# LexiVault: A repository for psycholinguistic lexicons of lesser-studied languages

**Hind Saddiki,*** **Samantha Wray,†** **Daisy Li†**
* New York University Abu Dhabi, UAE
† Dartmouth College, New Hampshire, USA
hind.saddiki@nyu.edu, samantha.c.wray@dartmouth.edu, daisy.li.26@dartmouth.edu

**Abstract**

This paper presents LexiVault, an open-source web tool with annotated lexicons and rich retrieval capabilities primarily developed for, but not restricted to, the support of psycholinguistic research with key measures to design stimuli for low-resource languages. Psycholinguistic research relies on human responses to carefully crafted stimuli for a better understanding of the mechanisms by which we learn, store and process language. Stimuli design captures specific language properties such as frequency, morphological complexity, or stem likelihood in a part of speech, typically derived from a corpus that is representative of the average speaker's linguistic experience. These measures are more readily available for well-resourced languages, whereas efforts for lesser-studied languages come with substantial overhead for the researcher to build corpora and calculate these measures from scratch. This stumbling block widens the gap, further skewing our modeling of the mental architecture of linguistic processing towards a small, over-represented set of the world's languages. To lessen this burden, we designed LexiVault to be user friendly and accommodate incremental growth of new and existing low-resource language lexicons in the system through moderated community contributions while abstracting programming complexity to foster more interest from the psycholinguistics community in exploring low-resource languages.

**Keywords:** psycholinguistics, low-resource languages, software tool

## 1. Introduction

Psycholinguistic research is concerned with characterizing the representations and mechanisms allowing the human mind to learn, store and process language. There are many measures that have been demonstrated to be relevant to targeted experimental investigations of psycholinguistic research, including but not limited to: word, lemma, stem, root, with frequency and length calculated for each, (New et al., 2006; Kliegl et al., 2004; Yap and Balota, 2015), part of speech (POS) and morphological paradigm information, including morphemic transition probability (Taft, 1979b,a; Baayen et al., 1997; Solomyak and Marantz, 2010), and orthographic and phonological similarity (Andrews, 1997; Adelman et al., 2013; Perea, 2015). Some of these are trivial for the average researcher lacking computational or natural language processing skills to estimate for high-resource languages. For example, Google n-grams are available for word frequencies in major world languages such as English, French, Spanish, and Chinese (Michel et al., 2011). More sophisticated measures are available for high-resource languages as well, often via web interface, e.g. the CELEX database (Baayen et al., 1995).

The availability of these quantitative measures for high-resource languages contributes in part to the fact that our understanding of the mental architecture of linguistic processing is largely modeled from the study of a select few languages, helmed by English. The over-representation of English in computational and psycholinguistic models obscures the fact that many of its features are cross-linguistically quite rare (Majid and Levinson, 2010). A recent survey of research has quantified this bias for acquisition research specifically by noting the field only includes research on 1.47% of the world's languages (Kidd and Garcia, 2022). Psycholinguistic researchers of languages beyond this small number must build corpora and calculate relevant lexical statistics from scratch before tackling their main research purpose. This scarcity stands alongside a similar trend in Natural Language Processing, where researchers also call for increased diversity of inclusion of data, both at the individual language level, and the typology of linguistic features included (Joshi et al., 2020).

Within that broader context, this paper describes the development of **LexiVault**[1], an open-source web tool with annotated lexicons and rich retrieval capabilities with key measures to support design of psycholinguistic stimuli for low-resource languages, starting with Modern Standard Arabic and Tagalog. These two languages were selected to develop and launch the tool because they cover a broad range of possible linguistic features. Tagalog is written using a Latin alphabet, using left-to-right text order, and exhibits phonological alternation as part of its morphology. Arabic utilizes the Arabic script, necessitating a further level of transliteration to be

[1] https://github.com/SAVANT-team/LexiVault



*LREC-COLING 2024*, pages 10459–10465
20-25 May, 2024. 

easily machine-readable. Unlike Tagalog, its morphophonology is more transparent. Furthermore, the inclusion of these two languages served an immediate need in psycholinguistic research as two languages of the project "Systematicity and Variation In Word Structure Processing Across Languages: A Neuro-Typology Approach" (SAVANT)[2]. This project aims to investigate how the mind and brain process word-internal structure, relying on precise measurements of frequencies of words, morphemes, and morphological and phonological processes. A study utilizing morpheme frequencies generated by LexiVault has already been published (Cayado et al., 2023), demonstrating a real-life use case for this tool. Data collection efforts continue for other languages, with the goal being to invest in access for the wider community in support of research in lesser-studied languages. To reach that objective, we **(1)** generated text features and statistics by inquiring after researchers' needs and surveying other tools being used in the pre-stimuli design stage to reduce overhead; **(2)** abstracted programming complexity wherever possible to maintain an accessible learning threshold for users with limited technical background and or resources; **(3)** adopted a modular, replicable resource building process to encourage interest in contributing more languages over time.

## 2. Motivation

Psycholinguistic research increasingly depends on quantitative measurements at all levels of language, typically derived from NLP tools. A corpus is intended to be representative of what is encountered by the average speaker (Brysbaert et al., 2017). These corpus-derived measures have served as the basis of seminal research on human language processing, including to support attested effects of latencies of response to word frequency (Brysbaert and New, 2009).

Semitic languages – including Modern Standard Arabic (MSA; Glottolog: `arab1395`) – have a long history of significant contribution to psycholinguistic models due to the difference inherent in processing non-contiguous morphemes as part of root-and-pattern morphology, in which morphemes are interleaved with, rather than affixed to, a stem (Prunet et al., 2000; Idrissi et al., 2008; Boudelaa et al., 2010). Arabic exists in a rich dialectal landscape, and development of the initial MSA architecture in our current tool allows for easy adoption of additional dialects, which are being prepared for inclusion.

Despite having over 100 million speakers as a national language of the Phillipines, and a growing interest in psycholinguistic research (Pizarro-Guevara and Garcia, 2023), Tagalog-Filipino, henceforth Tagalog (Glottolog: `taga1269`), has few computational resources (Cruz and Cheng, 2021). Tagalog exhibits several interesting lexical features that made it an ideal test case for developing a robust, searchable annotated lexicon. First, in addition to prefixation and suffixation, it utilizes infixation, in which a stem morpheme is split to accommodate an affix, circumfixation in which two affixes surround a stem, and reduplication, in which a stem partially or wholly repeats. Tagalog words may exhibit multiple of these processes simultaneously. Furthermore, the presence of morphophonological phenomena such as nasal coalescence means that stemming is not always a simple case of affix stripping.

While exploring software options to host and query the lexicons being developed, we observed that interfaces for psycholinguistic data were heavily skewed towards English and other overrepresented languages (Fearnley, 1997; Taylor et al., 2019). SketchEngine (Kilgarriff et al., 2014) is one of the most widely used multilingual resource platforms. Although designed to support lexicography research, it does offer a mature ecosystem to onboard new languages and a friendly interface to query and retrieve useful some statistics, but very few statistics at the lexical or sub-lexical level such as morphemic features. Additionally, there is the critical constraint of being a subscription-based commercial service. On the other side of our tool spectrum is Aralex (Boudelaa and Marslen-Wilson, 2010), a free-access interface specifically designed to address the lack of psycholingistic data for an understudied language - MSA in this case. Many of the data annotation and search features offered in Aralex served as inspiration in developing LexiVault, while we sought to overcome some of its shortcomings in terms of limited retrieval capability and lack of infrastructure to accommodate more languages. As we developed LexiVault, we sought to bring together desirable features from both SketchEngine and Aralex to satisfy our design criteria, as summarised in Table 1.

## 3. Building Language Resources

As we compiled the list of word features to be generated for each corpus, given the eventual goal of wide linguistic diversity and the presence of language-specific attributes such as root-and-pattern morphology, we sought to establish a baseline structure for language lexicons with the minimum required components to be viably useful in a psycholinguistic study. This would allow for wider accommodation when onboarding severely resource-poor languages beyond the ones planned

[2]https://savant.qmul.ac.uk/

| Design Requirements | SketchEngine | Aralex | LexiVault |
|---|---|---|---|
| Psycholinguistic data | | ✓ | ✓ |
| Multi-language support | ✓ | | ✓ |
| Freely accessible | | ✓ | ✓ |
| User-friendly data retrieval | ✓ | | ✓ |
| User contribution support | ✓ | | ✓ |

Table 1: Project requirements and key tools inspiring LexiVault's design

for this particular project. As is often the case with low-resource languages, reliable annotation tools may not be available to produce a richer set of features. Therefore, we have established a baseline process and structure as follows, with more advanced processing for the target languages in following subsections:

- Build a sizeable corpus, previous work suggests a 16 million word minimum to be psycholinguistically representative (Brysbaert and New, 2009)
- Tokenize the corpus to word-level and generate raw word frequencies
- Perform grapheme-to-phoneme (G2P) transcription, preferably to the International Phonetic Alphabet (IPA)
- Generate character bi-tri-grams with frequencies
- *If possible*, perform rule-based stemming and validate with an expert informant
- *If stemming is possible*, generate morpheme frequency and morpheme-to-whole word transition probability
- Normalize frequencies to parts-per-million

Phonological Corpus Tools was used (Hall et al., 2019) to calculate the following measures: phonotactic probability, bigram probability, and minimal pairs. Phonotactic probability is calculated as the average unigram or bigram positional probability across a word (Vitevitch and Luce, 2004). Our three measures of interest are implemented using PCT's API functionality offline and included as annotation to individual items of the corpus.

The functions took as input the lexicon and a phonological feature matrix. The default feature matrix was used for Tagalog. Two modifications to the features were performed in the PCT GUI for MSA. Binary features for *pharyngealization* and *length* were added as these are phonemically contrastive in Arabic.

Finally, when more mature computational tools are available, additional annotations of interest such as POS tagging can be pursued in addition to the minimal list here. Both MSA and Tagalog included this, as described below.

### 3.1. Modern Standard Arabic (MSA)

The MSA lexicon was constructed from the Arabic subset of the GeoWAC family of corpora developed by Dunn and Adams (2020), primarily based on Web and Twitter data amounting to 618 million MSA words. It is our intention to balance the content out with more free-access MSA data sourced from newswire and other domains.

Despite the aforementioned interest in MSA and related languages due to their typologically rare root-and-pattern morphology, most existing computational resources for Arabic lack both root and pattern morpheme annotation, collapsing these into a singular 'stem' morpheme. We address that in our processing pipeline by utilizing Camel-Tools (Obeid et al., 2020) a Python suite of tools to perform a wide range of NLP tasks specifically designed for MSA, providing a rich set of annotation from stems, lemmas, POS to root and pattern, and further attributes detailed in the Camel-Tools documentation.

- Normalizing the text to address letter variants and vowel inconsistencies common in MSA script, especially in Web and social media texts
- Running Camel-Tools morphological analysis to produce stems, lemmas and POS tags, with respective reported accuracies of 95.7% and 95.5%
- Retrieving roots and patterns mappings within the Camel-Tools morphological database, with validation from an Arabic-speaking informant fluent in MSA to address error-or-blank entries
- Annotating words with phonemic transcription in IPA with Phonemizer (Bernard and Titeux, 2021)

### 3.2. Tagalog

The Tagalog lexicon was constructed from pre-existing corpora that were processed and annotated as described here. The existing sources that comprise the corpus include: 175 million words of Web data and newswire (Cruz and Cheng, 2021), 52 million words of Web data (Zuraw, 2006), 28.6 million words of Web data and Twitter (GeoWac) (Dunn and Adams, 2020), 2.1 million words of Wikipedia (Wray et al., 2022).

After consolidating these resources, the following processing pipeline was followed:

- Given the level of multilingualism in the Phillipines, code-switching and mixed-language text was common in the corpus. Language detection was performed at the sentence level and non-Tagalog text was discarded. This was performed with the python $langdetect$[3] port of the

[3]https://pypi.org/project/langdetect/

Java $language - detection$ module[4]
- POS tagging, using (Go and Nocon, 2017), a Stanford POS tagger (Toutanova et al., 2003) with a reported accuracy of 96.15%, after which reduction of full sentences to a lexicon of unique word-tag pairs was performed
- Stemming via in-house rule-based stemmer developed and validated with assistance of native Tagalog-speaking informant
- Annotating words with phonemic transcription in IPA, using Epitran (Mortensen et al., 2018)

## 4. Designing LexiVault

The design for LexiVault provides: **(1)** user-friendly retrieval interface for existing psycholinguistic lexicons, and **(2)** onboarding support for new language lexicons. We translated these objectives to technical requirements in keeping with the broader aims of the project for a system that is:

**REQ.1.** Freely accessible and open-source for ongoing enhancements

**REQ.2.** User friendly and rich in filtering parameters for precise data retrieval

**REQ.3.** Minimally technical for onboarding new languages to the interface

### 4.1. Architecture

To address **REQ.1.** and **REQ.3.**, we opted for an open-source setup (Fig.1) combining Github[5] to house the source code in a stable platform with contributor-friendly versioning support, and Streamlit[6] for a web development framework that completely insulates the front-end JavaScript and HTML scripting with a well-documented Python API for ease of use by contributors who are more likely to be familiar with Python as a popular programming language in text-processing-heavy disciplines.

Additionally, Streamlit functionality further simplifies **REQ.3.** with its modular Multipage App organization for seamless addition of multiple languages. With LexiVault set up with a main landing page and a multipage structure for each of the onboarded languages, the code is organized much like a file system tree, with modular language pages in a dedicated directory, automatically linked to a sidebar navigation on the web app. With this modular structure in mind, we created a **Template Page.py** for contributors to simply clone in the Github repository, and make simple edits to the Python script to connect that page to the appropriate language lexicon in the database, then, if necessary, tweak any of the search features to align with the attributes available in said language lexicon.

[4]https://github.com/shuyo/language-detection

[5]https://docs.github.com/en

[6]https://docs.streamlit.io/

### 4.2. Functionality

To satisfy **REQ.2.**, Streamlit offers a number of interactive widgets and an intuitive interface to accommodate advanced search features, display results on-screen in a paginated, customizable tabular form and offer users the option to export the results for use in stimuli design or further analysis.

**Flexible search parameters** The interface (Fig.2) offers advanced search functionality tailored to the language.[7] There is a toggle option to enforce strict or approximate matches. Category widgets like the POS options all dynamically link to the data source to display tags actually available in the lexicon. The searchkey text field handles one or multiple space-separated keys and all results are compiled into one query for efficiency.

**Batch retrieval and export** When planning stimuli, users often compile lists of candidate words to retrieve measures for. While Aralex (Boudelaa and Marslen-Wilson, 2010) only allows individual searches, forcing the user to reset after each one, LexiVault accommodates batch retrieval with a drag-and-drop file upload function: users upload a list of searchkeys in a simple text file, set the desired parameters, and get a complete set of results neatly compiled for each entry in their input file. When results are displayed, LexiVault offers the users an **Export as .csv** option to save their results in an easy-access format to address another user-experience limitation observed when using Aralex: in the absence of an export option, users need to get creative to copy the results from the web interface.

## 5. Conclusion and Next Steps

This paper introduces LexiVault, a web tool for storing and searching lexicons of lesser-resourced and under-researched languages. These lexicons are annotated with psycholinguistic measures in mind.

The tool has been launched with two language lexicons prepared: Modern Standard Arabic and Tagalog. LexiVault is designed to expand, with community assistance, and several languages are already in progress. This includes several dialects of Arabic, including Gulf and Egyptian. Bangla, a language with highly divergent phonemic and graphemic forms that will showcase the tool's lexical statistics, is also in progress.

As we continue development on language resources, the tool itself will keep evolving, informed by real user feedback from psycholinguists. Some planned functionality enhancements include: a tab-based navigation within each language page to

[7]For instance, the MSA interface supports input in either Arabic script or Buckwalter transliteration http://www.qamus.org/transliteration.htm

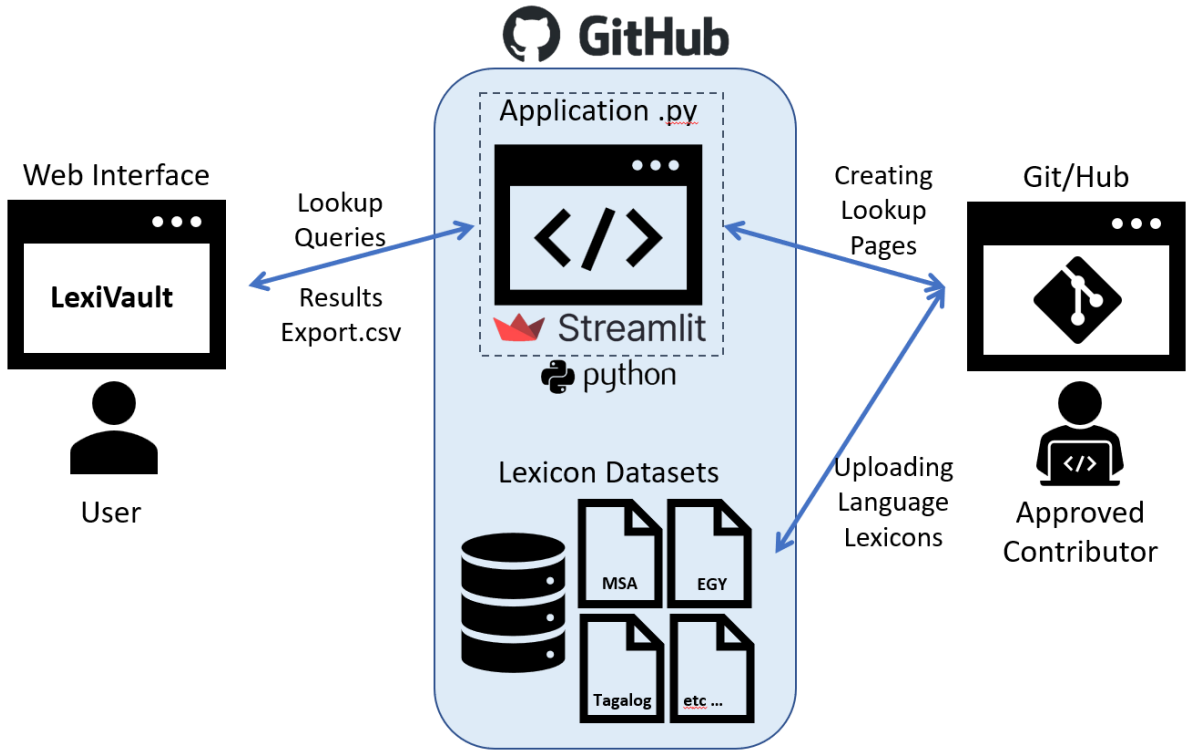


Figure 1: Overview of the LexiVault open-source achitecture.

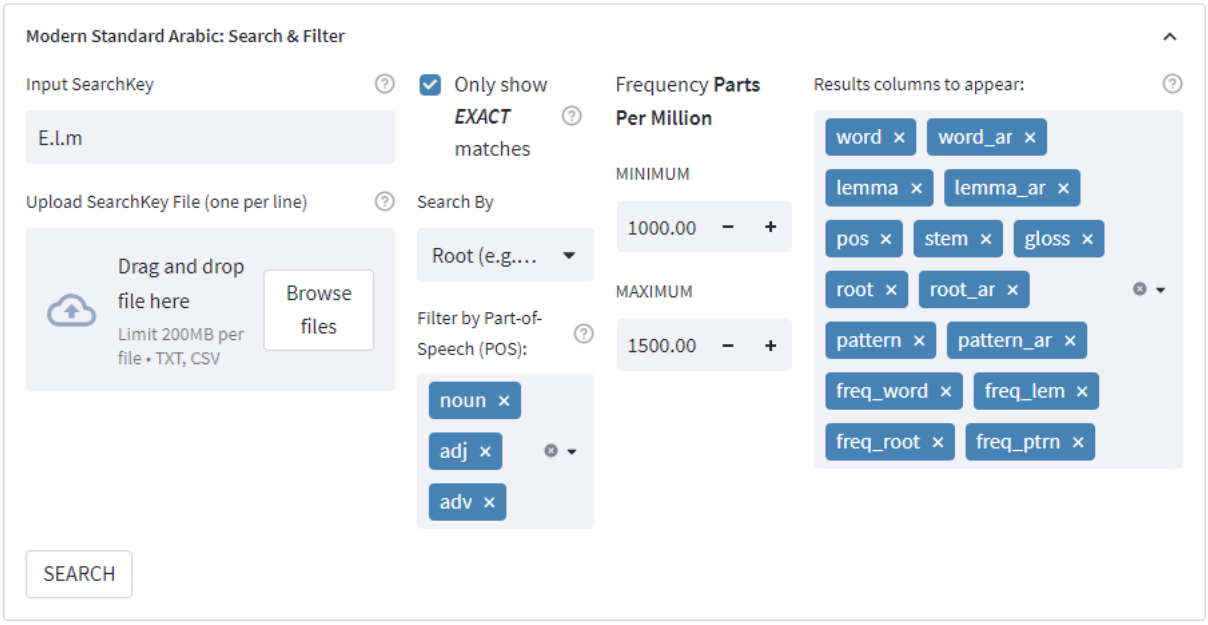


Figure 2: Example search parameters on the MSA language page.

toggle between word-level measures and n-gram measures which demand a slightly different set of search parameters; augmenting the matching option with regex-like capability for the searchkey function to detect more elaborate patterns.

In preparation for wider contribution, we're developing a moderation protocol to ensure quality and structure compliance both for on-boarding new languages and augmenting existing ones. Documentation is also underway to support users in their creation of new language pages, as well as video tutorials andor workshops for both contributors and end-users.

## 6. Acknowledgements

The support of the Economic and Social Research Council (ESRC) Grant no.: ES/V000012/1 for the SAVANT project is gratefully acknowledged.

## 7. Bibliographical References

James S Adelman, Suzanne J Marquis, Maura G Sabatos-DeVito, and Zachary Estes. 2013. The unexplained nature of reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(4):1037.

Sally Andrews. 1997. The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic bulletin & review*, 4(4):439–461.

R Harald Baayen, Ton Dijkstra, and Robert Schreuder. 1997. Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of memory and language*, 37(1):94–117.

R Harald Baayen, Richard Piepenbrock, and Léon Gulikers. 1995. The CELEX lexical database [webcelex]. *Philadelphia, PA: University of Pennsylvania Linguistic Data Consortium*.

Mathieu Bernard and Hadrien Titeux. 2021. Phonemizer: Text to phones transcription for multiple languages in python. *Journal of Open Source Software*, 6(68):3958.

Sami Boudelaa and William D Marslen-Wilson. 2010. Aralex: A lexical database for modern standard arabic. *Behavior Research Methods*, 42(2):481–487.

Sami Boudelaa, Friedemann Pulvermüller, Olaf Hauk, Yury Shtyrov, and William Marslen-Wilson. 2010. Arabic morphology in the neural language system. *Journal of cognitive neuroscience*, 22(5):998–1010.

Marc Brysbaert, Paweł Mandera, and Emmanuel Keuleers. 2017. Corpus linguistics. *Research methods in psycholinguistics and the neurobiology of language: a practical guide*, pages 230–246.

Marc Brysbaert and Boris New. 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(4):977–990.

Dave Kenneth Tayao Cayado, Samantha Wray, and Linnaea Stockall. 2023. Does linear position matter for morphological processing? evidence from a tagalog masked priming experiment. *Language, Cognition and Neuroscience*, 38(8):1167–1182.

Jan Christian Blaise Cruz and Charibeth Cheng. 2021. Improving large-scale language models and resources for Filipino. *arXiv preprint arXiv:2111.06053*.

Jonathan Dunn and Ben Adams. 2020. Geographically-balanced gigaword corpora for 50 language varieties. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2528–2536.

Stephen Fearnley. 1997. MRC Psycholinguistic Database search program. *Behavior Research Methods, Instruments, & Computers*, 29:291–295.

Matthew Phillip Go and Nicco Nocon. 2017. Using Stanford part-of-speech tagger for the morphologically-rich Filipino language. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, pages 81–88.

Kathleen Currie Hall, J Scott Mackie, and Roger Yu-Hsiang Lo. 2019. Phonological CorpusTools: Software for doing phonological analysis on transcribed corpora. *International Journal of Corpus Linguistics*, 24(4):522–535.

Ali Idrissi, Jean-François Prunet, and Renée Béland. 2008. On the mental representation of Arabic roots. *Linguistic Inquiry*, 39(2):221–259.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state andfate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Evan Kidd and Rowena Garcia. 2022. How diverse is child language acquisition research? *First Language*, 42(6):703–735.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography*, 1:7–36.

Reinhold Kliegl, Ellen Grabner, Martin Rolfs, and Ralf Engbert. 2004. Length, frequency, and predictability effects of words on eye movements in reading. *European journal of cognitive psychology*, 16(1-2):262–284.

Asifa Majid and Stephen C. Levinson. 2010. WEIRD languages have misled us, too. *Behavioral and Brain Sciences*, 33(2-3):103–103.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Google Books Team, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, et al. 2011. Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182.

David R Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Boris New, Ludovic Ferrand, Christophe Pallier, and Marc Brysbaert. 2006. Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic bulletin & review*, 13:45–52.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.

Manuel Perea. 2015. Neighborhood effects in visual word recognition and reading. *The Oxford handbook of reading*, 1.

Jed Sam Pizarro-Guevara and Rowena Garcia. 2023. Philippine Psycholinguistics. *Annual Review of Linguistics*, 10.

Jean-François Prunet, Renée Béland, and Ali Idrissi. 2000. The mental representation of Semitic words. *Linguistic inquiry*, 31(4):609–648.

Olla Solomyak and Alec Marantz. 2010. Evidence for early morphological decomposition in visual word recognition. *Journal of Cognitive Neuroscience*, 22(9):2042–2057.

Marcus Taft. 1979a. Lexical access-via an orthographic code: The basic orthographic syllabic structure (BOSS). *Journal of Verbal Learning and Verbal Behavior*, 18(1):21–39.

Marcus Taft. 1979b. Recognition of affixed words and the word frequency effect. *Memory & Cognition*, 7:263–272.

Jack E Taylor, Alistair Beith, and Sara C Sereno. 2019. LexOPS: An R Package and User Interface for the Controlled Generation of Word Stimuli.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 human language technology conference of the north american chapter of the association for computational linguistics*, pages 252–259.

Michael S Vitevitch and Paul A Luce. 2004. A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, & Computers*, 36(3):481–487.

Samantha Wray, Linnaea Stockall, and Alec Marantz. 2022. Early form-based morphological decomposition in Tagalog: MEG evidence from reduplication, infixation, and circumfixation. *Neurobiology of Language*, 3(2):235–255.

Melvin J Yap and David A Balota. 2015. Visual word recognition. *The Oxford handbook of reading*, 1:1–36.

Kie Zuraw. 2006. Using the web as a phonological corpus: a case study from Tagalog. In *Proceedings of the 2nd International Workshop on Web as Corpus*.