MaltLex: A database of visual lexical decision responses to 11,000 Maltese words

Jonathan Geary (University of Arizona); jonathangeary@email.arizona.edu

Introduction: In a lexical decision "megastudy", researchers collect responses to a wide range of words and non-words (e.g. differing in morphological complexity) in order to produce a massive database via which others can subsequently test novel hypotheses by analyzing a subset of the total dataset. Megastudies circumvent many of the shortcomings of traditional experiments and have grown in popularity in recent years (Keuleers and Balota 2015): visual lexical decision megastudies have been conducted on English (Balota et al. 2007), French (Ferrand et al. 2010), Malay (Yap et al. 2010), Dutch (Brysbaert et al. 2016), and Cantonese (Tse et al. 2017). To date, no megastudy has focused on a language which productively uses nonconcatenative morphology, which poses novel challenges for theories of word recognition (e.g. Frost et al. 1997). A Semitic language, Maltese not only uses nonconcatenative morphology, but its speakers have borrowed extensively from Indo-European languages such that roughly half the lexicon comprises Sicilian, Italian, and English loanwords which largely use concatenative morphology (Bovingdon and Dalli 2006), creating further challenges for theories of lexical processing. We report the creation of a database of Maltese visual lexical decision responses and demonstrate its use in replicating an analysis of the effect of etymology (Semitic vs. non-Semitic borrowings) on lexical decision speed.

Methods: In total, we have collected approximately 237,000 lexical decision responses from 104 native or near-native speakers of Maltese (M_{Age} = 24.0 years, range = 18−77 years; 53 participants identified as female, 51 as male; 87 identified as right-handed, 17 as left-handed) to 11,000 real Maltese words and 11,000 non-words. The real-word targets were randomly selected from Korpus Malti v3.0 (Gatt and Čéplö 2013), a 250-million-token corpus of written Maltese that we trimmed to remove non-Maltese texts and non-words (e.g. URLs), and supplemented by other written sources. The selected words were also checked against Ġabra, a Maltese lexical database containing 16,593 lemma-based entries (Camilleri 2013), and vetted by a native speaker. The final set of real-word targets consisted of 6,451 Semitic Maltese words, 4,439 non-Semitic words, and 110 words of uncertain etymology (Aquilina 1987-1990), and included both uninflected and inflected forms. Real-word targets ranged in frequency from 0−20,385.4 occurrences per million words in Korpus Malti (*M* = 36.1 occurrences per million words) and in length from 2−21 letters (*M* = 7.1 letters). For each real-word target, we constructed a non-word matched in length and in frequency-weighted neighborhood density (M_{Real} = 92.9, M_{Nonce} = 88.5 occurrences per million; Welch's *t*-test: $t(21{,}998) = -0.47$, *n.s.*). A native speaker vetted all potential non-word targets.

Participants completed 1−35 total sessions (*M* = 5.8 sessions), up to three sessions per day, during each of which they judged the lexicality of 200 visually-presented real words and 200 non-words to produce a total of 9−13 lexical decisions per item (*M* = 10.7 decisions). We excluded data from participants whose average RT exceeded 1,500 ms or accuracy rate fell below 80%.

Analysis of lexical stratum: In a visual masked priming lexical decision study, Geary and Ussishkin (2018) found that Maltese readers responded faster to Semitic-origin words than to non-Semitic borrowings, independent of frequency, neighborhood density, and word length. We replicate this with a larger dataset (10,890 versus 96 different words) by analyzing log RTs to real-word targets on trials where participants responded correctly, using the lme4 package (Bates et al. 2015) in R (R Core Team 2019) to fit an LMER model and assessing significance using the lmerTest package (Kuznetsova et al. 2016) to simulate Satterthwaite approximations for degrees of freedom. The model included lexical stratum (Semitic vs. Non-Semitic; reference: Non-Semitic), log frequency, log frequency-weighted neighborhood density, age, trial number, session number, and same-day session number as fixed effects; subjects and targets as random effects; and by-subjects random slopes for lexical stratum. The effect of lexical stratum was significant ($t(191.2) = -7.13$, $p < 0.001$), with participants reliably faster to judge Semitic words (*M* = 847 ms) than non-Semitic words (*M* = 852 ms). While the effect size is considerably smaller (5 versus 30 ms), the replication of this effect motivates further analyses which will assess whether, for instance, the "lexical stratum" effect may in fact reflect targets' overall morphological complexity.

References

Aquilina, J. 1987-1990. *Maltese-English-Maltese dictionary*. Malta: Midsea Books.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., and Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods* 39: 445−459.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67: 1−48.

Bovingdon, R. and Dalli, A. (2006). Statistical analysis of the source origin of Maltese. In Wilson, A., Archer, D., and Rayson, P. (eds.), *Corpus linguistics around the world*, 63−76. New York: Rodopi.

Brysbaert, M., Stevens, M., Mandera, P., and Keuleers, E. (2016). The impact of word prevalence on lexical decision times: evidence from the Dutch Lexicon Project 2. *Journal of Experimental Psychology: Human Perception and Performance* 42: 441−458.

Camilleri, J. J. (2013). *A computational grammar and lexicon for Maltese* (MSc Thesis). Chalmers University of Technology, Gothenburg, Sweden.

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., Augustinova, M., and Pallier, C. (2010). The French lexicon project: lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods* 42: 488−496.

Frost, R., Forster, K. I., and Deutsch, A. (1997). What can we learn from the morphology of Hebrew? A masked-priming investigation of morphological representation. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 23: 829−856.

Gatt, A., and Čéplö, S. (2013). Digital corpora and other electronic resources for Maltese. In *Proceedings of the international conference on corpus linguistics*. University of Lancaster.

Geary, J. A., and Ussishkin, A. (2018). Root-letter priming in Maltese visual word recognition. *The Mental Lexicon* 13: 1−25.

Keuleers, E., and Balota, D. A. (2015). Megastudies, crowdsourcing, and large datasets in psycholinguistics: an overview of recent developments. *The Quarterly Journal of Experimental Psychology* 68: 1457−1468.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2016). lmerTest: Tests in linear mixed effects models. [R package v. 2.0-32]. <https://CRAN.R-project.org/package=lmerTest>.

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Tse, C.-S., Yap, M. J., Chan, Y.-L., Sze, W. P., Shaoul, C., and Lin, D. (2017). The Chinese Lexicon Project: a megastudy of lexical decision performance for 25,000+ traditional Chinese two-character compound words. *Behavior Research Methods* 49: 1503−1519.

Yap, M. J., Rickard Liow, S. J., Jalil, S. B., and Faizal, S. S. B. (2010). The Malay Lexicon Project: a database of lexical statistics for 9,592 words. *Behavior Research Methods* 42: 992−1003.