Ведущая рубрики
Варвара Андреевна ОСИПОВА
редактор сайта Российской ассоциации электронных библиотек

БИБЛИОТЕКА ДЛЯ МЕДЛЕННОГО ЧТЕНИЯ

«Диглосса» – это многоязычная электронная библиотека, предоставляющая возможность читать философские тексты на русском языке и сравнивать их с оригиналом. Интерфейс библиотеки построен так, что на одном уровне параллельно друг другу в левой и правой частях экрана располагаются одинаковые смысловые абзацы книги в оригинале и в переводе, вместе с тем поддерживается функция морфологического анализа и перевода слов, а также озвучивание некоторых текстов.

В настоящее время библиотека содержит всего несколько десятков текстов. Это Бхагавит-гита на санскрите; произведения Гомера, Аристотеля, Фукидида, Платона и Прокла на древнегреческом; труды святых Ионна, Григория Паламы, Матфея, Марка и Луки на греческом койне; книги Эразма Роттердамского, Публия Корнелия Тацита, святого Фомы Аквинского, Августина Блаженного, Гая Светония Транквилла, Публия Овидия Назона, Петра Абеляра, Бенедикта Спинозы и Гая Юлия Цезаря на латыни; «Государь» Николы Макиавелли и «Божественная комедия» Данте Алигьери на итальянском; трактаты Ницше, Хайдеггера, Канта и Гегеля на немецком и произведения Марселя Пруста и Рене Декарта на французском.

Библиотеку представляет создатель ресурса – **Михаил Владимирович БЫКОВ** – веб-разработчик, член Российской ассоциации электронных библиотек.

Автор

Михаил Владимирович БЫКОВ
руководитель электронной библиотеки diglossa.ru

ЦЕЛЬ БИБЛИОТЕКИ

Электронная библиотека «Диглосса» (diglossa.ru) не ставит перед собой каких-либо научных, академических целей. Она предназначена для помощи тем, кто читает классические тексты, постоянно заглядывая в источник. «Диглосса» — библиотека для тех, кто читает медленно, почти по слогам. И даже просто по слогам: вряд ли можно читать Платона и Аристотеля в оригинале быстрее. Вдобавок, это традиция. Именно так и читались тексты во всех средневековых университетах, византийских школах, от Пифагора и до распространения книгопечатания. Идея «Диглоссы» родилась на семинаре А.В. Ахутина (см. bibler.ru), где тексты Платона и Канта читались хотя и по-русски, но с постоянным заглядыванием в источник, по паре абзацев текста за занятие.

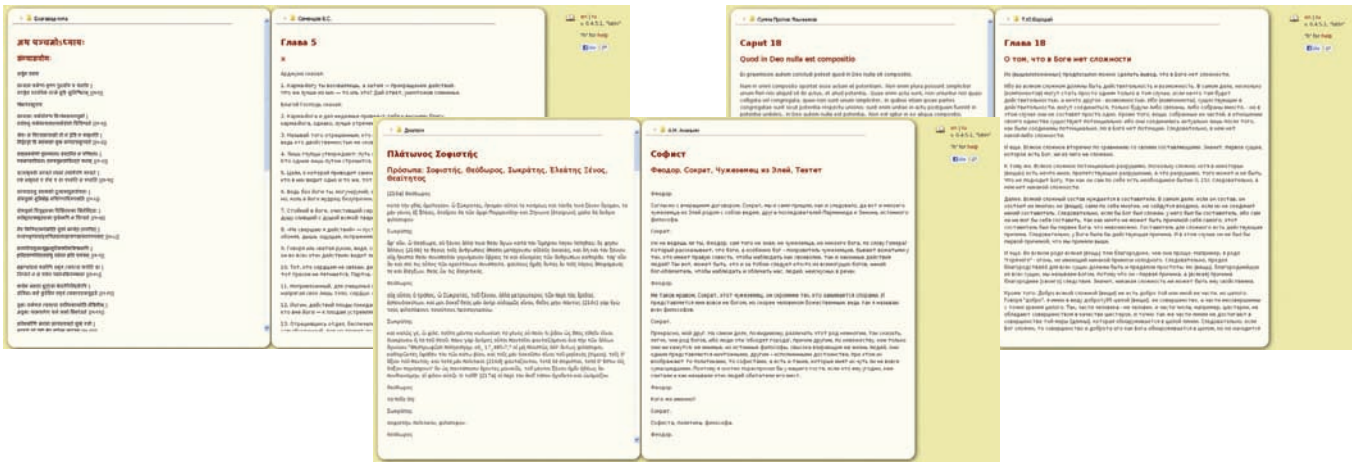
«Диглосса» предназначена для того, чтобы учиться у авторов-философов. Сейчас на сайте diglossa.ru несколько сот посетителей в день, но всего несколько из них, пять или семь, остаются на сайте дольше получаса, т.е. читают тексты. Этого пока очень мало. Библиотеку, тем не менее, можно использовать как любой другой корпус текстов для любой академической цели, поскольку она имеет для этого все возможности.

ОСНОВНЫЕ ПРИНЦИПЫ БИБЛИОТЕКИ

«Диглосса» использует несколько характерных принципов:

- тексты хранятся в «плоских» файлах (.txt), в юникоде utf-8, в системе контроля версий git[1];
- оригинальные и переводные тексты разбиты на абзацы, соответствующие друг другу по смыслу;
- используется морфологический анализатор слов «Морфей»;

1. Git – распределённая система управления версиями файлов. Проект был создан для управления разработкой ядра Linux; примером проектов, использующих Git, является Android.



- «Диглосса» является распределённой библиотекой, сообществом;
- все коды и тексты публикуются под открытой лицензией GNU GPL[2], используются стандартные методы разработки.

ИСТОЧНИК – «ПЛОСКИЙ ТЕКСТ»

Тексты хранятся в виде «плоских» файлов. Это нетривиально: все современные сетевые ресурсы устареют, если не будут постоянно модифицироваться. Посмотрите на сайты пятилетней давности: устаревают даже самые мощные. А чтобы сделать версию для планшета, нужно опять получить немалое финансирование. Рано или поздно на очередном витке финансирования не будет, и работа уйдёт в никуда.

В «Диглоссе» все преобразования происходят на лету, в момент загрузки текста в систему, так, как это нужно в данный момент и для данной цели. Тексты доступны в формате .txt, хотя читатель на сайте видит довольно сложную картинку. На другом сайте, или на планшете, или в телефоне он увидит другую картинку, но текст будет тот же самый. Сейчас, например, происходит преобразование текста в структуру «текст — абзац — предложение — слово» с возможностью подключения либо морфологического анализатора, либо воспроизведения звука. Тот же текст в иной веб-библиотеке, на ином ресурсе может обрабатываться совершенно по-иному.

Это важно — исходной формой текста является простейшая форма, собственно текст, лишённый какого-либо оформления и какой-либо логической разметки, или имеющий минимально необходимую. Поэтому текст остаётся модифицируемым, поддерживаемым, его легко обновить, заметив недостаток, дополнить, исправить опечатку. Всё остальное сделает система — и для читателя, и для редактора процесс полностью прозрачен. Это очень эффективный подход.

Хранение текста в системе контроля версий git автоматически сохраняет историю изменений текста. Видно, кто и когда создал данный текст, кто и зачем внёс исправления. Можно принять изменения, можно отказаться от них в пользу предыдущей версии в случае ошибки. Текст автоматически становится ресурсом для коллективной работы, авторизованным вплоть до изменения любой буквы.

Хранение «плоских текстов» — не догма. Например, в диалогах Платона есть короткие строки — имена участников, которым принадлежат реплики. Их хорошо бы выделить, например цветом. Этот принцип, однако, сложно выдержать строго. Иногда преобразования слишком сложны, и создание преобразователя трудоёмко. Например, на «Диглоссе» есть словарь хайдеггеровских терминов. Он очень непрост, и он единственный в своём роде. Намного удобнее хранить его в специальной логической разметке, нежели писать очень сложный парсер (от англ. *parser* — это программа или алгоритм, осуществляющие грамматический разбор) плоского текста — причём для единственного файла.

В общем, хранение плоских текстов и преобразование их на лету — это скорее элемент общей культуры работы с текстом, нежели технический принцип.

ПАРАЛЛЕЛЬНЫЕ АБЗАЦЫ

Все тексты «Диглоссы» хранятся в виде файлов, разбитых на абзацы, соответствующие друг другу по смыслу (сейчас это выполняется вручную). С повышением качества переводов, когда появится надежда, что внутри одного абзаца количество предложений источника и перевода будет равным, можно будет отслеживать соответствие не абзаца, а предложения. Сейчас это возможно не для 100% текстов.

МОРФОЛОГИЧЕСКИЙ АНАЛИЗАТОР MORPHEUS

Morphéus — простой морфологический анализатор, который разрабатывается с применением стандартных принципов разработки веб-приложений и имеет открытую лицензию GNU GPL. 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 исключений и сами исключения (самые ▶

2. GNU General Public License (Универсальная общественная лицензия GNU) – лицензия на свободное программное обеспечение, созданная в рамках проекта GNU в 1988 г. Цель GNU GPL – предоставить пользователю права копировать, модифицировать и распространять (в том числе на коммерческой основе) программы, а также гарантировать, что и пользователи всех производных программ получат вышеперечисленные права.

распространённые). А для латыни — словарь, многие исключения, и обработка более ста парадигм. Для санскрита — лишь начальные утилиты обработки словаря. Morpheus включает в себя более 3000 тестов-спецификаций. Для генерации латинских текстов использовалась программа Уильяма Уитеккера Words.

Алгоритм работы простого морфологического анализатора таков: каждая словоформа проходит через сито парадигм и связанных с ним правил, и на первом шаге выявляются парадигмы, способные породить данную словоформу и возможные словарные формы, на втором выбираются только те парадигмы, которые порождают действительно существующее в словаре слово и, наконец, результат кешируется в базе данных. Вдобавок есть механизм заполнения базы словоформ исключениями (терминами, не требующими никакого анализа, вне системы парадигм). Таким образом Morpheus превращает «Диглоссу» из свалки текстов в корпус. Все тексты могут быть преобразованы в любой стандартный формат, используемый сегодня в корпусной лингвистике.

СТАНДАРТНЫЕ ПРОСТЫЕ ТЕХНОЛОГИИ

«Диглосса» построена с использованием технологий, применяемых при разработке проектов со свободными лицензиями, например GNU GPL, и использует ту же культуру разработки. Библиотека 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.

Библиотека построена по принципу «как можно проще». Использование форматов XML, RDF, Java, Oracle в 90% лингвистических и филологических проектов излишне. Эти инструменты — слишком сложные и дорогие в поддержке и эксплуатации. А если вашу задачу можно решить «на коленке», используйте этот шанс. Ruby или Python вместо Java, NoSQL вместо Oracle, JSON вместо XML. Сложные же инструменты нужно использовать только там, где без них действительно никак нельзя обойтись.

ВОЛОНТЁРЫ И СООБЩЕСТВО

Хранение текстов в открытом доступе, свободное и удобное реплицирование базы данных и возможность создания любых API для передачи данных и связи приложений позволяют рассматривать «Диглоссу» как часть большой распределённой библиотеки, использующей общие исходные тексты для своих конкретных целей. Я хотел бы использовать и читать в библиотеке тексты, подготовленные компетентными людьми. По логике, если есть некоторая группа людей, которые занимаются, например, Аристотелем, то им удобно выложить свои наработки. Но проблема в другом: как их найти, и как понять, что тексты сделаны профессионально? Я когда-то видел в Сети текстовый файл с речами Лисия. Как теперь его найти? Как узнать, кто его сделал, когда и какие права использования?

В скором будущем читатель откроет страницу с текстом Аристотеля на своём ридере. И ему будет всё равно, откуда именно он в данный момент его получил. Но не всё равно, каково качество этого текста. В идеале, каждый текст должна готовить группа профессионалов, и готовить его постоянно и непрерывно, сменяя поколения публикаторов, добавляя комментарии, синхронизируя переводы, внося исправления, создавая справочный аппарат. Это типичная задача сетевого сообщества. Примеров успешного применения этой модели множество, их не счесть. А иного пути — например, рассчитывать на получение финансирования подготовки электронной формы того же Лисия — не существует, т.е. разовое финансирование возможно, но дело сделать таким путём нельзя.

«В идеале, каждый текст должна готовить группа профессионалов, и готовить его постоянно и непрерывно, сменяя поколения публикаторов, добавляя комментарии, синхронизируя переводы, внося исправления, создавая справочный аппарат. Это типичная задача сетевого сообщества»

Если возможность создания распределённой компетентной библиотеки существует, то она будет создана. Технических проблем объединения множества источников в единую библиотеку-подборку, любой компоновки и обработки, выбора необходимых именно вам текстов и даже фрагментов текстов не существует. Я хочу подчеркнуть, что говорю не о том, чтобы кто-то принял участие в самой «Диглоссе». Наша библиотека — открытый проект, и если кто-то захочет принять участие, я буду очень рад и окажу всяческое содействие в обучении и поддержку. Я имею в виду способ работы с текстами, который сам собой сложится в распределённую библиотеку, которая для читателя обернётся множеством разных и внешне не связанных ресурсов.

Я рассчитываю найти и связаться со всеми группами, читающими таким образом, и предложить им использовать «Диглоссу» в работе, добавить их тексты в нашу библиотеку или помочь создать подобный ресурс. Если вы занимаетесь медленным чтением с обращением к источнику, напишите, пожалуйста, мне: m.bykov@gmail.com. Тексты на китайском, японском, фарси, иврите, арабском, грузинском и армянском я сам подготовить не смогу, а предоставить их в доступ очень хочется. ■