diglossa.js - инструмент медленного чтения

в докладе описывается приложение для чтения и редактирования многоязычных текстов на основе нового формата е-книг, .dgl, подобного формату .epub, но построенному на основе markdown, а не html

описание проекта: http://diglossa.org/описание

diglossa.js - десктопное приложение для Window, MacOS, Linux. Мощная читалка и редактор структуры любых е-книг (.epub, .pdf, .fb2, .csv, .html, .md, .txt), в том числе е-книг на основе собственного формата .dgl. А также генератор многоязычных контекстных словарей. Позволяет читателю самому создавать dgl-книги, добавлять новые переводы и генерировать новые словари

diglossa.js позволяет подключать обычные словари, и генерирует свои многоязычные контекстные словари на основе dgl-книг. На сайте http://diglossa.org вы видите примеры подобных словарей, сайт может использоваться как справочник и контекстный онлайн-словарь. Здесь же вы можете загрузить примеры книг в формате .dgl (пока только в виде файла)

diglossa.js - бесплатна, имеет открытый код, свободную лицензию GNU GPL, модульную структуру и интуитивно ясный интерфейс

diglossa.js - построена на основе формата электронных книг .dgl - подобного .epub, но имеющего в основе не html, но markdown

Плагины позволяют использовать любой язык, помимо широко распространенных

основные возможности:

- импорт книг: .epub, .pdf, .fb2, .html, .md, .txt и - .dgl
  - импорт словарей - .sd, .dsl
  - автоматическое определение языка текста
  - вызов соответствующего языку словаря по alt-mouse-move
  - локальный и полнотекстовый поиск
  - закладки
- импорт и параллельное подключение переводов той же книги
- редактирование структуры книги
- автоматическая проверка синхронизации абзацев
- экспорт синхронизированных книг в формате .dgl (пока только в виде файла)
- генерация многоязычных контекстных фразовых словарей

markdown

dgl - формат электронных книг, подобный формату epub (см. http://idpf.org), но имеющий два отличия. Во-первых, он использует не html, но (псевдо)-markdown. Во-вторых, он не имеет специального файла оглавления .toc, но использует саму разметку markdown для построения оглавления. Во всем остальном .dgl копирует .epub, его можно назвать версией .epub-markdown, все возможности .epub сохранены.

псевдо-markdown - потому что из блоковых тегов используется только один - "<p>", или paragraph. Последовательность (псевдо) абзацев включает заголовки, абзацы, строки таблицы, строки списка, примечания и все подобное горизонтальное. Таким образом, текст книги превращается в последовательность абзацев. Благодаря этой особенности формата .dgl можно легко сопоставить два или более параллельных текста. В синхронизируемых текстах каждый абзац соответствует своей параллельной паре.

синхронизация и авто-синхронизатор

абзац может быть удален, скопирован, разбит на части, etc - см. соответствующий раздел документации. Однако, при синхронизации в текст книги исправления не вносятся, но создается синхронизатор. Это позволяет в любой момент откатить внесенные изменения, даже вернувшись к редактированию книги после редактирования или чтения иной книги, или рестарта диглоссы. авто-синхронизатор значительно ускоряет работу по проверке синхронизации абзацев подключаемой книги. при экспорте пакета книг их синхронизаторы теряются, и публикуется лишь окончательная форма книг.

Синхронизация абзацев книги - ручная работа. Но после синхронизации разделов, которая в любом случае должна выполняться вручную, возможно запустить автоматическую проверку синхронизации текстов. Реализованы два механизма авто-преверки синхронизации абзацев - сложный, с проверкой соответствия стемов абзаца, требующий коннекта с сетью и создания словарей соответствия стемов словоформ языковых пар. И простой, на основе сравнения длины строки абзаца, числа фраз и пр. формальных признаков. Оказалось, что простой метод дает неплохие результаты и вполне годен к применению в реальной практике. Собственно, авто-синхронизация есть убойная фича, делающая все вышеописанное реальным.

плагины

Диглосса группирует словоформы языка на основе стемов словоформ. Здесь стем - понятие не грамматики языка, а механизма обработки строк. Это просто подстрока, группирующая словоформы максимально плотно. А какие словоформы попадут в группу - не имеет значения. (Это не лингвистика, а обработка строк). В плагине реализуются два метода - анализатор и синтезатор. Анализатор в простейшей форме есть просто обычный стеммер. Всего могут быть три формы анализатора - a) стеммер, b) сложный стеммер, дающий два или несколько вариантов стема и c) полноценное разложение словоформы на все возможные цепочки последовательностей разбиения словоформы на подстроки. (На длинных словоформах это может быть весьма резурсозатратно). Затем происходит проверка наличия стемов или любых подстрок в словарях. Если метод a) не дает результата, стартует метод b), затем метод c). В конце синтезатор собирает полученные результаты проверки в компактную форму. Диглосса вызывает анализатор и отображает результат, полученый от синтезатора, не имея представления о том, что происходит в плагине.

ближайшие цели

- опубликовать пример плагина, и руководство по написанию плагинов для любого языка

- свести текст, параллельные тексты и справочный аппарат е-книги к единому, всегда актуальному источнику в системе контроля версий, автоматически клонируемому на компьютер читателя. (т.е. публикация в системе контроля версий, вдобавок к экспорту пакета в файл).

- дать возможность читателю самому добавлять новые параллельные переводы и справочные материалы к единому актуальному источнику. И, напротив, загружать себе лишь необходимые ему части е-книги. В том числе иметь возможность совмещать переводы и иные части книги, имеющие различные и несовместимые в обычной е-книге лицензии.

code & download

code: https://github.com/mbykov/diglossa.js

download: https://github.com/mbykov/diglossa.js/releases/latest

На момент публикации есть только пакеты для Linux, в том числе для Alt-Linux также.