A Statistical Analysis of the Source Origin of Maltese – Abstract

Roderick Bovingdon
99 Chetwynd Road
Merrylands NSW 2160
Australia

roderick_bovingdon@hotmail.com

Angelo Dalli
NLP Research Group[1]
University of Sheffield
United Kingdom

angelo@dcs.shef.ac.uk

The most recent theories relating to the original source of the Maltese language point to a direct Sicilian-Arabic connection (Agius, 1990; Agius, 1993; Agius, 1996; Brincat, 1994). This paper presents the results of the first ever large-scale statistical analysis of Maltese using the newly formed Maltilex Corpus (Rosner et al., 1999; Rosner et al., 2000). Traditional etymological and categorical analyses were supplemented with data mining techniques to provide accurate results confirming traditional subjective notions.

The Maltilex Corpus is made up of a representative mixture of newspaper articles, local and foreign news coverage, sports articles, political discussions, government publications, radio show transcripts and some novels. As of the time of writing, the corpus had over 1.8 million words and almost 70,000 different word forms, making it the largest digital corpus of Maltese in existence.

A random sample of 1000 unique word forms was selected from the corpus and the etymology and category class noted down for every word in the sample. Words falling under multiple category classes were duplicated and a single category was entered for every word to ensure unique etymology/word class pairs. Weights were added to every entry representing the number of category classes associated with a particular word to maintain accurate statistics.

A data matrix with 1034 entries was thus obtained, representing all possible etymology/word class pairs for the sample word forms. The data matrix was analysed using a custom written data mining tool to extract statistics about the relationship between etymology and word classes in Maltese. Overall statistics about the source language origins of Maltese, together with the most commonly occurring word classes were also extracted. The use of a data mining tool enabled us to analyse the data from two different perspectives – word class distribution for every etymological class and vice-versa.

Maltese grammar and morphology remain to this day largely Arabic, but with distinct Romance and English morphological accretions (Aquilina, 1973; Aquilina, 1979). Traditionally the Romance element in Maltese was thought to commence with the coming of the Normans in Medieval ages and intensified during the long period of rule under the Knights of Malta. Pre-Norman Arabic content appears to have been heavily Sicilian based (Brincat, 1995).

The greatest linguistic inroads from the Italian mainland occurred during the early days of the British rule when large numbers of political refugees sought and were granted asylum during the unification of Italy (Friggieri, 1979). The most recent linguistic influence on Maltese is English. English has steadily and increasingly affected Maltese, adding another language facet to the overall structure of Maltese (Mifsud, 1995).

Despite the relatively very recent accretions from English and the vastly different morphological structure of the two languages, the assimilation of English lexemes into a Maltese mould occurs with the least possible disturbance to Maltese morphology, especially with English verbs. Interestingly, contemporary Arabic is adopting similar assimilative patterns as Maltese in its borrowings from the English-American lexis.

This study clearly shows that, in addition to other aspects, Italian lexical influence upon present day Maltese has exceed the Arabic content in a quantitative sense. Such development has also enriched Maltese from a purely root based morphology, with the additional productive Romance feature of catenation (Schweiger, 1994).

[1] With support from the Maltilex Project and the Department of Computer Science and Artificial Intelligence at the University of Malta.

References

Agius, Dionisius A. 1990. Il-Miklem Malti: a contribution to Arabic lexical dialectology, British Society for Middle Eastern Studies.

Agius, Dionisius A. 1993. Reconstructing the Medieval Arabic of Sicily. *Languages of the Mediterranean*, Brincat, Joseph M. Msida, University of Malta. 119-129.

Agius, Dionisius A. 1996. Siculo Arabic. *Library of Arabic linguistics* (Kegan Paul International) 12.

Aquilina, Joseph. 1973. *The structure of Maltese*. Msida, University of Malta.

Aquilina, Joseph. 1979. *Maltese-Arabic Comparative Grammar*. Msida, University of Malta.

Brincat, Joseph. 1994. Gli albori della lingua maltese: il problema del sostrato alla luce delle notizie storiche di al-Himyari sul periodo arabo a Malta. *Languages of the Mediterranean*, Msida, University of Malta.

Brincat, Joseph. 1995. 870-1054: Al-Himyari's Account and its Linguistics Implications. Msida, University of Malta

Friggieri, Oliver. 1979. Storja tal-Letteratura Maltija. Msida, University of Malta.

Mifsud, Manwel. 1995. *Loan verbs in Maltese a descriptive and comparative study*. Studies in Semitic languages and linguistics. Leiden: Brill.

Rosner, Michael et. al. 1999. Linguistic and Computational Aspects of Maltilex. *Proc. of the ATLAS Symposium*, Tunis.

Rosner, Mike, Ray Fabri, Joe Caruana, M Lougraïeb, Matthew Montebello, David Galea, and G. Mangion. 1999. Maltilex Project, University of Malta.

Rosner, Mike, Ray Fabri, and Joe Caruana. 2000. Maltilex: A Computational Lexicon for Maltese, Msida, University of Malta.

Schweiger, Fritz. 1994. To what extent is Maltese a Semitic Language?. *Languages of the Mediterranean*, Msida, University of Malta.