Computational Morphology for Maltese: From Theory to Practice.

Computational linguistic tools provide the foundation for various applications, allowing the analysis of the structure of text and providing various applications with the necessary linguistic information. The availability and importance of computational linguistic tools for Maltese is increasing. In this work, we focus particularly on morphology for the Maltese language, with the attempt to create a computational model that is able to recognise and analyse grammatical formations of words.

A morphological analyser for Maltese must be able to deal with different types of formations; in particular, Maltese morphology exhibits both stem- based and root-based word formation processes. The “mixed” nature of Maltese morphology emerges clearly from a recent model of Maltese inflection by Fabri (2009). Certain assumptions for this model include information that is currently not available at a computational level (such as a lexicon with roots and stems already identified). It is also limited in scope dealing only with inflectional verb formations. Although several studies are required to bridge the linguistic knowledge to the computational one, certain heuristics and behavioural aspects can already be extracted from such work in order to ‘assist’ the computational morphology for Maltese.

Automatic derivation of morphological information presents an interesting challenge from a computational perspective. In principle, morphological information can be extracted once two or more words are associated together, and morphological rules can be extracted on the basis of similar patterns occurring regularly (for example Goldsmith (2000); Baroni (2010); Habash (2010) among others).

In this paper, we propose a preliminary approach to learning morphological patterns in unrestricted Maltese text based on the same rationale. By using the Levenshtein distance algorithm we plan to create several lists of orthographically similar words. Each set of words is then aligned, providing a number of ‘formation-shapes’ which are superimposed to detect similarities and extract formations automatically. This generic method can be supplemented with finer-grained linguistic knowledge. For instance, in computing Levenshtein distance, one can decrease the penalty for certain types of CV movement as observed by Fabri (2009), thus allowing for closer “relations” between certain word changes.

Through this work, we plan to discuss the different implications present in the proposed work, and the problems faced in the creation of a morphological analyser for Maltese. At this stage, it is not possible to evaluate such a system on the basis of metrics such as f-measure or accuracy, since we do not have an evaluation corpus available, and thus we are not in a position to present such results. However, we will present preliminary results from observations of the system’s strengths and weaknesses, in terms of different morphological associations as determined by the above techniques.

References

BARONI, M., MATIASEK, J. & TROST H. (2002). Unsupervised discovery of morphologically related words based on orthographic and semantic similarity. In: *Proceedings of the ACL-02 workshop on Morphological and phonological learning - Volume 6*

FABRI, R. (2009). Stem allomorphy in the Maltese verb. In: *Ilsienna - Our Language* - Vol. 1/2009, 1-20

GOLDSMITH, J. (2001). Unsupervised learning of the morphology of a natural language. In: *Computational Linguistics* – Vol. 27 n.2, 153-198

HABASH, N. & RAMBOW, O. (2005). Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. In: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (ACL' 05).