# Classifying the origin of Maltese nouns – A cross-language approach employing phonotactics

Maltese is a prime example of a language that emerged through extensive language contact, joining the two linguistic worlds of Semitic and Italo-Romance languages. Previous studies have demonstrated this on the basis of comparative methods (Comrie, 2009; Comrie & Spagnol, 2016; Lucas & Čéplö, 2020), mostly focusing on the non-concantenative and concatenative morphology of Maltese, with broken plurals such as *kelb-klieb* 'dogs' belonging to the first and sound plurals such as *fjura-fjuri* 'flowers' belonging to the second language family. The present study aims to extend earlier comparative studies by applying a computational method to the classification of a word's origin.

To do so, we trained a simple two-layer neural network (NDL, Baayen et al. (2011)) to classify 2-phones from 2347 Tunisian Arabic nouns from Gugliotta and Dinarelli (2020) as Semitic and to classify 2-phones from 2347 Italian nouns from the `subtlex-it` corpus[1] as Non-Semitic. Subsequently, the trained network was tasked with the classification of 6511 Maltese singular and plural nouns from Nieder et al. (accepted) as belonging to the categories `semitic` vs. `non-semitic`.

The network achieved an overall high classification accuracy of 97% in the training data. When the network was required to classify Maltese nouns, which were unknown to the network, as Semitic vs. Non-Semitic, the overall classification accuracy was at 70.76%. When inspecting the classification in more detail, we find that nouns with a Non-Semitic origin show a smaller classification probability as Semitic than those with a Semitic origin (39.0% vs. 84.6%, $\Delta = 45.6\%$, z-value = 34.34, p < 0.001). Moreover, we were interested how classification depended on the nouns' morphological class (sound plural, broken plural, singular sound, singular broken).

While all nouns with a Semitic origin across all morphological classes were classified as Semitic, plurals with a Non-Semitic origin showed a higher classification probability as Semitic when they show a broken pattern. This indicates that Semitic nouns in Maltese are more similar to their Semitic relatives than Non-Semitic nouns to their Non-Semitic relatives in terms of their phonotactic characteristics.

Our results show that it is possible to classify Maltese nouns using a simple two-layer network with a training based on Tunisian and Italian nouns only. The network is sensitive to the phonotactics of individual languages and can use language-specific phonotactic knowledge to classify a language that is the result of extensive language contact.

# References

Baayen, R. H., Milin, P., Durdevic, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. [Publisher: American Psychological Association]. *Psychological review*, *118*(3), 438–481.

Comrie, B. (2009). Maltese and the world atlas of language structures [Publisher: John Benjamins]. *Introducing Maltese Linguistics*, 3–12.

Comrie, B., & Spagnol, M. (2016). Maltese loanword typology. In G. Puech & B. Saade (Eds.), *Shifts and patterns in Maltese* (pp. 315–330). De Gruyter Mouton. https://doi.org/10.1515/9783110496376-012

Gugliotta, E., & Dinarelli, M. (2020). TArC: Incrementally and Semi-Automatically Collecting a Tunisian Arabish Corpus. *LREC*.

Lucas, C., & Čéplö, S. (2020). Maltese. https://doi.org/10.5281/zenodo.3744525

Nieder, J., Chuang, Y.-Y., van de Vijver, R., & Baayen, R. (accepted). A Discriminative Lexicon Approach to Word Comprehension, Production and Processing: Maltese Plurals. *to appear in Language*.

---

[1]downloaded from http://crr.ugent.be/programs-data/subtitle-frequencies