Руководство по массовой загрузке словарных данных

Версия 5

# 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.

На данный момент используется кодировка UTF-8, разделить — табуляция.

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

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

* часть речи всех лексем строки и параметры лексемы переводимого языка
* словоформы переводимого языка
* переводы (параметры лексем и словоформы целевого языка)

Важно: программа использует определенные специальные символы при разборе файла (описано ниже), поэтому в самих словарных данных (т.е. в том, что содержится в источнике) таких спецсимволов следует избегать или экранировать их. Следующие символы являются специальными: [, ], ", @, |. Кроме того, для маркировки расширенных комментариев используется последовательность «\*N», где N — число.

# Лексемы

В первом столбце указывается синтактическая категория (часть речи), соответствующая лексеме. Одновременно наличие такой записи является маркером новой лексемы переводимого языка: 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]

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. В каждой строке должна присутствовать только одна синтактическая категория!

Параметрами лексемы могут быть указания на особенности лексемы (например, pl — указание того, что лексема употребляется только во множественном числе) или значение классификатора словоизменения. Таких параметров может быть несколько, каждый указывается в отдельной паре квадратных скобок. Параметры, указанные в первом столбце (в отличие от части речи), распространяются только на лексему переводимого языка.

Важно: все параметры (перечисленные как в этом, так и в последующих разделах), записываемые в квадратных скобках, должны иметь уникальное обозначение.

# Словоформы переводимого языка

Во второй и, при необходимости, последующие столбцы вносятся слова переводимого языка. Словоформы, перечисленные во втором столбце считаются соответствующими оригиналу, а словоформы, перечисленные в последующих столбцах могут представлять собой обработанные варианты оригинала. 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]

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 также используются для каждой словоформы в столбце и не могут быть переопределены (что означает, что для ввода словоформ в другой системе записи 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.

Способ обозначения языков, диалектов и систем записи— произвольный, но обозначения должны быть единообразны в рамках одного файла. 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 [параметр] "комментарий"

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.); если грамматическая категория не указана, будет записана «словарная» грамматическая категория для данной синтактической категории в данном языке;
* диалект («перекрывает» «основной» диалект, чтобы не добавлять столбцы под каждый диалект) — означает, что данная словоформа является усреднённой для этого диалекта (ВНИМАНИЕ: если в «основном» диалекте словоформа также присутствует, он должен быть перечислен среди параметров);
* 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». Кроме того, если требуется записать несколько параметров одного вида, каждый из них необходимо указывать в отдельной паре квадратных скобок. Например, диалекты: «[Soi] [Ala-L]».

В комментарии к словоформе может быть приведены, например, обстоятельства произношения (для примеров от информантов). Комментарий должен идти после всех параметров словоформы.

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

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

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

Для преобразованных словоформ (перечисленных в третьем и последующих столбцах переводимого языка) указание параметров и комментариев не допускается, поскольку для них используются параметры и комментарии оригинальной словоформы.

Количество преобразованных словоформ должно совпадать с количеством оригинальных словоформ, т.е. если разные оригинальные словоформы в преобразованном виде выглядят одинаково, преобразованная словоформа должна быть записана столько раз, сколько было разных оригинальных словоформ. Верно и обратное: оригинальную словоформу нужно повторить столько раз, во сколько разных она преобразовывается.

# 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]

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

Система записи имеет такой же смысл, что и для словоформ переводимого языка, и также не может быть переопределена.

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:

[параметры лексемы] перевод [диалект] "комментарий\_перевода"

Поскольку из каждой такой строчки извлекается сразу лексема, словоформа и перевод (как связь), дополнительная информация может присутствовать как для лексемы, так и для перевода. Дополнительная информация для лексемы записывается в квадратных скобках перед словом, а дополнительная информация для перевода — после слова.

Для создаваемой лексемы-перевода можно задать параметры лексемы (те же, что и для переводимой лексемы).

Например, для ижорского слова «očkad» в качестве перевода будет использовано русское слово «очки». Оно употребляется только во множественном числе, поэтому можно записать:

[pl] очки

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. (применяется для обеих сторон перевода);
* 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" @ [параметры лексемы] перевод1 [диалект] "комментарий\_перевода" | [параметры лексемы] 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. №9 в <http://slovari.yandex.ru/order/en/>):

[рел.] @ чин, степень священства

В этом случае «[рел.]» — это тема.

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 (внутри одной ячейки) 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, но целесообразность этого под вопросом.