# Gold standard for English/Arabic

Version: 1.0, Date: 15/04/11

NOTE: Check for a newer version of the data at
www.ims.uni-stuttgart.de/~sajjad/resources.html

The data is released under a Creative Commons license. We request a citation of the following paper if the data is used in a publication.

Sajjad, Hassan; Fraser, Alexander; Schmid, Helmut (2011). An Algorithm for Unsupervised Transliteration Mining with an Application to Word Alignment. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-11).

The word pairs in the gold standard are extracted from a freely available English/Arabic parallel corpus of United Nations (UN) [1]. The data is available at http://www.euromatrixplus.eu/multi-un/

We randomly take 200,000 parallel sentences from the UN corpus of year 2000, word align it and extract the list of word pairs from it.

The English/Arabic gold standard is in a single line format, where each line contains a word pair and its tag (indicating whether a word is a transliteration). The words and tags are separated by a tab like the following:
english arabic tag

There are four kinds of tags.
1. All transliteration pairs have tag "ti".

The non-transliterations are categorized in three groups.

2- Word pairs which differ by one or two characters to qualify as transliteration pairs. They have tag "d".

Example: BALBONI بالوني/baloni, BERNARD برنار/bernar

3- In Arabic, article "al" and conjunction "wao" are attached with the word. There are cases in the list of word pairs where if an Arabic word is considered without "al" and "wao", it is a perfect transliteration of its corresponding English word. We tag these word pairs as "tm".

Note that the tag ”tm” is really a special type of near transliteration which we otherwise mark with ”d”.

4- The non-transliterations other than point 2 and point 3 are tagged ”ma”.
A word pair which contains both point 2 and point 3, they are tagged ”ma”.

Note: In the paper, we have not analyzed the results using different non-transliteration categories. We have merged them under non-transliteration pairs. However we hope that they are useful for the analysis of future transliteration mining methods.

Reference:
[1] Andreas Eisele and Yu Chen. 2010. MultiUN: A multilingual corpus from United Nation documents. In Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta. European Language Resources Association (ELRA).