EnoLEX: A Diachronic Lexical Database for the Enggano language

Gede Primahadi W. Rajeg [1, 3], Daniel Krauße [2], Cokorda Pramartha [3]

[1] University of Oxford, gede.rajeg@ling-phil.ox.ac.uk; primahadi_wijaya@unud.ac.id

[2] Lattice-CNRS, PSL, ENS, daniel.krausse@ens.psl.eu

[3] CIRHSS, Udayana University, cokorda@unud.ac.id

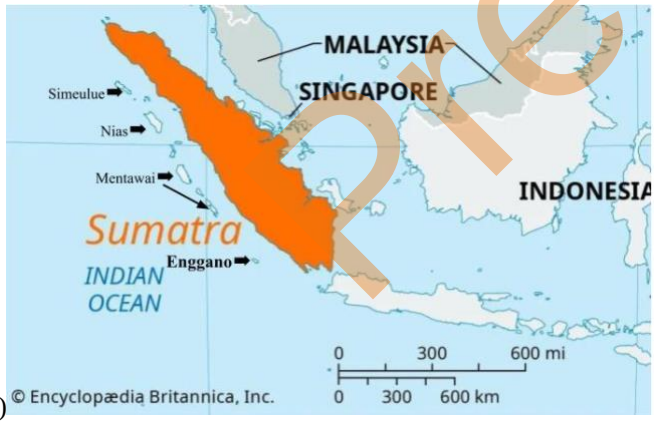
Abstract

In an effort to document and understand the diachronic change of the Enggano language, we present EnoLEX, a diachronic lexical database comprising legacy material starting as early as 1850. We draw on records of different shapes and sizes written in several source languages. EnoLEX is complemented with fieldwork materials of contemporary Enggano gathered between 2018-2024. The curation process began with the etymologisation of word forms to ensure that historically unrelated forms are assigned individual IDs even though they were glossed the same in the original records. The second step encompassed the retro-digitisation of the Enggano-German and Dutch-Enggano dictionaries. Further necessary steps were (i) the semantic mapping of the English gloss to the Concepticon catalogue, (ii) orthographies standardisation across the wordlists, and (iii) phonemic transcription. The final database is implemented using *Golang* (back-end) and *React* (front-end). EnoLEX is the first diachronic database of its kind for an endangered language in Indonesia, spanning written materials of more than 150 years and providing Proto-Malayo-Polynesian and Proto-Austronesian reconstructions that are directly linked to the Austronesian Comparative Dictionary. EnoLEX thus allows historical, statistical, dialectal, and comparative analyses of phonological, semantic, and lexical change within the same language. Our work contributes to the growing presence of lexical databases for linguistics and researchers investigating human prehistory and cultural evolution.

Keywords: historical database, lexical database, Austronesian, Enggano

1. Introduction

Enggano (Glottocode: *engg1245*, ISO code: *eno*) is an endangered Austronesian language in Indonesia spoken by around 1,500 speakers on Enggano island. The island is the southernmost one in the chain of Barrier Islands (see Figure 1a, adjusted from Encyclopædia Britannica), stretching northward through Mentawai, Nias, and Simeulue islands, off the west coast of Sumatra. Despite Enggano's threatened condition, whereby speakers increasingly shift towards Indonesian (cf. Arka et al., 2022), it is still vital in the three central villages: Meok, Apoho, and Malakoni (Figure 1b).

(a) 

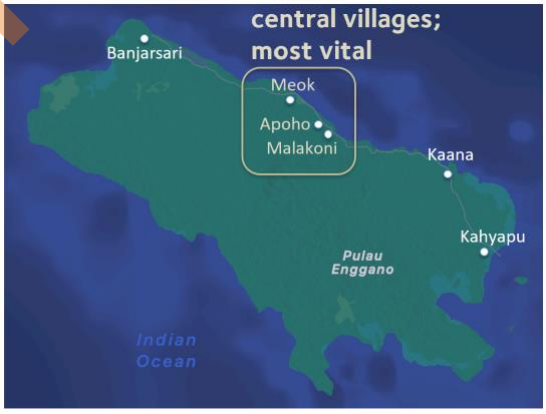
(b) 

Figure 1. Map of Enggano (b) in the context of the Barrier Islands chain off the west coast of Sumatra (a).

The Enggano language has attracted interest from the 1850s up to the present, resulting in rich legacy materials, especially those gathered during the colonial period of Indonesia. These legacy materials include early wordlists (cf. Krauße, *forthcoming* a), early attempts to determine its genetic affiliation (cf. Krauße, *forthcoming* b) as well as grammar sketches (Kähler, 1940) and Enggano texts produced by German linguist Hans Kähler from his field trip to Enggano in the 1930s. Kähler's Enggano dictionary (1987) was published posthumously with the assistance of his student Hans Schmidt. Most works of the 20th and 21st century on Enggano focus on linguistic aspects, particularly in resolving the long-standing debate on the status of Enggano within the Austronesian language family (§2) (Krauße, *forthcoming* b; Billings & McDonnell, 2024; Edwards, 2015; Nothofer, 1986; cf. Hemmings, 2024; Rajeg et al., 2024).

Our team seeks to document and understand the diachronic change of Enggano. For this, a diachronic lexical database comprising legacy material from as early as 1850 up to the contemporary language is essential, which we call "EnoLEX",

based on the ISO code *eno* for 'Enggano'. The legacy material consists of records of different shapes, sizes, and source languages (Table 1). Some wordlists of the 19th century are as short as 24 words (Boewang, 1854), while others are very detailed, one even comprising over 1,000 lexemes (Helfrich & Pieters, 1891). Records before the 19th century are almost exclusively written in Dutch, but there is also material in Italian (Modigliani, 1894) and Malay (Helfrich & Pieters, 1891). One wordlist only exists in handwritten form (Brouwer, n.d.), whereas all others were published in a machine-readable format. Material of the 20th and 21st centuries continues in Dutch (Helfrich, 1916), German (Capell, 1982; Kähler, 1987), and Indonesian (Amran et al., 1979; Kasim et al., 1987; Yoder, 2011). Our database is complemented with fieldwork material of contemporary Enggano gathered between 2018-2024. This collection allows us to see the detailed lexical development of Enggano from the 19th to the 21st century.

Although lexical databases for other languages of the Indonesian archipelago exist, such as the Austronesian Basic Vocabulary (Greenhill et al., 2008), Jakarta Lexical Database (Tadmor & Gil, 2015), TransNewGuinea (Greenhill, 2015), and LexiRumah (Kaiping & Klamer, 2018; Kaiping et al., 2019), their diachronic focus is either mainly areal or cross-linguistic. EnoLEX is the first of its kind because its historical analysis revolves around a single endangered language in Indonesia (§4.1). EnoLEX also implements core elements of the Cross-Linguistic Data Format (CLDF) into diachronic datasets (§4.2 & §4.3) (Forkel & List, 2020). The entries in EnoLEX with records of more than 150 years are directly linked to established Proto-Malayo-Polynesian and Proto-Austronesian reconstructions in the Austronesian Comparative Dictionary (Blust et al., 2023). Our database therefore facilitates historical, statistical, dialectal, and comparative analyses of phonological, semantic, and lexical change within the same language. The database is implemented for the user interface using *Golang* (back-end) and *React* (front-end) (§5.2).

2. Motivation

The Enggano language has sometimes been referred to as a 'linguistic puzzle' (Butters, 2021, p. 25; Kraußе, 2024). Several features that are either absent from or at least atypical of Austronesian languages have been noted, such as:

- lexical nasalization and word-level nasal harmony (cf. Smith, 2017; Hemmings & Tan, *forthcoming*)
- noun articles and grammatical cases in Old Enggano (cf. Hemmings, *forthcoming*)
- very low lexical cognacy with other Austronesian languages (Edwards, 2015, p. 76)

Scholars have therefore raised questions about the genetic affiliation of the Enggano language. The two principal hypotheses used to be that it is either an Austronesian language with very unusual but regular sound changes (Dyen, 1965; Nothofer, 1986; Mahdi, 1988, pp. 59–61) or that it is a non-Austronesian language with loanwords from Austronesian languages (Capell, 1982, p. 6; Blench, 2014).[1] More recent publications (Edwards, 2015; Smith, 2017; Billings & McDonnell, 2024; Kraußе, *forthcoming* b) have presented sufficient evidence that Enggano is indeed an Austronesian language, although the internal subgrouping of the Sumatran languages including Enggano is still debated.

Our motivation for EnoLEX is to understand the dynamics of phonological, morphological, and lexical changes in Enggano. We try to answer the following research questions:

- Why do the old wordlists deviate so much from each other?
- What caused the rapid lexical replacements within such a small language?
- What influences from other languages were there in the past?
- What did the dialectal situation of Enggano in the 19th century look like?
- Can we detect more Austronesian cognates that have so far been overlooked?

Only through large-scale comparison of all available materials from different periods are we able to draw conclusions and answer our questions above. In a broader context, EnoLEX contributes further cross-linguistic data to the growing interest in the creation of large-scale lexical databases, such as the Lexibank project[2] (List et al., 2022). In addition, EnoLEX offers a new dimension, namely diachrony, to the predominantly cross-linguistic datasets existing to date. Finally, EnoLEX will preserve the Old and Contemporary Enggano language for future generations.

3. Data sources

Table 1 provides information on each record that is included in EnoLEX. Many of the sources in Table 1 were collated in the first phase of the Enggano project. In the second phase, we added, among others, the fully digitised dictionary by Kähler (1987), the Dutch-Enggano dictionary by Helfrich (1916), as well as the Enggano "Holle List" (hereafter HL) (Stokhof & Almanar, 1987; Rajeg, 2023a). An external team of research assistants from Indonesia manually entered all

[1] About 150 years ago, Rosenberg (1878, p. 217) had already remarked that "the [Enggano] language does not have the slightest resemblance with the idioms of the neighbouring peoples". Two decades later, Modigliani (1894, p. 295) attempted a comparison of Enggano and the Nicobarese languages, which are now classified as Austro-Asiatic, but except for superficial physical similarities between the two peoples, he could not demonstrate any closer relationship.

[2] https://lexibank.clld.org/

entries of the Kähler dictionary into an online database storage (Rajeg et al., 2023). The next step for Kähler's digitised dictionary involved the correction of unrecognised characters after the transcribed entries had been exported into a .csv file for further processing. For example, the acute accent (◌́) had been rendered as an equal sign above a vowel (e.g., *epā̄ra* instead of *epára* 'climbing plant' [Kähler, 1987, p. 239]), and the common transcription symbol for diphthongs (:) had been recognised as a subscript left arrow (e.g., *kiko?a̲i̲xa* instead of *kiko?a:ixa* 'becoming night' [Kähler, 1987, p. 156]). Following the manual correction of such entries[3], the German definitions of the Enggano forms were translated into English, which were then checked by a research assistant fluent in German and English. The Dutch-Enggano dictionary (Helfrich, 1916) was retro-digitised through OCR with further manual corrections, including the manual correction of the English translations by a research assistant fluent in Dutch and English.[4] Finally, the Enggano words in the HL were OCR-ed, manually proofread, and computationally matched for their English equivalents from the digitised reference HL[5] (Stokhof, 1980; Rajeg, 2023b). Our contemporary data (2018-2024) is drawn from recent fieldwork that includes traditional stories and wordlist elicitation materials (Sangian et al., 2024).

Table 1: Comparison of all wordlists in EnoLEX.

Published	Collected	Author(s)	Lexemes	Dialect	Place
—	ca. 1850	Brouwer	104	northwest	Barhau
1854	1840-1850	Boewang	22	?	?
1855	1854	Van der Straaten & Severijn	199	northwest	Karkau
1855	1852	Von Rosenberg	154	northwest or south	Barhau
1864	1863	Walland	276	north	?
1870	1865-1870	Francis	91	northwest	Barhau?
1879	—	Oudemans	141	?	—
1888	1885	Helfrich	~460	south	Kioyo
1891 & 1893[6]	1891	Helfrich & Pieters	1,012	southeast & northwest	Pulau Dua & Karkua
1894	1891	Modigliani	484	southeast?	Kayaapu
1987	1895	Stokhof [7]	878	southeast?	Pulau Dua
1916	1891	Helfrich	1,407	southeast & northwest	Pulau Dua & Karkua
1979	1978	Amran et al.	168	?	?
1982	—	Capell	91	?	?
1987	1937-1938	Kähler	8,876[8]	south[9]	Kioyo
1987	1983?	Kasim et al.	179	west	Malakoni & Banjar Sari
2011	2010	Yoder	722	west	Meok
2024	2018-2024	Sangian et al.	3,819[10]	west	Meok

4. The curation process of EnoLEX

We report on three main aspects of data curation for EnoLEX: etymologisation (§4.1), concept mapping of the English gloss to a cross-linguistic semantic catalogue (§4.2), and orthography standardisation with IPA transcription (§4.3)

4.1 Etymologisation

All wordlists with Enggano lexemes were initially stored in a Google Sheet, which facilitated work on different devices for efficient collaboration among our team members. For the initial curation process, we sorted around 4000 lexemes attested from 1850 to 1895 etymologically (cf. below for the second phase) to ensure that historically unrelated forms are assigned individual IDs even though they are glossed the same in the original records. For example, forms like *kihu* 'ant' (Cognate ID 131 in Figure 2) (Helfrich, 1888) and *akiaki* 'ant' (ID 132) (Brouwer, n.d.) are assigned different Cognate ID despite having the same gloss.

[3] The R codes to pre-process Kähler (1987) digitisation output are available at https://github.com/engganolang/kahler-1987.
[4] The R code to pre-process Helfrich (1916) is available at https://github.com/engganolang/helfrich-1916-wordlist.
[5] Link to the digitised Holle list: https://engganolang.github.io/digitised-holle-list/
[6] The original list was published in 1891, and an erratum followed in 1893.
[7] We only know about this list thanks to its publication by Stokhof and Almanar as part of the so-called "Holle list" collection (1987). According to the colophon next to the Enggano list, it was supposedly collected on Pulau Dua by a person named Abs vd Noord, of whom we do not know anything. From the minutes of the General and Board Meetings of Batavian Society of Arts and Sciences, we learn that an Enggano wordlist was collected in 1895 by H. P. van der Horst (NBG, 1895, p. 100), while the name Abs vd Noord is not mentioned. As this date coincides with the year of the Holle list, it is unlikely that there exist two different lists.
[8] This figure consists of 3,316 stems/headwords and 5,560 example forms (i.e., derived forms of the stems, phrases, sentences).
[9] Kähler (1940, p. 81) states that there was no dialectal variation when he worked with his consultants, however he mentions that his data is the closest to the former southern dialect from the Kioyo village.
[10] This figure comes from the unique combination of derived/complex forms and their root forms extracted from the Contemporary Enggano FLEx database.

A tibble: 8 × 8

ID	`Cognate ID`	Year	`Given as`	`Common transcription`	`IPA phonemic transcription`	English	Sources
<int>	<int>	<fct>	<chr>	<chr>	<chr>	<chr>	<chr>
966	131	1888	kihu	kixu	kiçu	ant	Helfrich 1888
967	131	1891	kiho	kixo	kiço	ant	Helfrich & Pieters 1891
968	131	1891	kiho	kixo	kiço	ant	Helfrich & Pieters 1891
969	131	1895	èkihŏ	Ekix0	ɛkiçɔ	ant	Stockhof 1987
970	131	1987	ekixo	ekixo	ekiço	ant	Kähler 1987
971	132	<1855	akiaki	akiaki	akiaki	ant	Brouwer <1855
972	132	1855	akie akie	akii akii	aki: aki:	ant	vd Straten & S. 1855

Figure 2. A snippet of the database entry (in RStudio) capturing forms glossed as 'ant' with different Cognate IDs.

The column **Given as** in Figure 2 contains the original orthography/transcription of the words while the standardised, common orthography and its IPA phonemic transcription (§4.3) follow. Thus, we compare *kihu* to *kiho* (Helfrich & Pieters, 1891), *èkihŏ* (Stokhof & Almanar, 1987), and *ekixo* (Kähler, 1987). Meanwhile, *akiaki* is in a separate line having the same Cognate ID with *akie akie* (Van der Straaten & Severijn, 1855).

In a later stage, we compared the records of the 19th century to more recent data of the 20th and 21st centuries. Seemingly similar forms with too divergent semantics are stored in a note. For example, Kähler (1987) has *e?akĩ?akĩ* 'swallow' (name of a bird) (see the **Note for Cognate ID** column in Figure 3), which does not appear in any other wordlist but phonetically corresponds to the words for 'ant'.

A tibble: 8 × 9

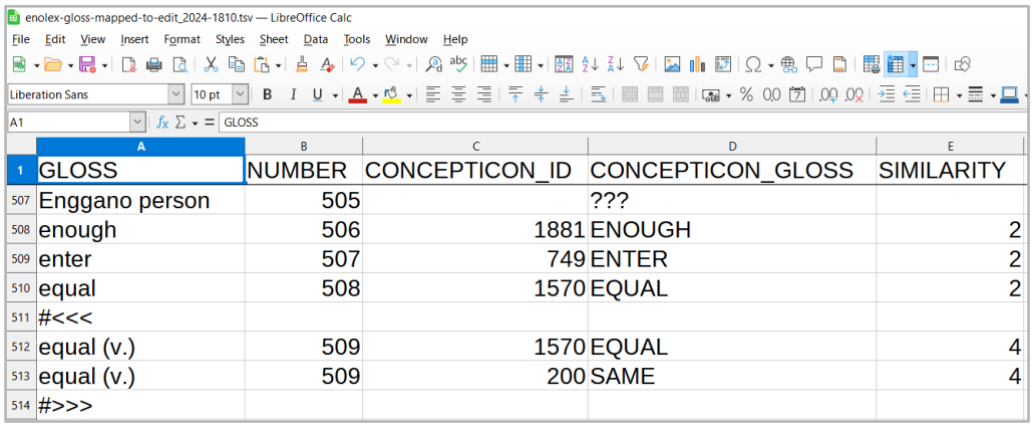
ID	`Cognate ID`	Year	`Given as`	`Common transcription`	`IPA phonemic transcription`	English	Sources	`Note for Cognate ID`
<int>	<int>	<fct>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>
966	131	1888	kihu	kixu	kiçu	ant	Helfrich 1888	NA
967	131	1891	kiho	kixo	kiço	ant	Helfrich & Pieters 1891	NA
968	131	1891	kiho	kixo	kiço	ant	Helfrich & Pieters 1891	NA
969	131	1895	èkihŏ	Ekix0	ɛkiçɔ	ant	Stockhof 1987	NA
970	131	1987	ekixo	ekixo	ekiço	ant	Kähler 1987	NA
971	132	<1855	akiaki	akiaki	akiaki	ant	Brouwer <1855	in Kähler (1987) e?akĩ?akĩ means 'swallow' (bird)
972	132	1855	akie akie	akii akii	aki: aki:	ant	vd Straten & S. 1855	in Kähler (1987) e?akĩ?akĩ means 'swallow' (bird)

Figure 3. A snippet of the database entry (in RStudio) including notes for certain Cognate IDs.

The etymologisation process was also necessary to eliminate wrong glosses. This is the case of *oemau*, which is also glossed as 'ant' by von Rosenberg (1855) but appears in no other list with this meaning. In fact, it is very likely that this form is the same as *oemahaoe* 'don't know' (Helfrich & Pieters, 1891; Stokhof & Almanar, 1987) and *ũmãhãũ* 'don't know, maybe' (Kähler, 1987, p. 285), indicating that von Rosenberg's consultant might have been unaware of the Enggano term for 'ant'. In such cases, *oemau* would not appear under the forms glossed as 'ant' (Indonesian: *semut*) in EnoLEX, but under those glossed as 'don't know' (Indonesian: *entah*). Similar steps had to be taken for Capell's (1982) list, which includes a large number of incorrect glosses, as outlined in Krauße (*forthcoming* b).

4.2 Linking the English Gloss to the Concepticon

Following the recommendation of the Cross-Linguistic Data Format (CLDF) (Forkel & List, 2020; List et al., 2022), we map the English gloss in EnoLEX onto a reference semantic catalogue, namely the *Concepticon* (List et al., 2016, 2023), for cross-linguistic comparability of semantic concepts. To (semi-)automate the linking of our glosses with the Concepticon, we used the *pyconcepticon* Python module (Forkel, 2022). As mentioned by Tjuka (2020), *pyconcepticon* helps with the initial mapping that needs further manual verification. Figure 4 shows the initial output of the concept mapping before manual editing.



	A	B	C	D	E
1	GLOSS	NUMBER	CONCEPTICON_ID	CONCEPTICON_GLOSS	SIMILARITY
507	Enggano person	505		???	
508	enough	506	1881	ENOUGH	2
509	enter	507	749	ENTER	2
510	equal	508	1570	EQUAL	2
511	#<<<				
512	equal (v.)	509	1570	EQUAL	4
513	equal (v.)	509	200	SAME	4
514	#>>>				

Figure 4. A snippet of the output of Concept mapping with *pyconcepticon*.

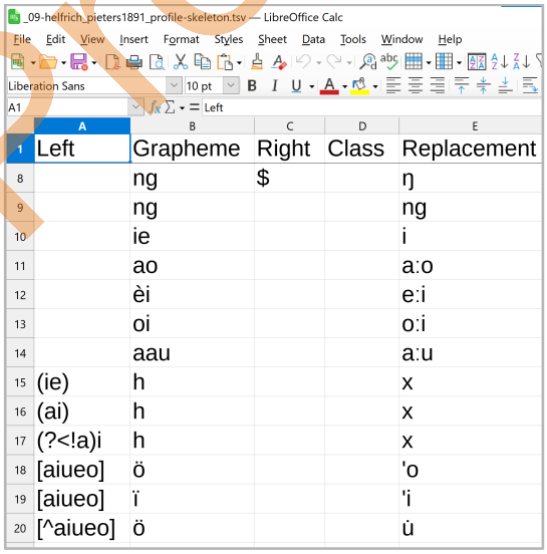
The **GLOSS** column shows the English glosses while **NUMBER** is the ID for each gloss. The markers #<<< and #>>> in the GLOSS column indicate the beginning and the end of a gloss that *pyconcepticon* recognised as a possible mapping onto more than one concept (cf. the **CONCEPTICON_GLOSS** column). For example, the English gloss 'equal (v.)' in Figure 4 (see rows ID 512 and 513) receives two possible concept mappings (i.e., EQUAL and SAME) by the *pyconcepticon* algorithm (these two concepts have a relatively high similarity score between themselves for 'equal (v.)', see the **SIMILARITY** column). In this case, we need to actively resolve this manually by consulting the concept description on the Concepticon web page and then select the most appropriate one.

Meanwhile, the marker **???** (see the row ID 507 for 'Enggano person') indicates that *pyconcepticon* does not know how to map the gloss with the reference concept. In such cases, we must find an alternative mapping on the Concepticon web page or leave the cell empty in the CONCEPTICON_GLOSS column. Once this concept curation is finished[11], the concept mapping table in Figure 4 will be rejoined into the main table so that the English gloss is linked to the relevant Concepticon concept and web page via the reference URL (i.e., https://concepticon.clld.org/parameters/) combined with the CONCEPTICON_ID (see example output with this Concepticon link in Figure 7 for *èkihŏ* 'ant'). The final semantic curation process is matching the Concepticon concept with its relevant semantic field in the Concepticon web page. We did this automatically by pulling the raw Concepticon data from its GitHub repository[12] and then join the semantic field category by matching the CONCEPTICON_GLOSS.

4.3 Orthography standardisation and IPA phonemic transcription

The third step was the standardisation of different orthographies across the wordlists using a set-up common orthography system for each period. For example, *(e)coignò* (Modigliani, 1894) represents the same lemma as *kok²njok* (Helfrich & Pieters, 1891) and *kŏnjŏk* (Stokhof & Almanar, 1987), all glossed as 'hungry'. The standard orthography renders these forms as *koiñO*, *ko'ñok*, and *kOñOk*, respectively (see further below for their representation in IPA). Such standardisation allows for easier comparison across wordlists of different periods. The corresponding expression *ekō iˀiōˀou* from Kähler (1987, p. 55) would later be linked to the same ID, which shows that the actual meaning is 'I am hungry' (lit. 'the hunger is on me').[13]

The orthography standardisation involves creating an orthography profile following the common transcription[14] set up by the team in our project (cf. Hemmings et al., 2023 for the challenges in Enggano orthography development). Using the *qlcData* R package (Moran & Cysouw, 2018), we created an orthography profile skeleton (Figure 5) for each period/author. We then manually edited the values of the **Grapheme** and/or the **Replacement** columns. For instance, the *qlcData* package by default splits words into single characters/graphemes but our team would manually correct the sequence of certain vowels as being diphthong. In that case, we added the vowel sequence in the Grapheme column (e.g., the string *aau* in row 14 in Figure 5) with its corresponding Replacement (i.e., *aːu*) to match the common transcription.


_09-helfrich_pieters1891_profile-skeleton.tsv — LibreOffice Calc

File Edit View Insert Format Styles Sheet Data Tools Window Help

Liberation Sans 10 pt

A1 = Left

	A	B	C	D	E
1	Left	Grapheme	Right	Class	Replacement
8		ng	$		ŋ
9		ng			ng
10		ie			i
11		ao			aːo
12		èi			eːi
13		oi			oːi
14		aau			aːu
15	(ie)	h			x
16	(ai)	h			x
17	(?<!a)i	h			x
18	[aiueo]	ö			ˈo
19	[aiueo]	ï			ˈi
20	[^aiueo]	ö			u̇

Figure 5. A snippet of the orthography profile skeleton for Helfrich and Pieters (1891).

[11] https://bit.ly/enolex-concepticon
[12] https://github.com/concepticon/concepticon-data
[13] The contemporary Enggano form is *kõ ĩẽ'*.
[14] The long, tidy format can be accessed at https://bit.ly/eno-long-format-ortho, which is derived from the wide version: https://bit.ly/eno-transcription-wide using the R codes available at https://bit.ly/eno-common-ortho-processing.

Another functionality of *qlcData* is that it can handle contexts surrounding (i.e., to the left and right of) the Grapheme to condition the replacement (i.e., contextual replacement using *regular expressions* [regex]). For example, the grapheme *ng* (rows 8-9 in Figure 5) will be replaced differently depending on whether it appears at the end of a string (row 8 with the regex anchor $) or elsewhere (row 9) (see also rows 17-20 for other examples of contextual replacement). Any changes made to the profiles are tracked on the GitHub repository for EnoLEX[15]. Once this was done, we used this edited orthography profile to transliterate the original transcription into the common transcription (a snippet of the results can be seen in the **Common transcription** column in Figure 2 and Figure 3).

Since our team has also mapped the common transcription with the IPA phoneme, we were able to semi-automatically map the orthography with the corresponding phonemes to retrieve an accurate IPA transcription for every word in each period (see the **IPA phonemic transcription** column in Figure 2 and Figure 3). Our example 'hungry' from above is thus represented as /koiŋɔ/, /koʔŋok/, and /kɔŋɔk/, respectively, which allows for an even more fine-grained comparison across wordlists. Some manual editing was still required for unrecognised characters to be mapped onto the IPA phonemes. The code for this step is made available at https://bit.ly/enolex-ortho. The output of the orthography profiling and IPA transcription is available at https://bit.ly/enolex-ortho-ipa.

5. Prototype development

The development of EnoLEX involved creating a mid-fidelity prototype (§5.1), implementing a system using *Golang* for the back end and *React* for the front end (§5.2), and utilising cloud computing infrastructure (§5.3) to ensure accessibility and scalability. This section outlines the steps taken and provides an overview of the figures representing different aspects of the prototype.

5.1. Mid-Fidelity Prototype

The mid-fidelity prototype tests and refines the core features of the EnoLEX database, focusing on (i) detailed user interfaces and interaction (for a more accurate representation of the user experience), and (ii) data management (Pramartha et al., 2023). The prototype serves to bridge the gap between initial wireframes and the final high-fidelity product (Engelberg & Seffah, 2002). Using Figma for the prototype development (available at https://s.id/26Ed9) allows for collaborative design and easy iteration based on feedback.

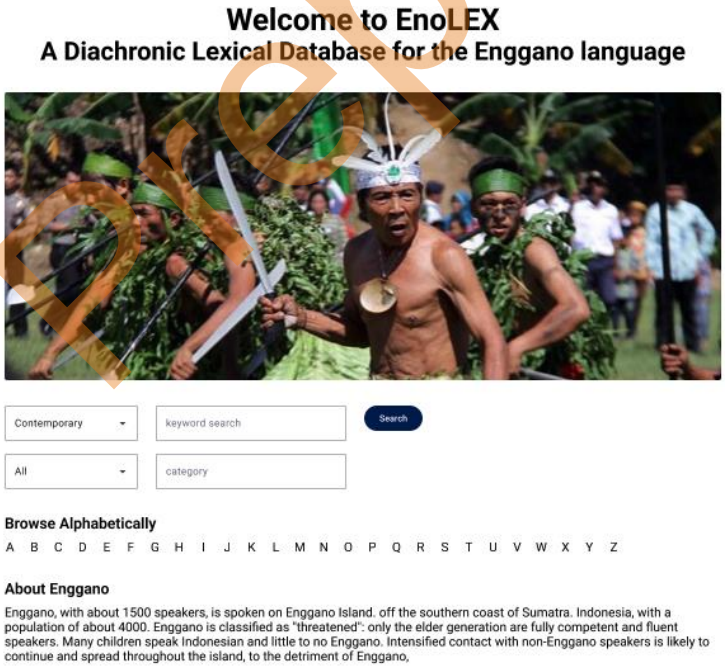


Figure 6. The landing page for the prototype.[16]

The landing page serves as the entry point for users accessing the database. It provides features and navigation options. The page is designed to be user-friendly and informative. Key elements of the landing page in the prototype include:

[15] https://github.com/engganolang/enolex
[16] Cover image source: https://bit.ly/eno-war-dance

- **Navigation Menu**: Links to different sections of the database, such as browsing and searching.
- **Keyword search**: A search bar that allows users to quickly find specific lexical entries by typing in queries.
- **Alphabetical browsing**: Users can select a letter to view all entries starting with that letter (cf. Figure 8).

The search function represents an essential component for users who wish to find particular words or phrases quickly. The design includes:

- **Search Input Field**: Users can type in their queries (see the "keyword search" bar in Figure 6).
- **Filter Options**: Additional filters to narrow down search results by date, sources, or other criteria (the "Contemporary" and "all" dropdown menus to the left of the "category" and "keyword search" search bars in Figure 6).
- **Results List**: Displaying the matched entries with information, such as the word, its gloss, and source details.
- **Pagination**: Enabling navigation through multiple pages of search results.

Figure 7 below shows the interface of results after a user query to retrieve relevant lexical entries via the keyword search bar. The results interface shows among others the original, common, and IPA transcriptions (§4.3) of the queried word and the link to the relevant webpage for the semantic concept in the Concepticon (§4.2).

Figure 7. Query results.

In addition to manually typing in the queries, users can also navigate through the entries alphabetically, which helps with finding specific words efficiently. Figure 8 below illustrates the output of the alphabetical browsing interface for lexical entries starting with the letter S. The forms in the **Word** column are in the common transcription (cf. §4.3). For example, *skOci* /skɔt͡ʃi/ 'rowboat' is standardised from the original *scotjie* and *sa:uda* /sauda/ 'snake' is standardised from *sauda*. The layout is designed to be intuitive, displaying the lexical items in a list format with relevant details such as their glosses, semantic field (§4.2), year of attestation, and source information. Key features include:

- **Entry List**: A comprehensive list of lexical entries, including the word, its gloss, and additional metadata.
- **Details Panel**: Clicking on an entry reveals more detailed information about the word, including its historical usage and related entries. The historical information in the EnoLEX database provides Proto-Malayo-Polynesian and Proto-Austronesian reconstructions that are directly linked to the Austronesian Comparative Dictionary.

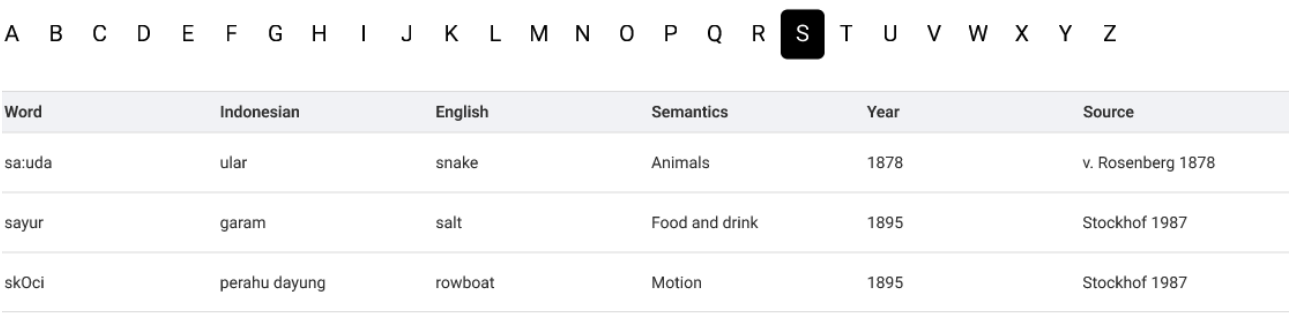


Figure 8. Interface of alphabetical browsing (S)

5.2. Implementation Using *Golang* and *React*

This step is still work in progress. Once the mid-fidelity prototype is finalised, we plan to implement the EnoLEX database using *Golang* for the back end and *React* for the front end. *Golang* was chosen for its efficiency, scalability, and

performance (Nabiil et al., 2023), which are essential for handling the extensive and complex data involved in a diachronic lexical database. *React*, on the other hand, offers a responsive and dynamic user interface (Kolomoyets & Kynash, 2023), making it easier for users to interact with the database. This combination allows for a robust, fast, and user-friendly application.

5.3. Cloud Computing Infrastructure

To enhance accessibility and ensure that the database can be used by researchers and linguists worldwide, EnoLEX is hosted on a cloud computing infrastructure (to be served on the web server of CIRHSS[17]). This setup allows users to access the database through any internet-connected browser, ensuring high availability and reliability. Utilising cloud services provides the flexibility to scale resources based on demand, ensuring optimal performance and cost-effectiveness. The cloud infrastructure also offers enhanced security measures to protect sensitive data contained within the database. In addition to presenting EnoLEX as a web page, we will maintain the raw data publicly on GitHub and archive them to Zenodo, which is standard practice adopted by the Lexibank project.

Acknowledgements

The study and databases reported in this paper are funded by the Arts and Humanities Research Council (AHRC), UK for the projects *Enggano in the Austronesian family: Historical and typological perspectives* (AH/S011064/1) and *Lexical resources for Enggano, a threatened language of Indonesia* (AH/W007290/1). We would like to thank our research assistants who helped with the digitisation of the Enggano-German dictionary: Ni Putu Wulan Lestari, Yul Fulgensia Rusman Pita, Fitriani Putri Koemba, Putu Dea Indah Kartini, Putu Wahyu Widiatmika, Ida Bagus Made Ari Segara, and I Gede Semara Dharma Putra, with Ida Bagus Gede Sarasvananda as the developer of the online entry system. Barnaby Burleigh and Rena Dusee provided useful feedback to the German-English and Dutch-English translations of the automated translations of the dictionaries by Kähler (1987) and Helfrich (1916). Throughout the development of EnoLEX, we have received very helpful comments from Erik Zobel, Charlotte Hemmings, Bernd Nothofer, Mary Dalrymple, I Wayan Arka, Engga Zakaria Sangian, Dendi Wijaya, and Sarah Ogilvie.

References

Amran, F., Madjid, J. E., Karim, M., & Sulistinah, S. (1979). *Etnografi penduduk Pulau Enggano: Sebuah laporan sementara*. Fakultas Sastra, Universitas Indonesia. https://lib.ui.ac.id/m/detail.jsp?id=20229144&lokasi=lokal

Arka, I. W., Arono, Wijaya, D., & Zakaria, E. (2022). *Critical ecological factors and Enggano vitality* [Presentation]. 14th International Austronesian and Papuan Languages and Linguistics (APLL) Conference, Berlin. https://enggano.ling-phil.ox.ac.uk/static/papers/APLL-14-Arka%20et%20al-1.pdf

Billings, B., & McDonnell, B. (2024). Sumatran. *Oceanic Linguistics*, *63*(1), 112–174. https://doi.org/10.1353/ol.2024.a928205

Blench, R. (2014). The Enggano: Archaic foragers and their interactions with the Austronesian world. *Unpublished Draft*. https://rogerblench.info/Language/Austronesian/Enggano/Enggano%20and%20its%20history.pdf

Blust, R., Trussel, S., & Smith, A. D. (2023). *CLDF dataset derived from Blust's 'Austronesian Comparative Dictionary'* (v1.2) [dataset]. Zenodo. https://doi.org/10.5281/zenodo.7741197

Boewang, D. (1854). Verslag omtrent het Eiland Enggano. *Tijdschrift Voor Indische Taal-, Land-En Volkenkunde*, *2*, 379–393.

Brouwer, D. J. (n.d.). *Woordenlijst van het weinige door de Inlanders op Engano te Barhao ons medegedeelde der Enganeesche taal* [Unpublished manuscript].

Butters, M. (2021). *Negation in Four Languages of Indonesia* [PhD Thesis]. University of Colorado at Boulder.

Capell, A. (1982). Bezirkssprachen im Gebiet im Gebiet des UAN. In R. Carle (Ed.), *GAVA': Studien zu austronesischen Sprachen und Kulturen Hans Kähler gewidmet* (pp. 1–15). Reimer.

Dyen, I. (1965). *A lexicostatistical classification of the Austronesian languages*. Waverly Press, Inc.

Edwards, O. (2015). The Position of Enggano within Austronesian. *Oceanic Linguistics*, *54*(1), 54–109. https://doi.org/10.1353/ol.2015.0001

Engelberg, D., & Seffah, A. (2002). A framework for rapid mid-fidelity prototyping of web sites. *Usability: Gaining a Competitive Edge*, 203–215.

Forkel, R. (2022). *pyconcepticon: Programmatic curation of concepticon-data* (3.0.0) [Python; OS Independent]. https://github.com/concepticon/pyconcepticon

Forkel, R., & List, J.-M. (2020). CLDFBench: Give Your Cross-Linguistic Data a Lift. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 6995–7002. https://aclanthology.org/2020.lrec-1.864

Greenhill, S. J. (2015). TransNewGuinea.org: An Online Database of New Guinea Languages. *PLOS ONE*, *10*(10). https://doi.org/10.1371/journal.pone.0141563

Greenhill, S. J., Blust, R., & Gray, R. D. (2008). The Austronesian Basic Vocabulary Database: From Bioinformatics to Lexomics. *Evolutionary Bioinformatics*, *4*, EBO.S893. https://doi.org/10.4137/EBO.S893

[17] https://www.cirhss.org/

Helfrich, O. L. (1888). De eilandgroep Engano. *Tijdschrift van Het Koninklijk Nederlandsch Aardrijkskundig Genootschap*, *5*, 272–314.
Helfrich, O. L. (1916). Nadere bijdrage tot de kennis van het engganeesch. *Bijdragen Tot de Taal-, Land- En Volkenkund*, *71*(1), 472–555. https://doi.org/10.1163/22134379-90001719
Helfrich, O. L., & Pieters, J. A. J. C. (1891). Proeve van een Maleisch-Nederlansch-Engganeesch woordenlijst. *Tijdschrift Voor Indische Taal-, Land- En Volkenkunde*, *34*, *35*, 539–623, 228–233.
Hemmings, C. (2024). *Verbal morphology in Enggano and Nias* [Presentation]. 16th Austronesian and Papuan Languages and Linguistics (APLL) Conference, Vrije Universiteit Amsterdam. https://enggano.ling-phil.ox.ac.uk/static/papers/APLL16_Hemmings.pdf
Hemmings, C. (*forthcoming*). Nominal Morphosyntax in Contemporary Enggano. In I. W. Arka, M. Dalrymple, & C. Hemmings (Eds.), *Enggano: Historical and Contemporary Perspectives*. ANU Press.
Hemmings, C., Arka, I. W., Sangian, E. Z., Wijaya, D., & Dalrymple, M. (2023). Challenges in Enggano Orthography Development. *Language Documentation and Description*, *23*(1), Article 1. https://doi.org/10.25894/ldd.329
Hemmings, C., & Tan, J. (*forthcoming*). Contemporary Enggano Phonology. In I. W. Arka, M. Dalrymple, & C. Hemmings (Eds.), *Enggano: Historical and Contemporary Perspectives*. ANU Press.
Kähler, H. (1940). Grammatischer Abriß des Enggano. *Zeitschrift Für Eingeborenen-Sprachen*, *XXX*, 81–117, 182–210, 296–310.
Kähler, H. (1987). *Enggano-Deutsches Wörterbuch*. Dietrich Reimer Verlag.
Kaiping, G. A., Edwards, O., & Klamer, M. (Eds.). (2019). *LexiRumah 3.0.0*. Leiden University Centre for Linguistics. https://lexirumah.model-ling.eu/
Kaiping, G. A., & Klamer, M. (2018). LexiRumah: An online lexical database of the Lesser Sunda Islands. *PLOS ONE*, *13*(10). https://doi.org/10.1371/journal.pone.0205250
Kasim, Y., Maksan, M., Arifin, S., & Zailoet, Z. (1987). *Pemetaan bahasa daerah di Sumatra Barat dan Bengkulu*. Pusat Pembinaan dan Pengembangan Bahasa. https://repositori.kemdikbud.go.id/1684/
Kolomoyets, M., & Kynash, Y. (2023). Front-End web development project architecture design. *2023 IEEE 18th International Conference on Computer Science and Information Technologies (CSIT)*, 1–5. https://doi.org/10.1109/CSIT61576.2023.10324238
Krauße, D. (2024). The Enggano language: Nothofer's contribution to solving a linguistic puzzle. In C. Bracks, A. Graf, & P. Keilbart (Eds.), *Towards the next 40 years of Southeast Asian Studies in Frankfurt: Essays in honour of Bernd Nothofer*. Iudicium.
Krauße, D. (*forthcoming* a). Early written records of Enggano. In I. W. Arka, M. Dalrymple, & C. Hemmings (Eds.), *Enggano: Historical and Contemporary Perspectives*. ANU Press.
Krauße, D. (*forthcoming* b). From past to present: Enggano as an Austronesian language. In I. W. Arka, M. Dalrymple, & C. Hemmings (Eds.), *Enggano: Historical and Contemporary Perspectives*. ANU Press.
List, J.-M., Cysouw, M., & Forkel, R. (2016). Concepticon: A Resource for the Linking of Concept Lists. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 2393–2400). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2016/summaries/127.html
List, J.-M., Forkel, R., Greenhill, S. J., Rzymski, C., Englisch, J., & Gray, R. D. (2022). Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data*, *9*(1), Article 1. https://doi.org/10.1038/s41597-022-01432-0
List, J.-M., Tjuka, A., Van Zantwijk, M., Blum, F., Ugarte, C. B., Rzymski, C., Greenhill, S., & Forkel, R. (2023). *CLLD Concepticon 3.1.0* (v3.1.0) [dataset]. Zenodo. https://doi.org/10.5281/ZENODO.7777629
Mahdi, W. (1988). *Morphophonologische Besonderheiten und historische Phonologie des Malagasy*. D. Reimer.
Modigliani, E. (1894). *L'isola delle donne*. Ulrico Hoepli. https://www.google.co.uk/books/edition/L_isola_delle_donne/gksCAAAAMAAJ?hl=en&gbpv=0
Moran, S., & Cysouw, M. (2018). *The Unicode cookbook for linguists: Managing writing systems using orthography profiles*. Language Science Press. https://doi.org/10.5281/zenodo.1296780
Nabiil, A., Makmur, B. H., Wijaya, R. W., Santoso Gunawan, A. A., & Edbert, I. S. (2023). Performance Analysis on Web Development Programming Language (Javascript, Golang, PHP). *2023 International Conference on Information Technology and Computing (ICITCOM)*, 6–11. https://doi.org/10.1109/ICITCOM60176.2023.10442358
NBG. (1895). Bestuursvergadering van Dinsdag 1 October 1895. *Notulen van de Algemeene En Directie-Vergaderingen van Het Bataviaasch Genootschap van Kunsten En Wetenschappen*, *33*, 94–108.
Nothofer, B. (1986). The Barrier Island languages in the Austronesian language family. In P. Geraghty, L. Carrington, & S. A. Wurm (Eds.), *FOCAL II: Papers from the Fourth International Conference on Austronesian Linguistics* (Vol. 94, pp. 89–107). Research School of Pacific and Asian Studies, Australian National University.
Pramartha, C., Mahendra, I. M. Y., Rajeg, G. P. W., & Arka, I. W. (2023). The development of semantic dictionary prototype for the Balinese language. *International Journal of Cyber and IT Service Management*, *3*(2), 96–

106.
Rajeg, G. P. W. (2023a). *CLDF dataset of the Enggano word list from 1895 in Stokhof and Almanar's (1987) Holle List* (1.0.0) [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.8038975
Rajeg, G. P. W. (2023b). *Digitised, searchable Holle List in Stokhof (1980)*. https://doi.org/10.5281/zenodo.7972273
Rajeg, G. P. W., Hemmings, C., & Arka, I. W. (2024, June 21). *Enggano middle voice: Evidence of Enggano as an Austronesian language* [Presentation]. 16th International Conference on Austronesian Linguistics (16-ICAL), De La Salle University, Manila. University of Oxford. https://doi.org/10.25446/oxford.26073907
Rajeg, G. P. W., Paramartha, C. R. A., Arka, I. W., & Dalrymple, M. (2023, June 22). *Enggano-German dictionary turns digital: Challenges and opportunities in retro-digitising historical materials of an endangered language* [Presentation]. Konferensi Linguistik Tahunan Atma Jaya Kedua puluh satu (KOLITA 21), Universitas Katolik Indonesia Atma Jaya, Jakarta. University of Oxford. https://doi.org/10.25446/oxford.25217018
Sangian, E. Z., Wijaya, D., Hemmings, C., Arka, I. W., & Dalrymple, M. (2024). *Documentation of Contemporary Enggano* [Unpublished Enggano corpus].
Smith, A. D. (2017). The Western Malayo-Polynesian Problem. *Oceanic Linguistics*, *56*(2), 435–490. https://doi.org/10.1353/ol.2017.0021
Stokhof, W. A. L. (Ed.). (1980). *Holle lists, vocabularies in languages of Indonesia, Vol. 1: Introductory Volume: Vol. Materials in Languages of Indonesia*. Dept. of Linguistics, Research School of Pacific Studies, The Australian National University. https://core.ac.uk/reader/159464813
Stokhof, W. A. L., & Almanar, A. E. (1987). *Holle lists, vocabularies in languages of Indonesia, Vol. 10/3: Islands off the west coast of Sumatra: Vol. Materials in Languages of Indonesia*. Dept. of Linguistics, Research School of Pacific Studies, The Australian National University. http://hdl.handle.net/1885/144589
Tadmor, U., & Gil, D. (Eds.). (2015). *Jakarta Lexical Database*. Jakarta Field Station, Department of Linguistics, Max Planck Institute for Evolutionary Anthropology. https://hdl.handle.net/1839/00-0000-0000-0022-7166-5
Tjuka, A. (2020, January 29). Adding concept lists to Concepticon: A guide for beginners. *Computer-Assisted Language Comparison in Practice*. https://doi.org/10.58079/m6kj
Van der Straaten, J., & Severijn, P. (1855). Verslag van een in 1854 bewerkstelligd onderzoek op het eiland Engano. *Tijdschrift Voor Indische Taal-, Land-En Volkenkunde*, *3*, 338–369.
von Rosenberg, C. B. H. (1855). Beschrijving van Engano en van deszelfs bewoners. *Tijdschrift Voor Indische Taal-, Land-En Volkenkunde*, *3*, 370–386.
von Rosenberg, C. B. H. (1878). *Der Malayische Archipel: Land und Leute in Schilderungen, gesammelt während eines driessig-jährigen Aufenhaltes in den Kolonien*. Gustav Weigel.
Yoder, B. E. (2011). *Phonological and phonetic aspects of Enggano vowels* [Master's Thesis, University of North Dakota]. https://commons.und.edu/theses/4457