LexiPhon: A Collection of Phonetically Transcribed Lexicons from Wikipedia

Amanda Doucette[1], **Timothy J. O'Donnell**[1,2,3], **Morgan Sonderegger**[1]
[1]McGill University, [2]Mila, [3]Canada CIFAR AI Chair
Montréal, Québec, Canada
amanda.doucette@mail.mcgill.ca, {morgan.sonderegger, timothy.odonnell}@mcgill.ca

Abstract

We introduce LexiPhon, an open-source dataset of phonetically transcribed lexicons for 87 languages derived from Wikipedia data with automated grapheme-to-phoneme (G2P) transcription, along with the open-source software used to create it. Each lexicon provides transcriptions generated by up to three G2P methods, crowdsourced transcriptions from WikiPron (Lee et al., 2020) where available, word frequencies calculated from Wikipedia, along with word lengths and phonological neighborhood densities. We introduce an internal validation metric based on phonological feature edit distance to ensure transcriptions are consistent within languages, as manual validation is not possible. This dataset fills a gap in the existing space of phonetic lexicons, with a much larger set of words per language than existing multilingual word lists, and more languages than existing lexicon datasets. The dataset, along with the software used to create it, are freely available on OSF at https://osf.io/rd9ma.

Keywords: lexicon, Wikipedia, phonetic transcription, G2P

1. Introduction

Although large quality-controlled phonetically transcribed lexicons are available for some widely studied languages, such as English, German, Dutch (Baayen et al., 1995), and French (New et al., 2004), they are not available in many less studied languages. Large-scale phonetic transcription data are useful in many areas of computational linguistic research, such as studying typological variation and linguistic universals. However, when studying cross-linguistic variation, high-quality lexicons in just a few languages do not provide enough data to draw typological conclusions. Existing resources, although very useful for some types of linguistic research, exhibit several inadequacies for phonetic typology.

Existing multilingual word lists contain hundreds or thousands of languages, but often only have phonetic transcriptions for a small word list in each, similar to a Swadesh list (Swadesh, 1955). For example, the Automated Similarity Judgment Program (ASJP; Brown et al., 2008) database contains over 6,000 languages, but only 40 words per language. The Global Lexical Database (GLED; Tresoldi, 2023) contains over 6,000 languages, but transcribes only 30-40 concepts per language. NorthEuraLex (Dellert et al., 2020), which covers 107 languages, provides a larger list of approximately 1,000 concepts per language, but is still not representative of an entire lexicon. Similarly, LexiBank (List et al., 2022) is a collection of lexical datasets with approximately 3,000 concepts transcribed for over 2,000 language varieties. Datasets with larger phonetically transcribed word lists do exist, like VoxCommunis (Ahn and Chodroff, 2022) with between 0.24 and 288 hours of speech transcribed for 36 languages, and VoxClamantis (Salesky et al., 2020) with transcriptions of Bible recordings in 635 languages, but these datasets also are not representative of an entire lexicon. Some languages in VoxCommunis with hundreds of hours of transcribed speech may be, but they are not comparable to the data for languages with only a few hours of speech. VoxClamantis, while including a large number of languages, only contains Bible translations, which will be missing many lexical items.

These existing lexical datasets (summarized in Table 1) have been used to study, for example, correlations between word length and phoneme inventory (Wichmann and Holman, 2023) or word length and phonotactic complexity (Pimentel et al., 2020), but lexical studies like these would benefit from a larger, more representative set of words per language. This would provide higher statistical power, allowing for negative results to be identified. Using a small subset of the lexicon can also radically change conclusions – Pimentel et al. (2020) found a negative correlation between word length and phonotactic complexity in NorthEuraLex (Dellert et al., 2020), which mainly contains morphologically simple words. However, Doucette et al. (2024) found a positive correlation in a larger sample of lexical data containing morphologically complex words.

In the absence of large phonetically transcribed lexicons, some previous studies have instead used orthography as a proxy to study phonetic phenomena in the lexicon (e.g. Piantadosi et al., 2011; Wu

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 5689–5700
11–16 May 2026.

Dataset	Citation	Languages	Data per Language	IPA?
LexiPhon		87	56,000 words	Yes
NorthEuraLex	(Dellert et al., 2020)	107	1,000 words	Yes
ASJP	(Brown et al., 2008)	6,000	40 words	No
Global Lexical Database	(Tresoldi, 2023)	6,000	30-40 words	Yes
LexiBank	(List et al., 2022)	2,000	3,000 words	Yes
VoxCommunis	(Ahn and Chodroff, 2022)	36	0.24 - 288 hours	Yes
VoxClamantis	(Salesky et al., 2020)	635	20 hours	Yes

Table 1: Existing multilingual lexical datasets.

et al., 2019; Mahowald et al., 2018). In languages that have straightforward orthography-to-phoneme mappings, this may be justified, but in many languages, orthography does not clearly correspond to phonetic transcription. Other studies have instead examined either a single language or a few languages where transcriptions are available (e.g. Hay and Baayen, 2003; Fratini et al., 2014), and speculated that their findings may also apply to a larger set of languages.

In this paper, we introduce LexiPhon – a dataset designed to fill this gap in lexical resources, as well as open-source software for creating additional lexicons. This collection of lexicons in 87 languages with an average of 56,899 words per language is derived from Wikipedia and supplemented with automated grapheme-to-phoneme transcriptions and community-written transcriptions from WikiPron (Lee et al., 2020). Additional information that is often needed in studies of the lexicon, such as word lengths, phonological neighborhood density, and number of vowels is also included, and the software is written to allow the dataset to be easily extended with additional measures or transcription methods. In this paper, we describe our methodology for creating these lexicons, including excluding non-words and foreign language words from these lexicons and our procedure for normalizing automatically generated phonetic transcriptions. We also provide summary statistics for the dataset and a validation of the phonetic transcriptions.

2. Data

Wikipedia is available in over 300 languages, 180 of which contain more than 10,000 articles (Wikipedia contributors, 2025). Because digital lexicons are either not available or very small for many of these languages, extracting this data from Wikipedia allows us to examine a much larger set of languages than would otherwise be possible and study languages where lexicons compiled by humans are not available. For each language, we download a "dump" – a backup of all pages in the language's Wikipedia in HTML format.[1]

We also use transcriptions from WikiPron (Lee et al., 2020), a database of human-written pronunciation dictionaries scraped from Wiktionary,[2] a collaborative multilingual dictionary. Because the phonetic transcriptions in Wiktionary are written by volunteer community members, they are sometimes inconsistent in the level of detail included in the transcription. For some languages, there are very few entries in Wiktionary, but in others there are millions of entries. In our dataset, we include WikiPron transcriptions when available.

3. Methods

3.1. Pre-processing

Wikipedia HTML dumps, like most raw HTML data, are messy and require some pre-processing to extract a lexicon. Following previous large web-based datasets, such as The Pile (Gao et al., 2020) and the RefinedWeb dataset (Penedo et al., 2023), we implement a pre-processing pipeline to convert the raw data to a usable format and filter out non-lexical items. However, lexicons are inherently different from large language model (LLM) training datasets such as the two just cited. Where an LLM dataset may be concerned with filtering data at the sentence or full-text level, we are concerned with filtering data at the word level. In this section, we describe the pre-processing pipeline implemented for LexiPhon.

First, we extract the text of the article from the HTML with tools from the Gensim Python package (Řehůřek and Sojka, 2010), leaving us with the raw text of each Wikipedia article. Next, we remove named entities from the text. Names are often borrowed from other languages or are in some way not representative of the language's lexicon. For example, the name of the Polish city Gdańsk occurs in English Wikipedia, but should not be included in an

[1] Retrieved 9/1/25 from https://dumps.wikimedia.org

[2] https://www.wiktionary.org/

English lexicon. To remove named entities, we use the ParaNames corpus (Sälevä and Lignos, 2024), which includes 16.8 million named entities in more than 400 languages. Many of these named entities are multiple words (e.g. "New Mexico" or "Pacific Ocean"), so we remove them before further segmenting the text into words. All names occurring in the ParaNames corpus (case-insensitive) were removed from each Wikipedia article before further processing.

Next, we split the text into individual words using the word_tokenize function from the NLTK Python package (Bird et al., 2009). Text is tokenized according to language-specific rules when available, defaulting to English tokenization otherwise. One language in our dataset, Japanese, does not split words on white space, so a specialized tokenizer is needed. For this we use the SudachiPy Python package (Takaoka et al., 2018), which is capable of tokenizing Japanese text.

The preliminary word list obtained after the tokenization step includes many tokens that should not be considered words in a language's lexicon. For example, some HTML fragments remain (e.g. "upright=0.9", ".jpg|alt=apricot"), as well as URLs embedded in article text (e.g. "http://www..."), numerals, typographical errors and misspellings (e.g. "4.65billion", "everybodys", "Ukranian"), and foreign-language words (e.g. "chanté" or "ogólnopolski" in English Wikipedia). As the dataset is too large to filter manually, we employ several methods to filter these non-words automatically.

First, we attempt to filter out all words containing characters outside the language's orthography. Hyperglot (Rosetta Type Foundry, 2025), a tool developed to check language support in fonts, contains a database listing the Unicode character codepoints needed to represent the standard orthography of over 700 languages. Tokens containing Unicode codepoints outside these sets were excluded. Tokens containing punctuation and numerals, with the exception of word-internal punctuation[3] were also excluded. For languages not included in Hyperglot, tokens including any Unicode numeral, punctuation, or symbol codepoint were excluded.[4]

Based on manual spot-checking of the English data, these rules result in the exclusion of most non-words and some out-of-language words. While filtering by orthography removes all words from languages with different orthographies, words from languages sharing the same orthography will remain, as will typographical errors and misspellings. To exclude these, we remove low-frequency tokens. Most typographical errors and misspellings are not expected to occur frequently, and while foreign language words are occasionally used in Wikipedia articles, they are typically rare. Token frequencies are calculated as the number of times the token occurs per million and are included in the final lexicons. Any token occurring less than once per million was excluded. This threshold was selected by manual inspection of excluded words in English – while this certainly has the effect of removing some legitimate low-frequency words from our lexicons, the majority of excluded tokens were not words in a manual inspection of the English tokens excluded by this rule. Furthermore, lexical frequency distributions contain a large number of very low-frequency words, resulting in frequency estimates that are highly dependent on sample size (Baayen, 2001). Many low-frequency words will not occur in a sample from Wikipedia at all, so rather than include some low-frequency words in our lexicons along with many non-words, we choose to exclude very low-frequency words altogether.

3.2. Grapheme to Phoneme Transcription

After obtaining a list of words for each lexicon, we convert the filtered list to IPA transcriptions using grapheme-to-phoneme (G2P) methods. There are many G2P methods available, often resulting in slightly different transcriptions of the same word. Furthermore, each G2P method works on a different subset of languages. Previous datasets using G2P transcription methods typically apply only one method per language: NorthEuraLex (Dellert et al., 2020) uses transcriptions based on handwritten rewrite rules, VoxCommunis (Ahn and Chodroff, 2022) uses the XPF (Priva et al., 2021) and Epitran (Mortensen et al., 2018) G2P systems, and VoxClamantis (Salesky et al., 2020) uses the Unitran G2P (Qian et al., 2010). Rather than choosing one method and limiting the number of languages included in the dataset, we use multiple G2P methods and include transcriptions for each in the languages they are available in. Although not an automatic G2P method, we also include the WikiPron transcriptions described above when available. The three G2P methods used in this dataset were chosen to maximize language coverage: XPF (Priva et al., 2021), Epitran (Mortensen et al., 2018), and CharsiuG2P (Zhu et al., 2022). The software released with the dataset allows for additional transcription methods to be added as needed.

XPF (Priva et al., 2021) is a rule-based G2P method that uses handwritten orthography-to-phoneme mappings to transcribe words. This rule-based system works well in languages with trans-

[3]Unicode codepoints: 0x0027, 0x055A, 0xFF07, 0x2032, 0x2035, 0x055F, 0x0836, 0x0970, 0x09FD, 0x0A76, 0x0AF0, 0x10B39, 0x110BB, 0x11174, 0x111C7, 0x1123D, 0x1144F, 0x114C6, 0x11643, 0x116B9, 0x1183B, 0x2027, 0x2043

[4]Unicode categories: Nd, Nl, No, Pc, Pe, Pf, Pi, Po, Ps, Sc, Sm, So, Zl, Zp, Zs

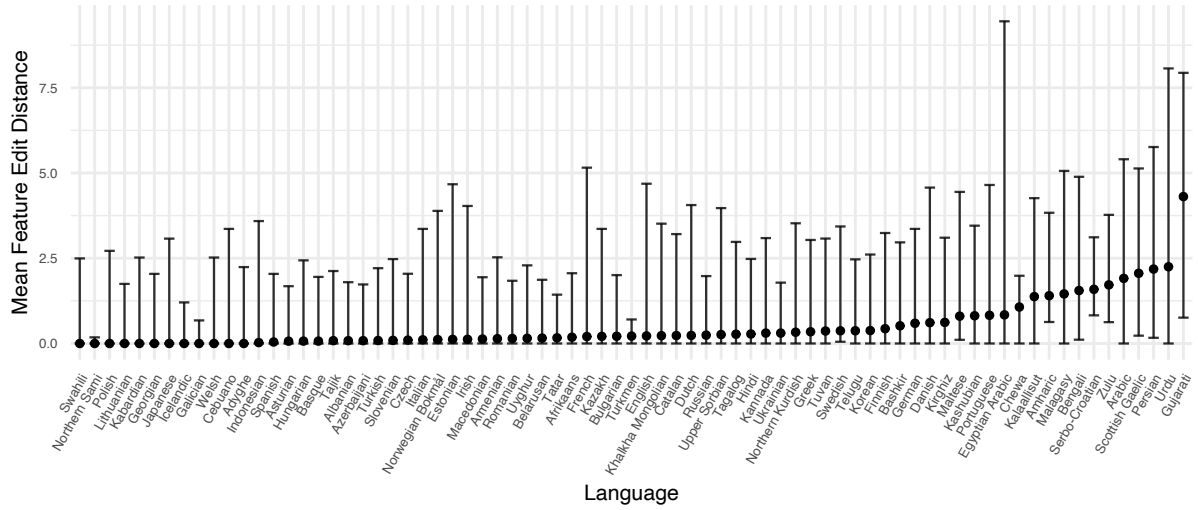

Figure 1: Mean feature edit distance between transcription methods for each language, with 95% quantiles. Languages with only one transcription method are excluded.

parent orthography-to-phoneme mappings, like Ukrainian and Turkish. Any out-of-vocabulary symbols are transcribed as "@", which allows us to filter out any words XPF is unable to transcribe.

Epitran (Mortensen et al., 2018) is another rule-based G2P method using handwritten mappings. However, it also includes pre- and post-processing steps that allow it to handle languages with less transparent orthographies, like English and German

CharsiuG2P (Zhu et al., 2022) is a multilingual neural transducer-based model trained on IPA transcription data for 100 languages. We use a pre-trained model hosted on Hugging Face to generate transcriptions.[5] This method is able to transcribe languages with less transparent orthography-to-phoneme mappings, but sometimes produces inaccurate transcriptions or outputs non-IPA symbols. It appears to give reasonable transcriptions in some low-resource languages like Northern Sami and Welsh, but performs poorly in other languages, like Swedish, where non-IPA symbols are frequently output.

3.3. Transcription Post-processing

These G2P methods, along with the human-written transcriptions from WikiPron, vary in the accuracy and level of detail included in the transcriptions they produce. For example, some include diacritics for stress, length, tone, and syllable boundaries, while others do not. Following Lexibank (List et al., 2022), an aggregation of many lexical datasets with different transcription systems, we apply a normalization procedure to each transcription in order to make them more comparable across G2P methods and languages. There are many linguistic transcription systems, often using variants of the same IPA character (e.g. dz, $\widehat{\text{dz}}$, and $\text{d}\widehat{\text{z}}$), which can be difficult to reconcile (Anderson et al., 2018). For the purposes of the validation described in the following section, we map symbols in the transcriptions to the IPA symbols included in the Panphon Python package (Mortensen et al., 2018). All affricates are mapped to their equivalent with a tie bar, tones transcribed as diacritics are mapped to tone letters, and pairs of confusable Unicode characters are mapped to a single character. We also remove syllable boundary and stress markers, as these occur inconsistently across languages, even within the same G2P system. Finally, we check if all symbols in the transcription are valid IPA symbols in Panphon. Transcriptions with invalid symbols are still included in the published lexicons, with an additional column marking whether the transcription contains only valid IPA symbols.

After processing the transcriptions, we also include several additional lexical statistics in the dataset: The length of each transcription, the number of vowels (as a simple proxy for number of syllables), and phonological neighborhood density, or the number of words in the lexicon that differ by exactly one IPA symbol (Luce and Pisoni, 1998). These measures are often useful in lexical research, and are included in other lexicons such as the French lexicon Lexique (New et al., 2004). Phonological neighborhood density data in particular is central to many studies of the lexicon (e.g. Stokes, 2010; Vitevitch and Luce, 2016), but existing resources only provide data for a limited number of languages – for example, the CLEARPOND database with 5 languages (Marian et al.,

[5] https://huggingface.co/charsiu/g2p_multilingual_byT5_small_100

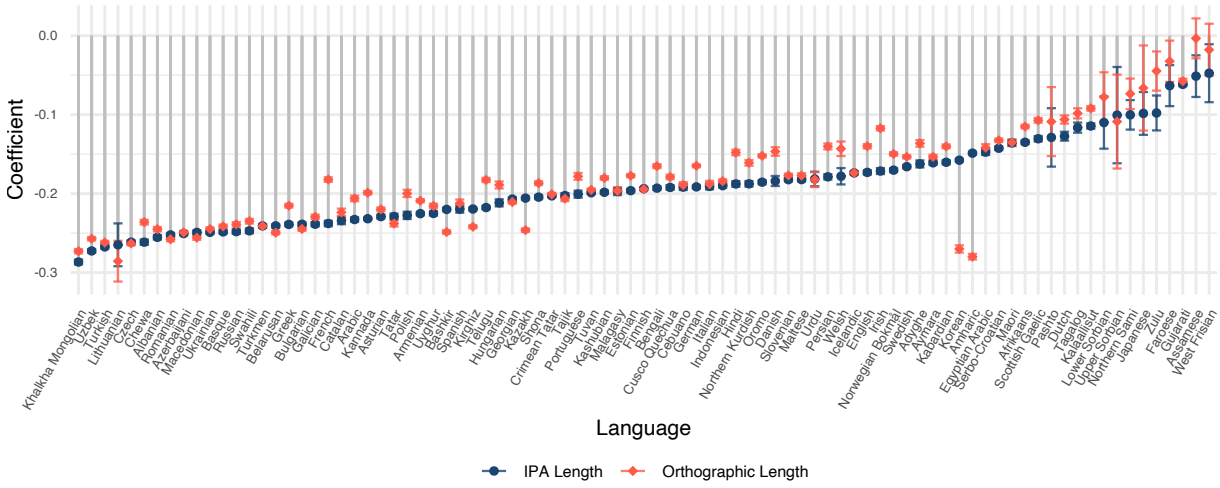

Figure 2: Linear regression coefficients for phonotactic complexity and word length in each language. IPA word lengths are the mean number of IPA symbols for all transcription methods available for each language.

2012), and Jiwar with 40 languages (Alzahrani, 2025). These measures are calculated automatically for each transcription method, and additional measures can be included in the lexicons using the software provided.

4. Validation

4.1. Distance Between Transcriptions

Due to the size of this dataset, manual validation of the IPA transcriptions is not possible. However, most languages in the dataset have multiple transcription methods available. While we cannot compare these transcriptions to gold-standard human-written transcriptions, we can compare them to each other. If the transcription systems are accurate, transcriptions of the same word should be similar, but because of pronunciation variations or slight differences in transcription systems, they are rarely identical.

In order to validate the automatically generated transcriptions, we calculate a phonological feature-based Levenshtein distance between transcriptions using Panphon (Mortensen et al., 2018). In this distance metric, the cost of substituting one IPA symbol for another corresponds to the number of phonological features that differ between them, rather than a fixed cost per substitution. For example, [a] is more similar to [ɑ] than it is to [i].

We calculate this distance between each transcription of each word, and take the mean of all distance measures for languages with more than two transcriptions per word, then normalize it by dividing by the mean length of all transcriptions. The resulting distance represents the average number of feature edits per character to transform one transcription into another. The distributions of these distances in each language are shown in Figure 1. Many languages have an average distance of less than one feature edit per character, consistent with minor phonetic differences between transcriptions, suggesting that transcriptions are fairly consistent in most languages. Languages with larger distances between transcriptions are less consistent, and should be examined more closely before using them in any analysis. It is possible that in these cases, one transcription system is systematically wrong.

4.2. Valid IPA Symbols

We also calculate the proportion of words containing valid IPA symbols for each language and transcription method, shown in Table 2. Transcription methods where this proportion is low for a particular language are perhaps unreliable. For example, 35% of Epitran transcriptions for Aymara include invalid IPA symbols, suggesting that these transcriptions may not be reliable.

4.3. Relationships Between Lexical Variables

As a final validation of this dataset, we examine several well-known relationships between lexical variables. Zipf's law of abbreviation (Zipf, 1935) states that there is a negative correlation between word length and frequency, which was been consistently demonstrated in many languages (Strauss et al., 2005; Bentz and Ferrer-i Cancho, 2016; Doucette et al., 2024), often using orthographic

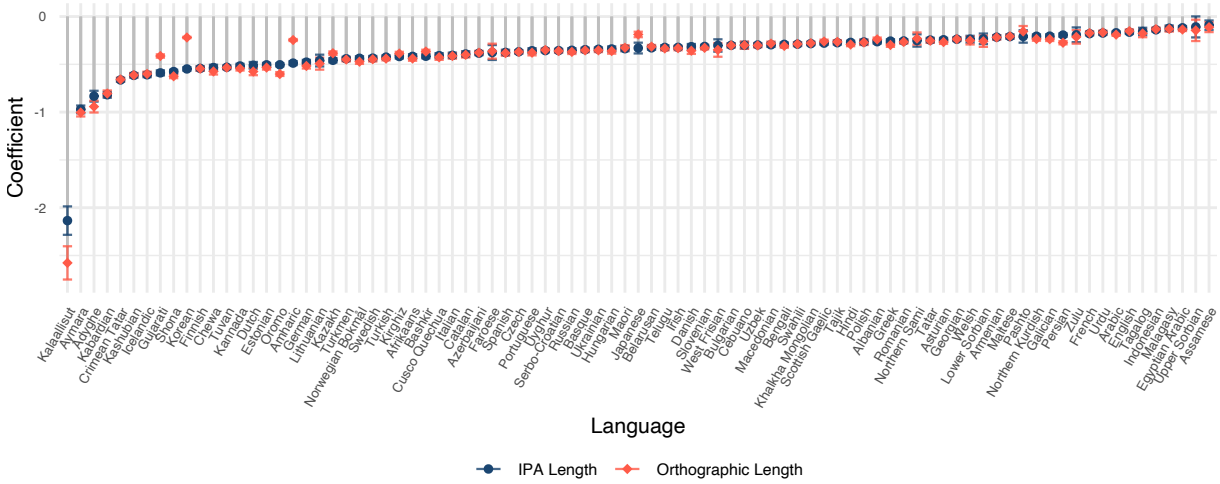

Figure 3: Linear regression coefficients for word frequency and word length in each language. IPA word lengths are the mean number of IPA symbols for all transcription methods available for each language.

length in large datasets, large multilingual word lists like ASJP (Brown et al., 2008), or datasets representing a specific subset of the lexicon like UniMorph (Batsuren et al., 2022). For each language, we fit two linear regression models: one predicting word length in phones from frequency, and another predicting orthographic word length from frequency. Regression coefficients for each language are shown in Figure 3. As expected, all languages in our dataset have a negative relationship between word length and frequency, although there is variation in the magnitude of the coefficient across languages. Following Doucette et al. (2024), we also fit linear mixed effects regression models with random intercepts and slopes by language, and find an effect of -0.38 (95% CI [-0.43, -0.32], $p < 0.001$, rand. eff. st. dev. = 0.25) in the IPA model and an effect of -0.38 (95% CI [-0.44, -0.32], $p < 0.001$, rand. eff. st. dev. = 0.29) in the orthographic model.

We also examine the relationship between word length and phonotactic complexity. Phonotactic complexity is calculated for each word following the method used by Pimentel et al. (2020) and Doucette et al. (2024), which results in a surprisal-like measure. Using data from NorthEuraLex (Dellert et al., 2020), Pimentel et al. (2020) find a consistently negative correlation between word length and phonotactic complexity both within and across languages. However, in the UniMorph dataset (Batsuren et al., 2022), which mainly contains morphologically complex words, Doucette et al. (2024) find a positive correlation within languages, and no clear relationship across languages. Again, we fit a linear regression model for each language, predicting phonotactic complexity from word length, and show the correlations in Figure 2. All coefficients are negative, suggesting that the positive correlations found in UniMorph are the result of sampling only morphologically complex words in the lexicon. We also fit a linear mixed effects regression model predicting phonotactic complexity from word length and mean word length within language, with random slopes and intercepts by language.[6] We find an effect of word length of -0.19 (95% CI [-0.20, -0.18], $p < 0.001$, rand. eff. st. dev. = 0.05) and an effect of mean word length of -0.05 (95% CI [-0.08, -0.02], $p = 0.001$), corroborating Pimentel et al.'s (2020) findings of a negative correlation between phonotactic complexity and word length both within and across languages.

5. The Lexicons

LexiPhon currently contains phonetically transcribed lexicons for 87 languages, listed in Table 2. These lexicons range in size from 13,298 to 158,274 words, with a mean of 56,899 and a median of 53,157. The distribution of the number of words per lexicon is shown in Figure 4.

For each language, we calculated the mean word length for each of the four G2P methods, listed in Table 2. A density plot of these means is shown in Figure 5. The three automated G2P methods (XPF, Epitran, and CharsiuG2P) have similar mean word length distributions, while WikiPron transcriptions skew toward shorter word lengths. WikiPron transcriptions are written by volunteers, which could explain this tendency towards shorter words: perhaps Wiktionary editors prioritize transcribing shorter,

[6] in `lme4` syntax: `phonComplexity ~ wordLength + meanWordLength + (1 + wordLength | language)`

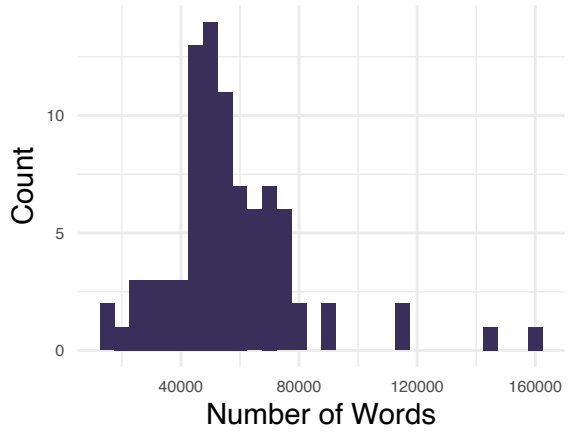

Figure 4: Number of tokens per language.

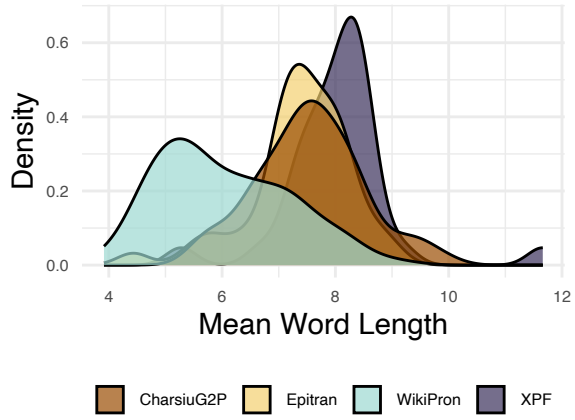

Figure 5: Density plot of mean word lengths within each language, for each G2P method.

more frequent words. This suggests that WikiPron transcriptions may not be representative of an entire language's lexicon and should be used with caution, particularly in languages where there are very few entries.

Each G2P method uses a different set of IPA symbols for transcription. While XPF and Epitran provide a fixed set of symbols for each language, CharsiuG2P and WikiPron are able to use any IPA symbol in their transcriptions. IPA symbol inventory sizes for each language are shown in Table 2, where we can see that CharsiuG2P and WikiPron generally use larger symbol inventories than XPF and Epitran. This suggests that these methods either provide narrower transcriptions than XPF and Epitran or are more prone to error. In order to confirm this, the transcriptions will need to be verified by native speakers.

6. The Software

Along with the lexicons, we release open-source software for both creating new lexicons and augmenting the existing lexicons with additional data. We provide a Python script that downloads a Wikipedia dump for a particular language, then runs the entire LexiPhon processing pipeline: Frequency counting, filtering, G2P transcription and normalization, and calculating word lengths, phonological neighborhood densities, and distances between transcriptions. Additional transcription methods can be included by providing a TSV file listing each word followed by its transcription. These additional transcriptions are normalized following the same process as those provided in LexiPhon. There is also an option to provide additional data as TSV files, which are added as additional data columns in the output lexicon.

7. Discussion

In this paper, we introduce a dataset of phonetically transcribed lexicons with word frequencies generated from Wikipedia in 87 languages, along with open-source software to generate new lexicons. These lexicons are much larger than previously available multilingual word lists, which contain only tens to hundreds of words. The code provided allows lexicons to be produced for any language in Wikipedia that is covered by one or more of the G2P methods. We also provide simple internal validations showing consistency between transcription methods in lieu of manual verification. While some individual transcriptions may be inaccurate, these lexicons can be used for comparing aggregate properties of lexicons across languages – for example, investigating relationships between phonological neighborhood density and word length. We hope these lexicons will provide a starting point for further development of multilingual phonetic lexicons. In future work, these datasets could be improved by human verification, by including more G2P methods, and by incorporating additional pronunciation dictionaries.

Table 2: LexiPhon languages, total number of words, number of WikiPron words, mean feature edit distance between transcriptions, number of phones for each transcription method, and proportion of words with only valid IPA symbols. EPI: Epitran; CS: Charsiu; WP: WikiPron.

			Number of IPA symbols				Proportion Valid IPA			
Lang.	N (WP)	Dist.	XPF	EPI	CS	WP	XPF	EPI	CS	WP
Adyghe	18810 (829)	0.18	–	–	143	63	–	–	0.99	0.99
Afrikaans	42845 (1363)	0.32	–	–	53	55	–	–	1	1
Albanian	51958 (764)	0.31	32	50	86	65	0.97	0.97	0.93	1
Amharic	73816 (282)	1.63	–	38	51	60	–	0.99	0.81	1
Arabic	58450 (5105)	2.09	–	52	82	59	–	0.69	1	1
Armenian	56124 (6493)	0.7	32	–	60	37	1	–	1	1
Assamese	49930 (1586)	–	–	–	–	46	–	–	–	0.87
Asturian	47890 (594)	0.15	24	–	–	31	1	–	–	1
Aymara	61256 (0)	–	33	–	–	–	0.35	–	–	–
Azerbaijani	63225 (1889)	0.3	30	42	66	57	0.95	0.94	0.99	1
Bashkir	55812 (1651)	0.93	35	–	49	100	0.99	–	1	0.98
Basque	58083 (2135)	0.33	26	–	46	36	0.88	–	0.88	1
Belarusan	66816 (3057)	0.37	46	–	56	58	0.98	–	0.61	0.91
Bengali	44818 (1722)	1.67	–	52	175	67	–	0.98	0.38	0.91
Bulgarian	54835 (6423)	0.42	45	–	95	53	0.99	–	1	1
Catalan	44742 (14225)	0.62	–	41	51	38	–	0.91	1	1
Cebuano	17477 (758)	0.78	–	23	–	36	–	1	–	1
Chewa	32323 (329)	0.94	–	63	–	41	–	0.96	–	0.59
Crimean Tatar	117033 (0)	–	51	–	–	–	0.92	–	–	–
Cusco Quechua	53159 (0)	–	29	–	–	–	0.5	–	–	–
Czech	70416 (12862)	0.41	32	36	69	45	0.96	0.97	0.96	0.95
Danish	52350 (2611)	1.24	–	–	135	103	–	–	0.29	0.98
Dutch	49203 (9572)	0.8	–	48	61	50	–	0.97	1	1
Egyptian Arabic	28557 (107)	1.68	–	–	56	54	–	–	1	1
English	39103 (16423)	0.8	–	47	44	60	–	1	1	1
Estonian	75862 (1005)	0.68	–	–	103	56	–	–	0.93	0.66
Faroese	74587 (1114)	–	–	–	–	79	–	–	–	0.85
Finnish	79246 (12899)	0.89	–	–	44	61	–	–	1	0.83
French	44515 (13803)	0.82	–	47	50	40	–	0.92	0.98	0.99
Galician	47401 (1664)	0.12	–	–	77	57	–	–	1	1
Georgian	66931 (8927)	0.5	32	–	43	34	0.99	–	1	1
German	54234 (3320)	1	–	46	88	62	–	0.99	1	1
Greek	50862 (3159)	0.86	23	–	36	33	1	–	1	1
Gujarati	52780 (1312)	4.37	–	–	200	66	–	–	0.84	0.89
Hindi	34948 (9589)	0.64	–	65	63	55	–	0.99	0.45	0.88
Hungarian	71818 (20042)	0.41	55	63	100	70	0.94	0.99	1	1
Icelandic	59025 (4296)	0.09	–	–	98	60	–	–	0.99	1
Indonesian	45077 (2810)	0.4	–	28	33	54	–	1	0.96	0.94
Irish	44690 (4129)	0.61	–	–	100	118	–	–	0.91	0.91
Italian	52010 (16564)	0.64	–	48	53	31	–	0.93	1	1
Japanese	48461 (1652)	0.35	–	–	109	39	–	–	0.97	0.46
Kabardian	48047 (513)	0.57	47	96	–	65	0.96	0.92	–	1
Kalaallisut	13298 (304)	1.45	34	–	–	30	0.73	–	–	0.98
Kannada	66756 (573)	0.78	61	–	–	58	0.89	–	–	0.95
Kashubian	76210 (539)	1.12	–	42	–	60	–	0.95	–	0.96

Table 2: LexiPhon languages, with Wikipedia codes, number of words, mean word lengths and number of phones for each G2P method. EPI: Epitran; CS: Charsiu; WP: WikiPron. *(continued)*

Lang.	N (WP)	Dist.	XPF	EPI	CS	WP	XPF	EPI	CS	WP
Kazakh	56909 (936)	0.84	–	36	39	90	–	0.97	1	0.99
Khalkha Mongolian	54662 (1679)	0.58	–	75	–	100	–	0.96	–	0.91
Kirghiz	60778 (381)	0.84	33	28	–	52	0.91	0.97	–	1
Korean	89774 (9186)	0.77	26	–	51	48	0.92	–	0.51	0.45
Lithuanian	74818 (1958)	0.17	–	–	166	95	–	–	1	0.96
Lower Sorbian	74881 (1051)	–	–	–	–	55	–	–	–	1
Macedonian	50744 (14973)	0.41	34	–	48	44	0.8	–	1	1
Malagasy	22661 (99)	1.88	21	–	–	53	0.66	–	–	0.64
Maltese	50969 (4646)	1.33	–	56	59	41	–	0.93	0.88	1
Maori	27467 (0)	–	–	33	–	–	–	0.85	–	–
Northern Kurdish	49861 (1164)	0.65	–	–	180	53	–	–	1	1
Northern Sami	78477 (1805)	0.05	–	–	113	71	–	–	1	0.98
Norwegian Bokmål	53168 (748)	0.59	–	–	66	65	–	–	1	1
Oromo	88679 (0)	–	–	98	–	–	–	0.87	–	–
Pashto	39477 (673)	–	–	–	–	60	–	–	–	0.98
Persian	36683 (5684)	2.29	–	44	65	94	–	0.75	1	0.9
Polish	70942 (19847)	0.35	–	38	80	43	–	0.91	0.92	0.93
Portuguese	45829 (14552)	1.24	–	43	59	43	–	0.98	1	1
Romanian	55331 (2955)	0.44	27	29	82	58	0.92	0.99	0.95	1
Russian	70231 (39334)	0.52	–	46	97	93	–	0.99	0.97	0.97
Scottish Gaelic	43532 (1914)	2.08	–	–	142	115	–	–	0.98	0.95
Serbo-Croatian	61701 (7030)	1.64	–	34	63	43	–	0.99	1	1
Shona	62974 (0)	–	–	44	–	–	–	0.89	–	–
Slovenian	69766 (2427)	0.46	–	–	69	37	–	–	1	1
Spanish	44008 (14384)	0.23	–	28	51	27	–	0.98	1	1
Swahili	40317 (74)	0.23	–	56	93	40	–	0.96	0.83	0.8
Swedish	53200 (1988)	0.8	–	38	126	85	–	0.98	0.8	0.81
Tagalog	44446 (5790)	0.79	–	–	45	28	–	–	1	1
Tajik	43851 (464)	0.33	29	33	–	42	0.97	0.89	–	1
Tatar	28362 (0)	0.29	33	–	39	–	0.95	–	1	–
Telugu	51089 (1624)	0.72	59	48	–	78	0.92	1	–	0.91
Turkish	70546 (3589)	0.43	44	35	64	102	0.97	0.93	1	0.99
Turkmen	64194 (110)	0.27	–	33	62	62	–	0.99	1	0.96
Tuvan	117139 (457)	0.64	39	–	–	81	0.79	–	–	0.97
Ukrainian	67602 (16285)	0.52	32	43	109	82	1	0.98	0.99	1
Upper Sorbian	158276 (256)	1.09	34	–	–	49	0.91	–	–	0.99
Urdu	26652 (2830)	2.71	–	26	–	87	–	0.29	–	0.95
Uyghur	47657 (721)	0.39	29	31	–	44	0.95	0.96	–	0.99
Uzbek	60247 (0)	–	31	51	–	–	0	0.96	–	–
Welsh	35787 (4768)	0.26	–	–	55	44	–	–	1	1
West Frisian	45467 (684)	–	–	–	–	51	–	–	–	1
Zulu	145363 (1020)	1.85	–	41	–	54	–	0.95	–	0.92

8. Bibliographical References

Cormac Anderson, Tiago Tresoldi, Thiago Chacon, Anne-Maria Fehn, Mary Walworth, Robert Forkel, and Johann-Mattis List. 2018. A cross-linguistic database of phonetic transcription systems. In *Yearbook of the Poznan Linguistic Meeting*, volume 4, pages 21–53.

R. Harald Baayen. 2001. *Word Frequency Distributions*, volume 18 of *Text, Speech and Language Technology*. Springer Dordrecht.

Christian Bentz and Ramon Ferrer-i Cancho. 2016. Zipf's law of abbreviation as a language universal. In *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics*.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly.

Cecil H Brown, Eric W Holman, Søren Wichmann, and Viveka Velupillai. 2008. Automated classification of the world's languages: a description of the method and preliminary results. *Language Typology and Universals*, 61(4):285–308.

Amanda Doucette, Ryan Cotterell, Morgan Sonderegger, and Timothy J. O'Donnell. 2024. Correlation does not imply compensation: Complexity and irregularity in the lexicon. In *Proceedings of the Society for Computation in Linguistics 2024*, pages 117–128.

Viviana Fratini, Joana Acha, and Itziar Laka. 2014. Frequency and morphological irregularity are independent variables. Evidence from a corpus study of Spanish verbs. *Corpus Linguistics and Linguistic Theory*, 10(2):289–314.

Jennifer Hay and Harald Baayen. 2003. Phonotactics, parsing and productivity. *Italian Journal of Linguistics*, 15:99–130.

Jackson L. Lee, Lucas F. E. Ashby, M. Elizabeth Garza, Yeonju Lee-Sikka, Sean Miller, Alan Wong, Arya D. McCarthy, and Kyle Gorman. 2020. Massively multilingual pronunciation modeling with WikiPron. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4223–4228.

Paul A Luce and David B Pisoni. 1998. Recognizing spoken words: The neighborhood activation model. *Ear and hearing*, 19(1):1–36.

Kyle Mahowald, Isabelle Dautriche, Edward Gibson, and Steven T. Piantadosi. 2018. Word forms are structured for efficient use. *Cognitive Science*, 42(8):3116–3134.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.

Tiago Pimentel, Brian Roark, and Ryan Cotterell. 2020. Phonotactic complexity and its trade-offs. *Transactions of the Association for Computational Linguistics*, 8:1–18.

Ting Qian, Kristy Hollingshead, Su-youn Yoon, Kyoung-young Kim, and Richard Sproat. 2010. A Python toolkit for universal transliteration. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50.

Stephanie F. Stokes. 2010. Neighborhood density and word frequency predict vocabulary size in toddlers. *Journal of Speech, Language, and Hearing Research*, 53(3):670–683.

Udo Strauss, Peter Grzybek, and Gabriel Altmann. 2005. Word length and word frequency. In Peter Grzybek, editor, *Contributions to the Science of Text and Language: Word Length Studies and Related Issues*, pages 277–294. Springer.

Morris Swadesh. 1955. Towards greater accuracy in lexicostatistic dating. *International journal of American linguistics*, 21(2):121–137.

Kazuma Takaoka, Sorami Hisamoto, Noriko Kawahara, Miho Sakamoto, Yoshitaka Uchida, and Yuji Matsumoto. 2018. Sudachi: a Japanese tokenizer for business. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*.

Michael S. Vitevitch and Paul A. Luce. 2016. Phonological neighborhood effects in spoken word perception and production. *Annual Review of Linguistics*, 2:75–94.

Søren Wichmann and Eric W Holman. 2023. Cross-linguistic conditions on word length. *PloS one*, 18(1):e0281041.

Wikipedia contributors. 2025. List of wikipedias — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=List_of_Wikipedias&

oldid=1262630444. [Online; accessed 3-October-2025].

Shijie Wu, Ryan Cotterell, and Timothy O'Donnell. 2019. Morphological irregularity correlates with frequency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5117–5126. Association for Computational Linguistics.

Jian Zhu, Cong Zhang, and David Jurgens. 2022. ByT5 model for massively multilingual grapheme-to-phoneme conversion.

George Kingsley Zipf. 1935. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Houghton Mifflin.

9. Language Resource References

Emily Ahn and Eleanor Chodroff. 2022. *VoxCommunis: A Corpus for Cross-linguistic Phonetic Analysis*. PID https://doi.org/10.17605/OSF.IO/T957V.

Alaa Alzahrani. 2025. *Jiwar: A database and calculator for word neighborhood measures in 40 languages*. PID https://github.com/AlaaAlzahrani/Jiwar_database.

R H. Baayen and R Piepenbrock and L Gulikers. 1995. *CELEX2 LDC96L14*. Linguistic Data Consortium. PID https://doi.org/10.35111/gs6s-gm48.

Khuyagbaatar Batsuren and Omer Goldman and Salam Khalifa and Nizar Habash and Witold Kieraś and Gábor Bella and Brian Leonard and Garrett Nicolai and Kyle Gorman and Yustinus Ghanggo Ate and Maria Ryskina and Sabrina Mielke and Elena Budianskaya and Charbel El-Khaissi and Tiago Pimentel and Michael Gasser and William Abbott Lane and Mohit Raj and Matt Coler and Jaime Rafael Montoya Samame and Delio Siticonatzi Camaiteri and Esaú Zumaeta Rojas and Didier López Francis and Arturo Oncevay and Juan López Bautista and Gema Celeste Silva Villegas and Lucas Torroba Hennigen and Adam Ek and David Guriel and Peter Dirix and Jean-Philippe Bernardy and Andrey Scherbakov and Aziyana Bayyr-ool and Antonios Anastasopoulos and Roberto Zariquiey and Karina Sheifer and Sofya Ganieva and HilariaCruz and Ritván Karahóǧa and Stella Markantonatou and George Pavlidis and Matvey Plugaryov and Elena Klyachko and Ali Salehi and Candy Angulo and Jatayu Baxi and Andrew Krizhanovsky and Natalia Krizhanovskaya and Elizabeth Salesky and Clara Vania and Sardana Ivanova and Jennifer White and Rowan Hall Maudslay and Josef Valvoda and Ran Zmigrod and Paula Czarnowska and Irene Nikkarinen and Aelita Salchak and Brijesh Bhatt and Christopher Straughn and Zoey Liu and Jonathan North Washington and Yuval Pinter and Duygu Ataman and Marcin Wolinski and Totok Suhardijanto and Anna Yablonskaya and Niklas Stoehr and Hossep Dolatian and Zahroh Nuriah and Shyam Ratan and Francis M. Tyers and Edoardo M. Ponti and Grant Aiton and Aryaman Arora and Richard J. Hatcher and Ritesh Kumar and Jeremiah Young and Daria Rodionova and Anastasia Yemelina and Taras Andrushko and Igor Marchenko and Polina Mashkovtseva and Alexandra Serova and Emily Prud'hommeaux and Maria Nepomniashchaya and Fausto Giunchiglia and Eleanor Chodroff and Mans Hulden and Miikka Silfverberg and Arya D. McCarthy and David Yarowsky and Ryan Cotterell and Reut Tsarfaty and Ekaterina Vylomova. 2022. *UniMorph 4.0: Universal Morphology*. European Language Resources Association. PID https://unimorph.github.io/.

Johannes Dellert and Thora Daneyko and Alla Münch, Alina Ladygina and Armin Buch and Natalie Clarius and Ilja Grigorjew, Mohamed Balabel and Hizniye Isabella Boga and Zalina Baysarova and Roland Mühlenbernd and Johannes Wahle and Gerhard Jäger. 2020. *NorthEuraLex: a wide-coverage lexical database of Northern Eurasia*. Springer. PID https://doi.org/10.1007/s10579-019-09480-6. Version 0.9.

Leo Gao and Stella Biderman and Sid Black and Laurence Golding and Travis Hoppe and Charles Foster and Jason Phang and Horace He and Anish Thite and Noa Nabeshima and Shawn Presser and Connor Leahy. 2020. *The Pile: An 800GB Dataset of Diverse Text for Language Modeling*. PID https://pile.eleuther.ai/.

Johann-Mattis List and Robert Forkel and Simon J Greenhill and Christoph Rzymski and Johannes Englisch and Russell D Gray. 2022. *Lexibank, a public repository of standardized wordlists with computed phonological and lexical features*. PID https://doi.org/10.5281/zenodo.15194559.

Viorica Marian and James Bartolotti and Sarah Chabal and Anthony Shook. 2012. *CLEARPOND: Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities*. PID https://doi.org/10.1371/journal.pone.0043230.

David R. Mortensen and Siddharth Dalmia and Patrick Littell. 2018. *Epitran: Pre-*

cision G2P for Many Languages. PID https://github.com/dmort27/epitran.

Boris New and Christophe Pallier and Marc Brysbaert and Ludovic Ferrand. 2004. *Lexique 2: A new French lexical database*. PID doi.org/10.3758/bf03195598.

Guilherme Penedo and Quentin Malartic and Daniel Hesslow and Ruxandra Cojocaru and Alessandro Cappelli and Hamza Alobeidli and Baptiste Pannier and Ebtesam Almazrouei and Julien Launay. 2023. *The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only*. PID https://doi.org/10.57967/hf/0737.

Uriel Cohen Priva and Emily Strang and Shiying Yang and William Mizgerd and Abigail Creighton and Justin Bai and Rebecca Mathew and Allison Shao and Jordan Schuster and Daniela Wiepert. 2021. *The Cross-linguistic Phonological Frequencies (XPF) Corpus manual*. PID https://cohenpr-xpf.github.io/XPF/.

Rosetta Type Foundry. 2025. *Hyperglot v.0.7.2*. PID https://github.com/rosettatype/hyperglot/releases/tag/0.7.2.

Elizabeth Salesky and Eleanor Chodroff and Tiago Pimentel and Matthew Wiesner and Ryan Cotterell and Alan W Black and Jason Eisner. 2020. *A Corpus for Large-Scale Phonetic Typology*. Association for Computational Linguistics. PID https://osf.io/bc2ns.

Jonne Sälevä and Constantine Lignos. 2024. *ParaNames 1.0: Creating an Entity Name Corpus for 400+ Languages Using Wikidata*. PID https://github.com/bltlab/paranames/releases/tag/v2024.05.07.0.

Tiago Tresoldi. 2023. *A Global Lexical Database (GLED) for Computational Historical Linguistics*. PID https://doi.org/10.5281/zenodo.7368116.