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.

Для каждого источника переводов следует использовать свой файл. Источник переводов — это публикация или иной источник, откуда получена информация по соответствию слов (лексем) переводимого языка словам (лексемам) других языков. Таким образом не допускается в один файл вносить переводы, например, из разных словарей.

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

* лексемы
* словоформы
* переводы

# 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 [параметры лексемы] [классификатор словоизменения]

Часть речи (синтактическую категорию) нужно указывать принятым обозначением: желательно английским, но можно и русским или финским и т. д. 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.

Поддерживаются следующие параметры лексемы:

* pl — указание того, что лексема употребляется только во множественном числе.

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.

Формат ввода словоформы:

словоформа "комментарий" [параметр]

Для словоформы можно задать комментарий (любой текст в кавычках) и один или несколько параметров (каждый в отдельной паре квадратных скобок, порядок следования не важен). Поддерживается задание следующих параметров:

* грамматическую категорию (падеж, число, род и т.д.);
* диалект («перекрывает» «основной» диалект, чтобы не добавлять столбцы под каждый диалект) — означает, что данная словоформа является усреднённой для этого диалекта;
* указание на конкретного информанта — означает, что данная словоформа является идиолектным примером произношения.

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

В комментарии к словоформе может быть приведены, например, обстоятельства произношения (для примеров от информантов).

Если для лексемы вводится сразу несколько словоформ или вариантов, их необходимо разделить символом «|».

Таким образом, с полным набором параметров это может выглядеть так:

словоформа1 "комментарий" [грамматическая\_категория] [диалект] [информант] | словоформаN "комментарий" [грамматическая\_категория] [диалект] [информант]

# Слова целевых языков (переводы)

В оставшихся столбцах перечисляются переводы лексем переводимого языка. В заголовке столбца указываются параметры «по умолчанию» для переводов, вносимых в столбец.

Формат заголовка столбцов:

язык (диалект) [система записи @ источник данных]

Указанный язык и диалект автоматически проставляется для лексем и словоформ соответственно, создаваемых на основании слов в столбце.

Система записи и источник данных имеют такой же смысл, что и для словоформ переводимого языка, и также не могут быть переопределены. В общем случае, указание источника для слов-переводов не имеет смысла, поскольку обычно сами слова соответствуют литературной норме целевого языка.

Способ обозначения языков, диалектов, систем записи и источников — произвольный. Все параметры, кроме наименования языка — необязательны.

Для каждой лексемы переводимого языка, указанной в первом столбце, необходимо перечислить переводы, ей соответствующие. Близкие по смыслу переводы могут быть объединены в семантические группы. При этом каждая группа или каждый отдельный перевод располагаются на новой строке (во всех последующих строках параметры лексемы и словоформы переводимого языка не вводятся).

Формат ввода отдельного перевода:

перевод [диалект] "комментарий\_перевода"

Для каждого перевода можно задать комментарий (любой текст в кавычках) и диалект (в квадратных скобках).

Диалект, в случае с переводами, имеет смысл, отличный от случая со словоформами. Для перевода наличие диалекта показывает, что этот перевод характерен только для определённого диалекта целевого языка. Пример для слова «счёт» (из головы):

bill [брит.] | check [амер.]

Комментарий даётся для ограничения перевода. Пример для слова «bill» (перевод №4 в <http://slovari.yandex.ru/bill/en/>):

программа "(концерта, мероприятия, спектакля)".

Внутри группы переводы необходимо разделить символом «|». Описание группы может предваряться комментарием (любой текст в кавычках) и параметрами, уточняющими применимость (каждый в отдельной паре квадратных скобок, порядок следования не важен). Поддерживается задание следующих параметров:

* тема — показывает стиль, отраслевую применимость и др.;
* диалект — указывает, что семантика характерна только для определённого диалекта переводимого языка.

Таким образом, группа с полным набором параметров это может выглядеть так:

[тема] [диалект] "комментарий\_группы": перевод1 [диалект] "комментарий\_перевода" | переводN [диалект] "комментарий\_перевода"

Пример для слова «order» (перевод сущ. №7 в <http://slovari.yandex.ru/order/en/>):

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

В этом случае «[амер.]» — это указание диалекта переводимого языка, а «”(в ресторане)”» — это комментарий к переводу. Слова «порция» и «заказ» (с учётом комментария!) объединены в семантическую группу.

Пример для слова «order» (перевод сущ. №11 в <http://slovari.yandex.ru/order/en/>):

[рел.] "(orders)": духовный сан

В этом случае «[рел.]» — это тема, а "(orders)" — комментарий, указывающий, что такой смысл присутствует при определённых условиях (в данном случае, при применении во множественном числе).

Пример для слова «as» (перевод №1 в http://slovari.yandex.ru/as/en/):

"вводит предикатив и др. члены предложения со значением": как | в качестве

В этом случае комментарий "вводит предикатив и др. члены предложения со значением" объясняет семантику (смысл) группы перевода.

Группа — необязательная сущность и перевод может существовать отдельно, не сопровождаясь никакими «групповыми» комментариями и ограничениями.

# Ввод дополнительной информации

Если какой либо элемент вводимой информации (например, комментарий), слишком длинен, можно использовать метки «\*1», «\*2» и т.д. Текст, который должен быть вставлен вместо меток вводится после всех столбцов (в каждой следующей ячейке) в виде «\*1: длинный текст». В каждой строке нумерация меток начинается с 1.

Прежде чем использовать метки, важно оценить осмысленность длинного текста в каждом конкретном месте. Важно помнить, что на данный момент через массовую загрузку не предполагается вводить, например, примеры использования слов и пространные рассуждения о смысле слова.

Кроме того, описанный способ массовой загрузки не предполагает ввод для целевых языков словоформ и их диалектной применимости. Теоретически это возможно (через введение объединяющих круглых скобок или через метки), но целесообразность этого под вопросом.