COMPUTATIONAL MORPHOLOGY

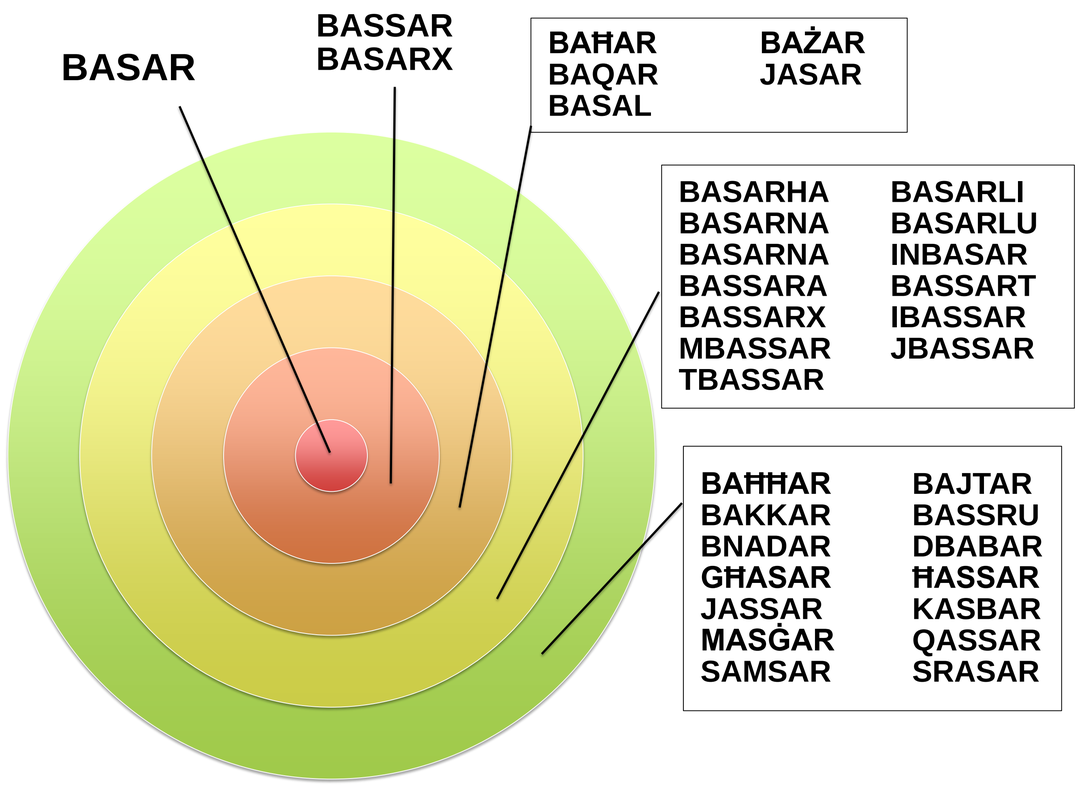
For Maltese: From Theory to Practice

Claudia Borg, Ray Fabri, Albert Gatt
Institute of Linguistics, University of Malta

Our aim is to create a **morphological analyser** to deal with different morphological formations (stem- and root-based), using **supervised and unsupervised machine learning techniques** to learn the formations automatically. We first need to establish **training data**. Our first attempt is to **cluster together morphologically related words**. Here, we discuss two of the techniques employed and some of the issues encountered.

Minimum Edit Distance

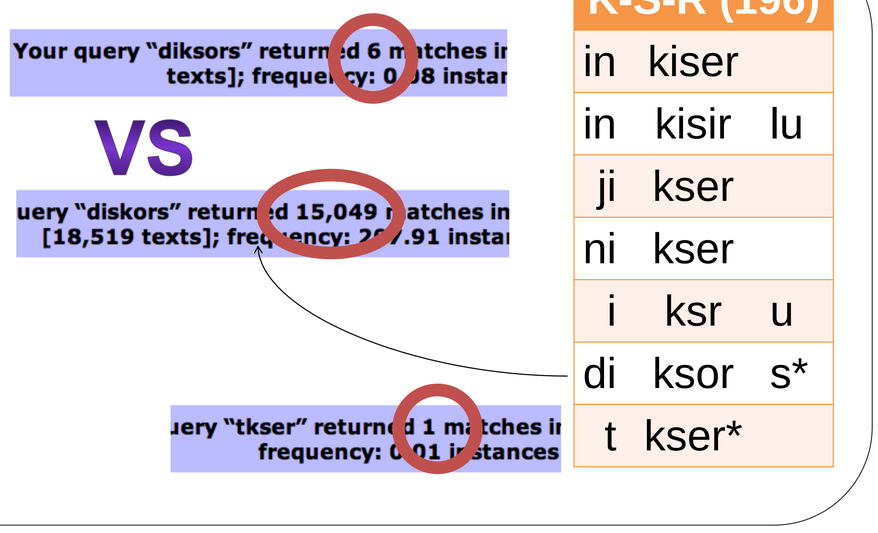
This technique clusters words together on the basis of orthographic similarity. Although we can see morphologically related words within close distance of BASAR, there are also unrelated words which will distract the learning task.



Consonant & Root–based Clustering

We now cluster by seeding the procedure with a sequence of three or four consonants (allowing only vowels to occur in between these key consonants). For words of Semitic origin, the clusters at times contain unrelated words. A closer look at these ‘errors’ shows that they are generally misspelt words occurring in the corpus. For words of Romance origin, the clusters produced are usually larger, containing groups of unrelated words. Further clustering is necessary to obtain a better grouping of related words.

Some of the words in the cluster K-S-R containing 196 words. If words are not present in the Corpus, then they are not present in our results. “Errors” in clustering can be easily identified manually.



Future Work

Development of exploration techniques whereby semantic analysis is taken into account. Although at this stage we do not have part-of-speech information, semantic analysis could be derived from the positioning of words within a sentence without the need of any additional linguistic knowledge.
Our first goal is to put together enough data to serve for training, testing and evaluation when we start applying machine learning techniques to automatically learn the morphological processes.
Once we have a morphological processor in place, it will be placed online and made available to use with other linguistic tools for Maltese.