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

Jonathan Geary
jonathangeary@email.arizona.edu
University of Arizona

THE UNIVERSITY OF ARIZONA

1. Introduction

We report the construction of a database of visual lexical decision to 11,000 Maltese words and 11,000 non-words, then demonstrate its use with two replications.

The "megastudy" approach: In a megastudy, researchers collect behavioral responses to a diverse range of stimuli (e.g. words of varying morphological complexity) to produce a database through which we can subsequently test novel hypotheses by analyzing a subset of the total dataset.

- Megastudies circumvent many of the shortcomings of traditional experiments (Keuleers and Balota 2015).
- For instance, they include a wider range of stimuli which more accurately reflect individuals' linguistic experience.
- Visual lexical decision megastudies have been conducted for Cantonese, Dutch, English, French, and Malay (Table 1).

<u>No megastudy has focused on a Semitic language; the use of nonconcatenative morphology in Semitic poses novel challenges for lexical processing (e.g. Frost et al. 1997)</u>.

Why Maltese? Maltese is a Semitic language, and much of the lexicon uses typical nonconcatenative morphology.

BUT Maltese speakers have borrowed heavily from Indo-European languages (Sicilian, Italian, and English), such that half the lexicon comprises loanwords which primarily use concatenative morphology (Bovingdon and Dalli 2006).

<u>Maltese's split lexicon thus presents further challenges for theories of word processing (e.g. Geary and Ussishkin 2018)</u>.

2. Methods

One hundred and four native or near-native speakers of Maltese participated in multiple visual lexical decision sessions.

- All participants were bilingual in Maltese and English: We had them complete the Bilingual Language Profile (BLP; Birdsong et al. 2012) to provide a composite measure of language balance.

Participants completed 1−35 sessions each (*M* = 5.8 sessions), during each of which they judged the lexicality of 200 visually-presented real Maltese words and 200 non-words.

- Real-word targets were selected randomly from Korpus Malti v3.0 (Gatt and Čéplö 2013), then checked against the Ġabra lexical database (Camilleri 2013) and vetted by a native speaker.
- Real-word targets included inflected and uninflected forms, and targets ranged in length from 2−21 letters (*M* = 7.1 letters).

We collected 9−13 judgments per target (*M* = 10.7 judgments).

4. Lexical stratum analysis

Geary and Ussishkin (2018) found that Maltese readers were faster to judge Semitic words ($N = 48$) than non-Semitic words ($N = 48$; difference = 30 ms); their sample size was small.

We compared RTs to Semitic ($N = 6{,}451$) and non-Semitic Maltese words ($N = 4{,}439$) using a LMER model that included lexical stratum (reference: non-Semitic) plus control predictors (e.g. CD, neighborhood density), assessing significance using Satterthwaite approximations for degrees of freedom via the lmerTest package (Kuznetsova et al. 2016) in R.

- **The effect of lexical stratum was significant ($t(191.5) = -7.75$, $p < 0.001$)**, with Semitic words ($M = 847$ ms) being responded to faster than non-Semitic words ($M = 852$ ms; difference = 5 ms).

3. Word frequency analysis

Brysbaert and New (2009) used the ELP dataset (Balota et al. 2007) to compare two measures of word frequency…

- Word Frequency (WF) – The number of times a word appears in a corpus;
- Contextual Diversity (CD) – The number of unique documents in which a word appears in a corpus.

…and found that CD better predicts visual lexical decision RTs.

- I Readers may actually be tracking CD across their experience, not WF, meaning that the word frequency effect is really a CD effect.
- II CD may simply better approximate linguistic experience than does WF.

We compared WF and CD (computed from Korpus Malti v3.0; Gatt and Čéplö 2013) by fitting a series of LMER models that included log-transformed WF, CD, or WF and CD as predictors (plus controls like neighborhood density), and then comparing their Akaike Information Criterion (AIC) values using the formula $\exp(-\Delta_{AIC}/2)$ (Burnham and Anderson 2004).

- **The CD model (AIC = 40,163) outperformed the WF model (AIC = 40,247; $p < 0.001$)**, but not the WF-and-CD model (AIC = 40,162; *n.s.*).

5. Discussion

Contextual Diversity (by itself) better predicts lexical decision RTs to Maltese words than does Word Frequency.

Semitic words are judged faster than non-Semitic words, though the effect size is smaller than previously observed.

These are but two analyses one could perform with MaltLex. **We aim to release the MaltLex dataset in late Summer 2020.**

Table 1 – Summary of visual lexical decision megastudies. ELP (Balota et al. 2007) and MLP (Yap et al. 2010) also included speeded naming tasks (not reported here).

	Language	Subjects	Real words	Non-words	Items/Session	Sessions/Subject	Datapoints
ELP (Balota et al. 2007)	American English	816	40,481	40,481	2,000 (1st), 1,372-4 (2nd)	2 sessions	2,749,324
FLP (Ferrand et al. 2010)	French	975	38,840	38,840	1,000	2 sessions	1,946,988
DLP (Keuleers et al. 2010)	Dutch	39	14,089	14,089	500 (1st-56th), 178 (57th)	57 sessions	1,098,942
MLP (Yap et al. 2010)	Malay	40	1,510	1,510	1,020	3 sessions	~122,400
BLP (Keuleers et al. 2012)	British English	78	28,730	28,730	500 (1st-56th), 230 (57th)	57 sessions	2,240,940
DLP2 (Brysbaert et al. 2016)	Dutch	81	30,016	29,601	500	62 sessions	2,495,448
CLP (Tse et al. 2017)	Cantonese	594	25,286	25,286	936-938	3 sessions	~1,670,000
MEGALEX (Ferrand et al. 2018)	French	96	28,466	28,466	356	50 sessions	2,596,095
MaltLex (Geary 2020)	**Maltese**	**104**	**11,000**	**11,000**	**400**	**1-35 sessions**	**237,094**

Acknowledgements

The author thanks Skye Anderson, Ray Fabri, Jessica Formosa, Kenneth I. Forster, Abigail Galea, Albert Gatt, Heidi Harley, and Victoria Sciberras Herrera for their support and expertise, as well as the Institute of Linguistics and Language Technology and the Department of Cognitive Science at the University of Malta for their assistance and for the use of their resources. Any errors are the responsibility of the author. Additionally, we wish to thank the organizers of the 33rd Annual CUNY Human Sentence Processing Conference for their handling of the COVID-19 situation.

This research was supported by a Doctoral Dissertation Research Improvement Award from the National Science Foundation (Award #BCS-1918143).

Selected References

Birdsong, D., Gertken, L. M., and Amengual, M. (2012). *Bilingual Language Profile: An easy-to-use instrument to assess bilingualism*. COERLL, University of Texas at Austin.

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., and New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods* 41: 977−990. http://doi.org/10.3758/BRM.41.4.977.

Burnham, K. P., and Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research* 33: 261−304. http://doi.org/10.1177/0049124104268644.

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

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. https://doi.org/10.1037/0278-7393.23.4.829.

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. https://doi.org/10.1075/ml.18001.gea.

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. http://doi.org/10.1080/17470218.2015.1051065.

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.