Руководство по массовой загрузке словарных данных

Версия 3

# Общая информация

Массовая загрузка словарных данных осуществляется через CSV-файлы. Для удобства, фактически ввод может осуществляться через XLS-файл, который можно конвертировать в CSV-файл.

CSV-файл состоит, по меньшей мере, из пяти столбцов:

* лексемы (1й столбец)
* словоформы
* параметры словоформ
* переводы
* параметры переводов

# Лексемы

В первом столбце указывается часть речи, соответствующая лексеме. Одновременно наличие такой записи является маркером новой лексемы: до тех пор, пока в этом столбце не будет встречено следующее указание части речи, все строки таблицы будут относиться к одной лексеме.

Под лексемой в данном случае следует понимать слово со всеми его словоформами и вариантами (разными формами и произношениями), имеющей некоторый набор смыслов. При этом одной лексемой окажутся только те слова с разным смыслом, которые являются полными омонимами, то есть ни совпадение основной формы слова, ни омографичность всех форм, не являются достаточным условием. Например, не будут одной лексемой слова, имеющие разные:

* части речи («а» как союз и «а» как междометие);
* ударения (например, «замо́к» и «за́мок»);
* способы словоизменения (именительный: «ласка», родительный: «ласк» (действие) против «ласок» (животное)).

Формат заполнения первого столбца: *часть речи [классификатор словоизменения]*.

Часть речи нужно указывать принятым обозначениям: желательно английским, можно и русским или финским. При загрузке каждому обозначению будет поставлен в соответствие одна из частей речи, имеющаяся в базе словаря, поэтому необходимо и достаточно обеспечение единообразия обозначений внутри одного CSV-файла. Например, если для обозначения существительного используется «n», при загрузке единожды можно будет выбрать, что это означает «существительное» и все слова, для которых указано «n», будут загружены как существительные.

Для классификатора словоизменения (например, номер спряжения или склонения) на данный момент нет ограничения на значения.

# Словоформы переводимого языка

Во второй и, при необходимости, последующие столбцы вносятся слова переводимого языка. Слова вводятся в виде словоформ, соответствующих лексеме, обозначенной в первом столбце. В заголовке столбца указываются параметры «по умолчанию» для словоформ, вносимых в столбец.

Формат заголовка столбцов: *язык (диалект) [система записи @ источник данных]*.

Указанные язык и диалект автоматически проставляются для словоформ в столбце. Для отдельных слов возможно указание другого диалекта, отличного от «основного» (см. ниже).

Указанные системы записи и источник данных также используются для каждой словоформы в столбце и не могут быть переопределены (что означает, что для ввода словоформ в другой системе записи или из другого источника нужно добавить отдельный столбец).

Указание системы записи позволяет определить, в частности, представление словоформы в словаре: фонетическая транскрипция (например, МФА), фонематическая транскрипция (например, упрощённая учебная транскрипция) или орфографическая запись.

Источник данных — это публикация или иной источник, откуда фактически получены словоформы, вводимые в столбец. При этом если словоформы были преобразованы в систему письма, отличную от используемой в оригинальном источнике, следует указать это явным образом, например: «Нирви (преобразован в МФА)». Не следует путать это с указанием источника самих переводов: в общем случае, источников словоформ может быть несколько, или вообще ни одного, поэтому источник перевода задаётся непосредственно на этапе загрузки.

Способ обозначения языков, диалектов, систем записи и источников — произвольный, поскольку при загрузке файла эти параметры выставляются вручную (т.е. они нужны только для человека, который будет загружать файл). Все параметры, кроме наименования языка — необязательны.

Для каждой лексемы переводимого языка, указанной в первом столбце, необходимо перечислить словоформы и варианты, ей соответствующие. Если для лексемы вводится сразу несколько словоформ и вариантов, их необходимо разделить символом «|». Для каждой из них можно задать комментарий (любой текст в кавычках) и один или несколько параметров (в квадратных скобках). Поддерживается задание следующих параметров:

* грамматическую категорию (падеж, число, род и т.д.);
* диалект («перекрывает» «основной» диалект, чтобы не добавлять столбцы под каждый диалект);
* указание на конкретного информанта (для приведения идиолектных примеров произношения).

Формат ввода параметров:

Аналогично части речи, значения параметров должны быть единообразны внутри одного CSV-файла. Например, если для обозначения нижнелужского диалекта можно использовать «Ala-L».

# Слова целевых языков (переводы)

В целом ввод слов целевых языков соответствует вводу словоформ переводимого языка, но есть и принципиальные отличия.

Затем для каждой лексемы необходимо перечислить все переводы, причём каждый перевод указывается в новой строке. Лексему и словоформы при этом повторять не нужно.

В последнем столбце для перевода может быть указан комментарий или (в квадратных скобках) указано ограничение. Ограничение может быть диалектное (т.е. такой перевод актуален только для определённого диалекта переводимого языка) или смысловое. Пример «смыслового» ограничения: <http://slovari.yandex.ru/кошка/перевод/>. Тут ограничением будет «(для лазания на столбы)» или «(электрического крана)».

В итоге пришёл к следующему способу указания перевода: группируем в одной строке близкие по смыслу переводы, как обычно делают в словарях.

Вот, например:

http://slovari.yandex.ru/order/en-ru

Здесь как раз такие группы под номерами и перечислены. К группе может быть приведена "тема" (см. напр., 10, 11, 14, 15 и др.) и комментарий (напр., 11). Группа — необязательная сущность, она как бы сверху накладывается на несколько переводов.

К отдельным переводам тоже может быть добавлен комментарий (яркиц пример — 33). Ну и диалектное ограничение (напр., 7 и др.).

Я, честно говоря, так и не продумал пару технических мелочей, но практически уверен, что для целей ввода данных можно принять следующий формат ячейки с переводами:

[тема] [диалект] "комментарий группы": перевод1 [диалект] "комментарий перевода" | перевод2 [диалект] "комментарий перевода"

Пример из order, 11:

[рел.] "(orders)": духовный сан

Пример из order, 11:

[амер.]: порция | заказ "(в ресторане)"

Пример из bill, 1 (там же):

[брит.]: счёт

Пример обратного перевода для «счёт» (из головы):

bill [брит.] | check [амер.]

Например, если комментарий очень большой, ставить вместо него "\*1", а в последнем столбце вводить текст этого комментария. В формате "\*1: текст комментария". Соответственно, если в одной строке вдруг много больших комментариев (что само по себе плохо уже), то делать \*2, \*3 и т.д.

Подчеркну, что комментарии - это не примеры перевода и не пояснения к конкретным случаям употребления (что-то типа "в тот вечер, когда информант называл это слово, он только вернулся из Таллинна, поэтому, возможно, это эстонизм") -

это всё отдельные сущности, которые мы будем вводить позже.

переводы, примеры использования переводов и примеры применения слов информантами - это три разных вещи