Многоязычная библиотека Диглосса (diglossa.ru) и

морфологический анализатор Морфей

Библиотека Диглосса (diglossa.ru)

Электронные библиотеки, существующие в настоящее время, бывают либо коммерческие, либо созданные академическими институтами, например Национальный корпус русского языка, Персей, либо созданные сообществом интернет-пользователей, например, библиотека Мошкова. В последнем случае это обычно просто свалка текстов неясного качества с неясными авторскими правами.

Диглосса стремится объединить в себе достоинства обоих подходов. Она стремится иметь все возможности академических корпусов текстов, оставшись проектом сообщества. Тот же принцип использует, например, Википедия. Это достигается применением простых и ясных, но жестких и обязательных правил-договоренностей. Вот некоторые из них:

Все тексты Диглоссы хранятся в системе контроля версий git в виде плоских фалов, и разбиты на абзацы, соответствующие друг другу по смыслу (сейчас это выполняется вручную). Это важно - исходной формой текста является простейшая форма, собственно текст, лишенный какого-либо оформления и какой-либо логической разметки. Все преобразования происходят на лету, в момент загрузки текста в систему, так, как это нужно в данный момент и для данной цели.

Это позволяет рассматривать Диглоссу как часть большой распределенной библиотеки, использующей общие исходные авторитетные тексты для своих конкретных целей.

Сейчас, например, происходит преобразование текста в структуру текст-абзац-предложение-слово, с возможностью подключения либо морфологического анализатора, либо воспроизведения звука. Но повторюсь, преобразование может быть любым и диктоваться конкретной целью.

Диглосса построена с использованием технологий, применяемых при разработке проектов со свободными лицензиями, например GNU GPL. Она сама имеет эту же лицензию, ее исходные коды открыты и доступны на Гитхабе, и использует ту же культуру разработки. Все тексты и коды могут быть легко клонированы либо как тексты в git (или любой системы контроля версий), либо как записи базы данных CouchDB (или любой No-SQL DB).

Диглосса полностью соответствует стандарту html5, и является приложением одной страницы. Страница никогда не перегружается, лишь подкачивается необходимая информация, при этом все страницы полностью индексируются любой поисковой машиной.

Переводы (правая страница) переключаются либо целиком, либо внутри одного абзаца, чтобы тексты переводов можно было удобно сравнить (используется shift-колесо мыши).

Каждое слово исходного текста (левая страница) является ссылкой, либо на справку морфологического анализатора Морфей, либо на воспроизведение звука данного слова.

Простой API позволяет выводить результаты - тексты, абзацы, предложения и слова - в любом необходимом формате.

Морфологический анализатор Морфей

Morpheus — (произносится с ударением на Eu) — простой морфологический анализатор, применяемый в многоязычной библиотеке diglossa.ru. Он разрабатывается с применением стандартных прицинов разработки веб-приложений и имеет открытую лицензию GNU GPL. А именно, он построен как набор утилит, которые могут быть объединены в цепочки, ведущие к необходимому результату. Например, сейчас для древнегреческого языка есть только утилиты, обрабатывающие словарь, утилиты исключениий и сами исключения (самые распространенные). А для латыни - словарь, многие исключения, и обработка более ста парадигм. Для санскрита - лишь начальные утилиты обработки словаря. Ключевые слова при разработке - agile, bbd-style. Morpheus включает в себя более 3000 тестов-спецификаций. Для генерации (латинских) тестов использовалась программа Уильяма Уитеккера "words".

В качестве образца для подражания при разработке я использую Natural Language Toolkit, используя, однако, Руби, а не Питон в качестве основного языка разработки. Все данные о парадигмах языка хранятся в JSON, открыты и легко читаются и модифицируются человеком. Обрабатываются не строки, а ruby (на сервере) или Javascript (на клиенте) объекты.

Алгоритм работы простого морфологического анализатора таков: Каждая словоформа проходит через сито парадигм и связанных с ним правил, и на первом шаге 1) выявляются парадигмы, способные породить данную словоформу и возможные словарные формы, затем на втором 2) выбираются только те парадигмы, которые порождают действительно существующее в словаре слово и, наконец, 3) результат кешируется в БД. Вдобавок есть механизм заполнения базы словоформ исключениями, (терминами, не требующими никакого анализа, вне системы парадигм.). Все происходит в NoSQL-БД (поскольку json), а именно в CouchDB.

Морфей превращает Диглоссу из свалки текстов в корпус. Все тексты с помощью простого API могут быть преобразованы в любой стандартный формат, используемый сегодня в корпусной лингвистике. Сейчас я использую в качестве образца форматы, описываемые в NLTK, с соответствующими изменениями, диктуемыми флективной природой используемых языков.

Волонтеры и сообщество

Ни Морфей, ни Диглосса не ставят себе какие-либо научные, академические цели. Они предназначены для помощи тем, кто читает классические тексты, постоянно заглядывая в источник. Я рассчитываю найти и связаться со всеми группами на доступной мне территории, читающими тексты таким образом, и предложить им использовать Диглоссу в работе. Их, однако, можно использовать как любой другой корпус текстов для любой академической цели.

Применяемые принципы позволяют легко наращивать и количество текстов в Диглоссе, и количество языков, используемых в Морфее. В настоящее время есть волонтеры по обоим направлениям. Но я, однако, рассчитываю на более активное участие сообщества в разработке.