корпусом, так что в идеале совокупность примеров всех употреблений составляют корпус. Он обязан быть полным в отношении учета абсолютно всех словоупотреблений. От списка слов (словника) его отличает наличие морфологической характеристики словоупотреблений. В конкордансе морфологическое описание реализуется как лемматизация, позволяющая приписать каждой словоформе граммемную характеристику. Это, в свою очередь, позволяет различить омонимичные формы. Реальной такая постановка задачи и стремление хоть сколько-нибудь приблизиться к данному прототипу возможны только в электронной форме сетевого ресурса. Именно такой конкорданс должен быть реализован в данном проекте.

Из указанных различий между прототипическим словарем и прототипическим конкордансом следует и то, что именно является для каждого из них объектом и единицей описания. Для словаря это значение и словарная статья, для конкорданса — словоупотребление и корпус. Таким образом, различие между конкордансом и толковым словарем заключается в том, что конкорданс не предполагает установления структуры (филиации) значений регистрируемых слов и не обязательно включает толкование этих значений. Его полнота, в частности, требует, чтобы он был словарем регистрирующего типа. Задача Конкорданса состоит в том, чтобы отразить все возможные различия и особенности употреблений, так как заранее неизвестно, какие из них могут оказаться значащими и значимыми. В этом отношении Конкорданс представляет собой базу для созданий различного рода **дифференциальных**, в том числе и **толковых**, словарей.

1. Общефилологические аспекты создания Конкорданса

1.1. Выбор источника

Поскольку всякий конкорданс есть особого рода лингвистически препарированный указатель к конкретному корпусу текстов, перед составителями встает проблема отбора филологически корректных текстов данного автора. Для электронного корпуса последнее означает оптимальное соответствие выбранному печатному изданию, для печатного — соответствие тем задачам, которые ставит перед собой конкорданс.

Из имеющихся изданий Ломоносова 11-томное академическое издание (**ПСС**) в наибольшей степени пригодно для намеченных целей, однако по целому ряду параметров оказывается неудовлетворительным и оно. Академическое издание непоследовательно отражает ломоносовское правописание: для разных томов был выбран различный орфографический режим, что привело к неоднородности корпуса, положенного в основу Конкорданса. Однако остальные издания еще менее пригодны с текстологической и лингвистической точек зрения, а попытка исправить 11-томное издание стала бы попыткой подготовки нового критического издания; между тем такая задача явно выходит за рамки обсуждаемого проекта.

1.2. Произведение и текст. Проблема неустойчивости текста

Совокупность авторских текстов исторически делится на произведения. Произведением считается автономный текст, выделенность которого определена автором или (достаточно часто!) его редакторами. Текстуальный состав произведения не является диахронически неизменным: сам текст еще при жизни автора и по воле автора может варьироваться. Современные научные издания (и в том числе ПСС Ломоносова) стремятся представить текст с учетом его авторской вариативности. Альтернативные фрагменты текста (то есть фрагменты, **не** сосуществующие ни в одном его синхронном срезе) представлены либо как другие редакции произведения (связные тексты), либо как варианты (набор разночтений к основной либо иной редакции).

Для задач настоящего Конкорданса принято решение учитывать в **алфавитной** части все словоупотребления, зафиксированные в ПСС и в дополнительных источниках. Вопрос о том, как учитывать вариативность в **частотном** словаре, остается открытым. Ясно, что нужно дать пользователю возможность, как минимум, учитывать частоту употребления того или иного слова в основном корпусе текстов (то есть в наборе «окончательных редакций»). Другие редакции (связные тексты) можно принимать за отдельные произведения и давать две статистики: с учетом и без учета других редакций. Однако в целом ряде случаев полный текст другой редакции нам неизвестен. Абсолютную частотность слов, появляющихся в вариантах (разночтениях) можно было бы учитывать таким же образом, однако в таком случае неясно, к чему мы приравниваем общее количество слов в данном произведении.

1.3. Проблема неоднородности корпуса

Не всякая наблюдаемая в корпусе вариативность имеет авторскую природу. Серьезную проблему для словаря представляет неоднородность, созданная непоследовательностью или множественностью редакторских подходов, реализованных в используемом корпусе.

По отношению к ПСС это, в первую очередь, проблемы, связанные с орфографическим режимом представления разных произведений: как и в других изданиях послевоенного времени, ПСС модернизирует текст Ломоносова. Эта модернизация проходит 3 стадии: (1) замена символов (букв), отмененных реформой 1918 г. (ять, фита, ижица, и десятиричное), буквами современного алфавита (*е*, *ф*, *и*, *и* либо *й*); (2) замена современными морфемами морфем, отмененных реформой 1918 г. (флексии *-аго*, *-яго*; префиксы *из-*, *без-*, *воз-* etc. перед глухими консонантами; и др.); (3) иные замены начертаний слов, приближающие ломоносовское написание к современному (например, *горкий* VS *горький*, *не льзя* VS *нельзя* — как видим, сказанное относится не только к буквенному составу слов, но и к слитному/раздельному написанию, прописным/строчным буквам и т.д.).

Мало того, что модернизированные написания не соответствуют ломоносовским — в ПСС отсутствует единый режим орфографической модернизации. Так, в материалах по русской грамматике в примерах не проведена модернизация типа (1) и (2). Модернизация типа (3) не проведена в 8-м томе, содержащем поэтические и риторические произведения, но проведена в остальных томах. Например, в 7-м томе ПСС в «Предисловии о пользе книг церьковных» прилагательное *церьковный*, вопреки всем источникам текста, дано в твердом варианте (*церковный*), а в поэме о Петре Великом, напечатанной в 8-м томе, то же прилагательное выглядит как *церьковный* (в соответствии с последним прижизненным изданием 1761 г.).

Единственный выход из ситуации — не учитывать вариативности буквенного состава слов, слитного/раздельного написания и написания с прописной/строчной буквы при сведении словоформ в лек-

сему (вокабулу). Случаи типа слитного/раздельного написания *не* с глаголами и другими частями речи потребуют специального лингвистического анализа, а в отдельных случаях, возможно, и текстологической проверки ПСС. К сожалению, в рамках данного проекта такая проверка не может быть проведена тотально. Однако в тех случаях, когда мы можем с уверенностью исправить текст ПСС, мы должны сделать в Конкордансе соответствующую помету и учесть верное чтение.

2. Разметка

2.1. Определение и классификация видов разметки

Электронный текст имеет не линейную структуру, а включает несколько параллельных слоев информации. С одной стороны, текст состоит из языковых элементов различного уровня (слова, фразы, предложения), с другой стороны, он состоит из структурных сегментов различных типов (заголовки, сноски, ремарки, стихи, цитаты, таблицы, формулы, страницы). Обычно при построении корпусов учитывается только языковое членение текста, но мы считаем, что структурное членение является не менее важным. Многие структурные элементы текста имеют яркие языковые особенности, которые требуют специальной формы представления в корпусе.

Разметка — это расстановка в тексте документа специальных маркеров (тегов), которые эксплицируют «скрытые» элементы информации, присутствующие в тексте. В зависимости от типа этой информации можно выделить следующие виды разметки:

1) **Метатекстовая** разметка.

Включает параметры, характеризующие текст в целом, в частности:

- автор (фамилия, имя, отчество);
- название произведения (заголовок, подзаголовок, incipit);
- короткое имя (используется для цитирования в примерах);
- дата создания произведения;
- жанрово-тематический класс произведения; и т. д.

2) **Структурная** разметка.

Эксплицирует логическую структуру текста, в частности:

- деление текста на структурные элементы (части, главы, действия, явления, реплики);
- заголовки структурных элементов;
- сноски, ремарки;
- стихотворные элементы (строфы, стихи);
- таблицы, формулы, рисунки, и т.д.

Многие из этих элементов имеют яркие языковые особенности и требуют специальной разметки.

3) **Форматная** разметка.

Описывает параметры оформления текста, включая:

- параметры шрифта (размер, жирность, курсив, разрядка, верхние/нижние индексы);
- параметры абзаца (выравнивание, отступы, межстрочный интервал);
- полиграфические и декоративные элементы (колонтитулы, виньетки);
- номера страниц и стихов; и т.д.

Форматная разметка нужна для адекватного представления текста в электронной библиотеке, а для лингвистического анализа многие элементы этой разметки оказываются ненужными или, наоборот, недостаточно информативными. Например, шрифтовое выделение заголовка носит чисто декоративный характер, а шрифтовое выделение отдельных слов может означать самые разные сущности: цитата, пример употребления, формула, часть слова и т.д. Абзацные отступы и выравнивание могут маркировать структурные элементы текста: заголовок, подпись, стихотворная строка определенного размера и т.д. Следовательно, для анализа текста нельзя просто игнорировать оформление, но нужно дать ему семантическую интерпретацию и поставить соответствующий структурный тег.

4) **Грамматическая** разметка.

Описывает признаки, характеризующие конкретное словоупотребление, включая:

- лексема (словарная форма);
- грамматические признаки лексемы (часть речи, одушевленность, переходность);
- грамматические признаки словоформы (число, падеж, наклонение, время, лицо).

Параллельно с грамматической информацией, эта разметка эксплицирует членение текста на языковые элементы (токены) различного типа — предложения, слова, знаки препинания, цифры.

2.2. Формат и технология разметки

Формат разметки текстов для корпуса должен учитывать многоплановость текста и наличие в нем нескольких параллельных слоев информации — метатекстовой, структурной и собственно лингвистической. Формат должен быть открытым, компактным, расширяемым, он должен быть совместим с существующими форматами разметки и легко интегрироваться с программами обработки. Ключевыми свойствами здесь являются открытость и со-

гласованность формата на всех этапах обработки. Именно это позволяет связать все операции в единую технологическую цепочку, на входе которой находится неразмеченный текст, а на выходе — размеченный корпус, из которого автоматически получается конкорданс.

Формат разметки текстов для корпуса был разработан на базе существующих стандартов представления текстовой информации для интернета (HTML+CSS) и стандартов кодирования лингвистической разметки для корпусов (TEI, XCES). После детального анализа существующих стандартов разметки корпусов был сделан вывод, что универсальные стандарты типа TEI или XCES являются слишком сложными, избыточными и неудобными для массовой разметки текстов. Напротив, формат HTML позволяет адекватно представить структурную, форматную и метатекстовую информацию, а также допускает использование нестандартных тегов. Поэтому в качестве формата разметки для корпуса было выбрано подмножество HTML плюс некоторые элементы TEI/XCES для кодирования грамматической разметки.

Технологическая цепочка подготовки корпуса включает следующие этапы обработки:

1) первичная разметка текста для представления в электронной библиотеке;
2) дополнительная структурная разметка и сегментация текста для корпуса;
3) грамматическая разметка и ее ручная постобработка (снятие омонимии, исправление разборов);
4) преобразование в базу данных, построение конкордансов и других производных.

Первоначально корпус текстов Ломоносова подготавливается в формате HTML со специальной разметкой, ориентированной на представление в электронной библиотеке. Этот формат включает достаточно полную метатекстовую, структурную и форматную разметку, необходимую для точного воспроизведения содержания и внешнего вида текста, но недостаточную для корпуса.

На следующем этапе в текст вносится дополнительная структурная разметка, которая маркирует фрагменты текста, требующие специальной обработки (заголовки, цитаты, примеры, комментарии, иноязычный текст и т.д.). Иногда такие фрагменты легко распознаются по специфическому оформлению (курсив, выравнивание), но часто их приходится определять и размечать вручную.

Далее текст пропускается через морфологический анализатор (парсер), который порождает для каждого слова некоторое множество вариантов разбора. После этого необходимо вручную проверить и исправить ошибки разбора, удалить неправильные варианты и добавить недостающие.

2.3. Элементы текста, требующие специальной обработки

1) Тексты/фрагменты на иностранных языках

Фрагмент на иностранном языке должен быть размечен при помощи специального тега, при этом должен быть определен язык фрагмента (латинский, немецкий, французский, греческий и др.). В идеале каждое слово должно быть приведено к словарной форме, для этого необходимо найти морфологический парсер для соответствующего языка. При этом приходится решать сложные филологические проблемы, связанные с тем, что орфография текста отличается от современной или непоследовательна. В крайнем случае, иноязычные слова могут быть даны как простой список словоформ без приведения к лексеме.

2) Переводы

Многие иноязычные тексты имеют переводы на современный русский язык и образуют как бы корпус параллельных текстов. Для удобства использования корпуса желательно связать оригинальные тексты с их переводами хотя бы на уровне предложений.

3) Цитаты из других авторов

Цитаты должны быть размечены в тексте при помощи специального тега, чтобы отличить авторский текст от заимствованного. Цитаты могут быть выделены по формальным признакам (курсив, кавычки и т.д.), но эти признаки довольно расплывчаты, поэтому требуется ручная разметка.

Заметим, что не всегда возможно отличить точную цитату от неточной, измененной автором при цитировании. В таком случае действует презумпция, что цитата считается авторским текстом, если не доказано обратное.

4) Примеры употребления (слова, фразы в грамматике)

Примеры употребления необходимо выделить при помощи специального тега, чтобы отличить от нормального употребления слова. Такие фрагменты похожи на цитаты, только взятые не из конкретного текста, а из некоторой модели языка. Формально примеры обычно выделяются курсивом, как и цитаты.

Семантическое различие между языковыми примерами и основным текстом достаточно очевидно даже для редакторов академического собрания. Так, в «Российской грамматике» в примерах сохраняется оригинальная орфография (с ятем, ером и т. д.), тогда как в основном тексте орфография модернизирована.

5) Фрагменты слов (слоги, буквы в грамматике)

Такие фрагменты очень похожи на примеры: они также оформляются курсивом и в них часто сохраняется оригинальная орфография. Однако, в от-

личие от примеров, такие фрагменты не считаются словами, а выделяются в отдельное подмножество (не-слова), аналогично формулам и цифрам. Проблема в том, что фрагмент слова может быть омонимичен нормальной лексеме (-у, -а, -ой, раз-), и поэтому должен быть специально размечен, чтобы случайно не попасть в список лексем.

6) Сокращения

Сокращения должны быть раскрываться, если этого возможно, например:

м. г. => м<илостивый> г<осударь>
муж. р. => муж<ескаго?> р<ода>
Е. И. В. => Е<го> И<мператорское> В<еличество>

Если раскрытие неоднозначно, после него ставится знак вопроса.

Некоторые сокращения уже раскрыты в редакторском тексте при помощи угловых скобок. Поэтому для разметки корпуса нужно выбрать другой знак, отличный от редакторских символов.

Сокращения представляют большую проблему для парсера и грамматической разметки. Парсер не может разобрать неполные или разорванные слова или порождает нелепые разборы, которые в любом случае приходится исправлять вручную. Не всегда возможно однозначно восстановить полную форму и привести ее к лексеме. В корпусе такое словоупотребление должно иметь признак, что форма сокращенная и тем самым недостоверная.

7) Цифровые комплексы

Цифровые комплексы обычно не разбираются и не заменяются на словесную запись, за исключением тех случаев, когда содержат фрагменты флексий (в 1754-м году, в 1-ой части). Это обусловлено тем, что не всегда возможно восстановить словесную форму числа, записанного цифрами. Все цифровые комплексы собираются в отдельное подмножество, аналогично фрагментам слов.

8) Формулы (в физике, химии, математике).

Формулы, переменные и другие элементы научной нотации должны размечены при помощи специального тега, чтобы они не смешивались с нормальными словами. Формулы могут содержать фрагменты, совпадающие с обычными словами (a quadratus plus b quadratus), но не должны интерпретироваться как слова. Все формулы собираются в отдельное подмножество, аналогично фрагментам слов.

2.4. Грамматическая разметка

Морфологический анализатор (парсер) — программа, выполняющая грамматический разбор текста, который включает в себя следующие задачи:

1) **токенизация** — разбиение текста на элементарные знаки (токены) и определение типа для каждого токена: слово, знак препинания, цифровой комплекс, тег разметки и т.д.

2) **сегментация** текста на предложения (а также клаузулы и др. виды сегментов);

3) **морфологический анализ** для слов, присутствующих в грамматическом словаре;

4) построение гипотез для нераспознанных слов (если это возможно).

Парсер является достаточно универсальной программой и мало зависит от конкретного языка и словаря. Вся конкретно-языковая информация о словоизменении записывается во внешних файлах в специальном формате и включает в себя две таблицы — таблицу парадигм + грамматический словарь.

Грамматический словарь — список лексем языка с приписанной им информацией о словоизменении. Каждая лексема в словаре содержит, как минимум, следующую информацию:

1) основа с указанием чередований;
2) постоянные признаки лексемы (часть речи, род, одушевленность, переходность, и т. д.);
3) номер парадигмы.

Модель словоизменения для русского языка основана на «Грамматическом словаре русского языка» А. А. Зализняка, который представляет собой наиболее авторитетный стандарт в данной области. В процессе создания парсера оказалось, что описание словоизменения в данном словаре недостаточно формально для программной реализации, и его пришлось детализировать и формализовать. В частности, пришлось расширить номенклатуру парадигм, разработать специальную нотацию для описания чередований в основе, точно описать схемы чередований, и т.д. В целом электронный словарь для парсера представляет собой отдельный продукт, заметно отличающийся от печатного издания по структуре и составу информации.

Парсер анализирует каждую словоформу по отдельности, без учета синтаксического контекста, и приписывает ей множество вариантов разбора (которое может быть пустым). Каждый вариант разбора содержит следующую информацию:

1) лексема (словарная форма);
2) грамматические признаки лексемы (часть речи, род, одушевленность, переходность);
3) грамматические признаки словоформы (число, падеж, наклонение, время, лицо);
4) номер парадигмы.

Основные проблемы при морфологическом анализе таковы:

Парсер дает только предварительную грамматическую разметку, которую необходимо проверить и исправить вручную. Основные проблемы, которые приходится решать при ручной обработке, таковы:

1) омонимия (грамматическая и лексическая), при которой словоформа получает несколько вариантов разбора;

2) отсутствие разбора, если лексема отсутствует в словаре или имеет нестандартную форму;
3) неправильный разбор.

Основная задача — устранить межлексемную омонимию, чтобы сгруппировать словоформы в словарные статьи. Что касается внутрилексемной омонимии, то устранить ее очень трудно, поскольку в русском склонении много омонимичных форм. Поэтому было принято решение, что грамматические формы внутри лексемы не размечаются и не различаются (к дому=в дому), за исключением особых случаев.

В процессе обработки текста приходится решать сложные лингвистические проблемы, связанные с различиями между языком эпохи Ломоносова и современным русским языком. Ломоносов широко использует церковнославянские формы и слова, отсутствующие в современном языке. Орфография текстов также весьма разнообразна и непоследовательна. Чтобы улучшить качество распознавания, необходима ручная настройка словаря и парсера, а также использование эвристических приемов и построение гипотез по аналогии.

3. Структура словарной статьи

3.1. Основные понятия

Лексема — множество словоформ с одинаковым лексическим значением.

Словоформа — вариант лексемы, имеющий определенное грамматическое значение.

Словоупотребление — конкретная словоформа в конкретном месте в корпусе текстов.

Словарная статья включает следующие основные зоны:

1) Заголовочное слово.
2) Грамматические пометы (часть речи, род, вид, переходность).
3) Краткая дефиниция (при необходимости).
4) Суммарная частота по всем текстам (возможно разбиение по типам/жанрам).
5) Примеры употребления с адресами и гиперссылками.

3.2. Заголовочное слово

Заголовочное слово представляет все варианты данной лексемы в компактном виде. Форма заголовочного слова выбирается по общепринятым правилам: для глаголов — инфинитив, для существительных — им. падеж ед. числа, для прилагательных — им. падеж ед. числа муж. рода и т. д.

В случае омонимии к заглавной форме добавляются цифровые индексы, в основном соответствующие словарю А.А. Зализняка, например: град[1] [осадки] — град[2] [город], мир[1] [спокойствие] — мир[2] [вселенная]. Индексы добавляются также в том случае, если омонимы имеют различные грамматические признаки (часть речи), например: знать[1] *с.ж.неод.* — знать[2] *г.нсв.* — знать[3] *вводн.* Многозначные слова при необходимости также могут быть разделены на подзначения, например: свет[1] — свет[2] [мир, бомонд], двор[1] — двор[2] [окружение монарха].

При группировке словоформ в словарную статью возникает много нетривиальных проблем, обусловленных широкой вариативностью исходного материала, в частности:

- какие формы можно считать вариантами одной лексемы?
- как удобнее группировать варианты?
- какой вариант выбрать в качестве основного (заголовочного слова)?

Исходя из соображений удобства, мы приняли следующие практические правила.

Орфографические и морфологические варианты по возможности нужно объединять в одну статью: знание=знанье, вариант=варьянт, кресло=кресла, зал=зала=зало.

Форма заголовочного слова должна включать все варианты: знание, знанье; или основной вариант и другие варианты в скобках: знание (знанье). В качестве основного варианта выбирается самый частотный или совпадающий с современным (для удобства пользователя).

Слова, которые могут писаться слитно и раздельно, желательно считать единой лексемой, а не разбивать на отдельные слова: вслед=в след, наверное=на верное, также=так же.

Исключениями из этого правила являются:

- частица не с финитными формами глагола: незнает = не+знает;
- клитики ж(е), ли(ль), б(ы): тыж = ты+же, онже = он+же, егоже = его+же или егоже (от иже), тыль = ты+ль, яб = я+б.

3.3. Грамматические признаки лексемы

Каждая лексема имеет признак «грамматический класс» (часть речи), который в основном соответствует словарю Зализняка и грамматической традиции. Далее указываются грамматические признаки, характерные для данного класса лексем: для глаголов — вид, для существительных — род и одушевленность, а также число для singularia и pluralia tantum. Грамматические признаки записываются с использованием достаточно понятных сокращений.

Группировка грамматических форм по лексемам соответствует грамматической традиции. Так, причастия и деепричастия включаются в парадигму глагола наравне с финитными формами. Степени сравнения включаются в парадигму

соответствующих прилагательных и наречий, кроме некоторых специфических форм (*больше, меньше, дальше*), которые считаются отдельными наречиями.

По практическим соображениям в ряде случаев мы решили игнорировать незначительные смысловые и синтаксические различия между парами омонимичных (полисемичных?) слов, которые в традиционной грамматике относятся к разным грамматическим классам, например:

1) многие субстантивированные прилагательные (*знакомый, приезжий, русский*) не отличаются от исходных прилагательных, поскольку они образованы по регулярной семантической модели;
2) не различаются мелкие оттенки значения для некоторых неизменяемых слов, например: *еще, уже* — наречие и частица, *ли, разве, даже* — частица и союз, и др.
3) прилагательные, не отличающиеся по смыслу от соответствующих причастий (*открытый, одетый, занятый* и т.д.), обычно считаются причастиями и включаются в парадигму соответствующего глагола.

3.4. Краткая дефиниция

Конкорданс не является толковым словарем и ориентирован на человека, владеющего русским языком. Общеязыковые слова в конкордансе даются без дефиниций и без филиации значений, поскольку эту информацию легко получить из любого толкового словаря.

Краткая дефиниция используется только для объяснения редких и устаревших слов, специальных терминов, а также для различения омонимов типа град[1] [осадки] — град[2] [город]. При необходимости дефиниция может быть расширена или дополнена ссылкой на словарь. Основные источники дефиниций для неизвестных слов:

- списки устаревших слов и комментарии в Собрании сочинений Ломоносова;
- Словарь русского языка XVIII века;
- Словарь Академии Российской.

Краткая дефиниция не пытается подменить полноценную энциклопедическую статью, а служит только для получения общего представления об описываемом предмете. Конкорданс предназначен для исследования языка, а не истории реалий.

3.5. Частота

Указывается абсолютная частота лексемы, т. е. количество всех ее употреблений во всех рассматриваемых текстах. При необходимости дается разбиение по основным типам/жанрам текстов.

3.6. Примеры употребления

Включает полный список всех употреблений данной лексемы, включая высокочастотные слова (предлоги, союзы). Каждое словоупотребление включает следующий набор признаков.

1) Контекст

Контекст — это связный фрагмент исходного текста, включающий данную словоформу и достаточный для понимания ее смысла и синтаксического окружения. Обычно контекст представляет собой целое предложение или клаузулу (связная часть сложного предложения), а при необходимости расширяется за указанные пределы. Сама словоформа выделяется жирным шрифтом.

2) Адрес

Адрес данного словоупотребления в корпусе, достаточный для его идентификации при цитировании и предназначенный для человека. Адрес включает в себя:

1) короткое имя произведения;

2) внутренний адрес — названия или номера явно выраженных структурных элементов текста (действие, явление, реплика, ремарка, заголовок) или меток (номер страницы, стиха).

Для драматических произведений внутренний адрес включает в себя: номер действия, номер явления, имя говорящего.

Для поэтических произведений внутренний адрес включает номер строфы и стиха.

При отсутствии структуры внутренний адрес может включать в себя номер страницы.

Адрес не всегда обеспечивает однозначную идентификацию текстового фрагмента (может быть несколько реплик одного лица на одной странице, ремарки не имеют имени), но вполне достаточен для цитирования и согласуется с принятой традицией.

3) Ссылка на текст

Прямая гипертекстовая ссылка на соответствующее место в корпусе текстов.

Предполагается, что в тексте все предложения пронумерованы независимо от структурной разметки и имеют метки, на которые можно сразу перейти.

Прямая ссылка на текст необходима, чтобы получить более широкий контекст для данного словоупотребления, если пример в конкордансе недостаточно информативен.

4) Тип фрагмента

Тип текстового фрагмента, присвоенный при первоначальной разметке, например:

- заголовок структурного элемента (часть, глава, реплика);
- ремарка;
- сноска;

- цитата;
- пример употребления слова;
- тип речи (проза/стихи);
- подпись под рисунком, и т.д.

Разные типы текстовых фрагментов имеют языковые особенности, которые должны быть предметом специального изучения.

4. Возможности использования конкорданса

Размеченный конкорданс представляет собой словарную базу данных, из которой путем различной проекции и группировки данных можно получать различные виды словарей и проводить объективные исследования авторского языка.

Форма базы данных открывает целый ряд возможностей, недоступных в традиционных бумажных словарях.

- динамический выбор примеров по любым параметрам,
- динамическая сортировка и группировка,
- быстрый переход из словаря в корпус текстов,
- просмотр и выдача словарной информации в различных форматах,
- генерация печатных словарей.

4.1. Динамический выбор примеров

Пользователь может создать для себя рабочее подмножество (проекцию) корпуса по любым текстовым и метатекстовым параметрам:

- жанр и тематика текста,
- дата написания,
- название произведения,
- тип текстового фрагмента (заголовок, сноска, цитата, проза/стихи),
- класс лексической единицы (русское/латинское/немецкое слово, цифры, фрагменты),
- грамматические признаки лексемы (часть речи, род, вид, переходность),
- контекст (соседние слова),
  и т. д.

4.2. Динамическая сортировка

Стандартной формой представления информации является алфавитно-частотный конкорданс, где лексемы отсортированы в алфавитном порядке, а примеры внутри статьи — по словоформе. Эта форма выбрана как наиболее нейтральная и понятная. На самом деле примеры внутри статьи могут сортироваться по любым существующим параметрам, например:

- метатекстовые параметры (жанр, автор, название, дата),
- тип текстового фрагмента,
- контекст, т.е. слово справа/слева (так называемый KWIC — keyword in context).

При необходимости можно сделать сортировку по нескольким параметрам, например: словоформа + контекст+жанр, жанр+словоформа+слово справа, и т.д. Видимо, наиболее удобной является сортировка примеров по признаку словоформа+контекст (слово справа+слово слева).

4.3. Группировка

Путем различной сортировки и группировки данных из одной и той же словарной базы можно получить следующие виды словарей:

- алфавитный конкорданс, где лексемы отсортированы в алфавитном порядке, а словоупотребления внутри статьи — по грамматической форме или по KWIC;
- частотный словарь, где лексемы сгруппированы в порядке убывания частоты,
- обратный алфавитный словарь,
- грамматический словарь, где лексемы сгруппированы по грамматическим признакам,
- словари отдельных произведений или типов речи, и т. д.

Кроме того, исследователь всегда сможет получить нужную ему проекцию словаря по запросу. Возможность динамического получения новых видов словарей (в том числе, не предусмотренных первоначальным замыслом) является совершенно новым словом в практике филологической работы.