ПРИЛОЖЕНИЕ 3

Алгоритм парсинга.

Слова расположены в словаре построчно:



Каждое слово находиться в начале предложения, т.е:



, то есть условие расположения слова в начале строки является обязательным, так как слово может находиться в рядах синонимов или в качестве пояснения, что все равно позволит найти синоним, но он уже может быть далек по смыслу от начального слова.

В связи с техническим аспектом, строку из словаря нужно перевести в верхний регистр, т.е:



Это сделано в связи с тем, что функция нормализации слова возвращает нормальную форму в верхнем регистре.

Когда мы нашли слово в словаре, необходимо распарсить строку и «вытащить» оттуда синоним.

Приведем используемые форматы строк.

Формат 1:



Формат 2:



В любом месте этих форматов может находиться Блок Объяснения, который помещается в скобочки и служит для объяснения смысла слова, и зачастую там находятся синонимы.

Сначала ищем формат 1, он является приоритетным.

То есть, в найденной строке ищем метку «см.», после которой начинается синоним. После метки следует пробел (см. расположение слов в словаре). То есть, нам нужно найти позицию метки, прибавить к ней четыре символа, и мы получим начало синонима. Конец синонима мы получим, найдя следующую метку – запятую с пробелом, либо если синоним один, то конец строки.

В словаре есть опечатки: иногда после синонима стоит не запятая и пробел, а просто пробел. Поэтому, можно пойти на хитрость – искать не до метки, а до пробела, а от запятых очистить функцией (приведу дальше по алгоритму).

Если формат 1 не найдем, ищем по формату 2 аналогично первому формату, то есть от метки ищем начало, и до пробела или до конца строки конец синонима.

Если не найден формат 2, значит синонимов нет.



Так как мы ищем конец синонима не до запятой и пробела, а до пробела, в синониме может находиться «мусор» - запятая или возможно скобка (бывает и такое, если синоним находится в поясняющем блоке).

Прототип функции, очищающей синоним от «мусора»:

string.**replace**(s, old, new[, maxreplace])

Чтобы не повторяться, код можно посмотреть в разделе Код.