



Penyusunan Korpus Bahasa Daerah

Gede Primahadi Wijaya Rajeg

University of Oxford, UK; Centre for Interdisciplinary Research on the Humanities and Social Sciences (CIRHSS) & *CompLexico* research group, Universitas Udayana

<https://orcid.org/0000-0002-2047-8621>

Konsinyasi Penyiapan Data Korpus Bahasa Daerah dan Pemetaan Bahasa (24 Oktober 2024)

Badan Pengembangan dan Pembinaan Bahasa, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi melalui Pusat Pengembangan dan Pelindungan Bahasa dan Sastra

Butir pembahasan

1. Konsep *korpus*
2. Target Bahasa
3. Sumber (data & daya)
4. Aksesibilitas + Infrastruktur
5. Contoh kasus:
 - a. Bahasa Enggano
 - b. Bahasa Bali
 - c. Bahasa-bahasa di Australia (LDaCA)

Butir pembahasan

1. Konsep *korpus* 
2. Target Bahasa
3. Sumber (data & daya)
4. Aksesibilitas + Infrastruktur
5. Contoh kasus:
 - a. Bahasa Enggano
 - b. Bahasa Bali
 - c. Bahasa-bahasa di Australia (LDaCA)

Konsep *korpus*

Bagian ini berasal dari sebagian salindia untuk MK Linguistik Korpus di Prodi S3 Linguistik Univ. Udayana.



sekumpulan sampel teks (i) **digital** yang (ii) bersifat **otentik**, (iii) **representatif** dan **berimbang**, serta (iv) **berukuran besar**
(Stefanowitsch 2020: 22-28; Gries 2017: 7)

Stefanowitsch, A. (2020). *Corpus linguistics: A guide to the methodology*. Language Science Press. <http://langsci-press.org/catalog/book/148>
Gries, Stefan Th. 2017. *Quantitative corpus linguistics with R: A practical introduction*. Second edition. New York: Routledge.

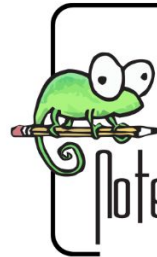
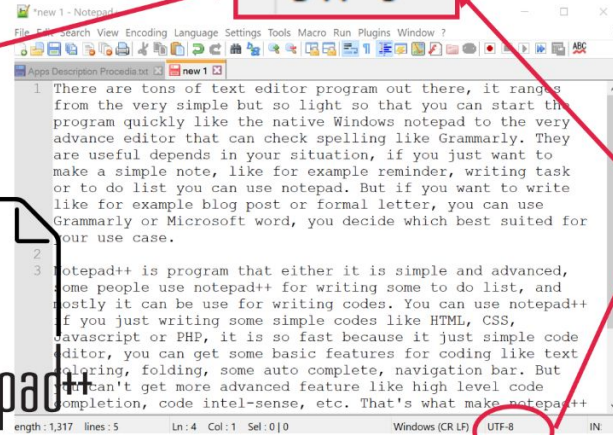
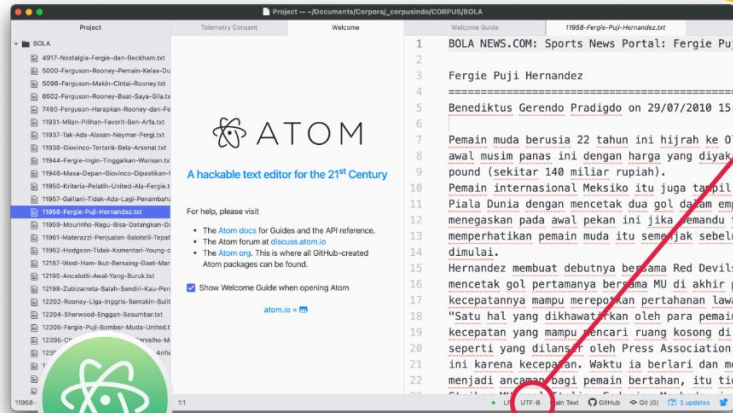
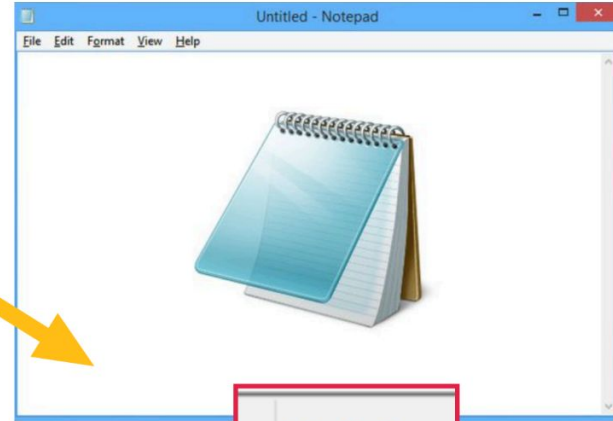


sekumpulan sampel teks (i) **digital** yang (ii) bersifat otentik, (iii) representatif dan berimbang, serta (iv) berukuran besar
(Stefanowitsch 2020: 22-28; Gries 2017: 7)

Stefanowitsch, A. (2020). *Corpus linguistics: A guide to the methodology*. Language Science Press. <http://langsci-press.org/catalog/book/148>
Gries, Stefan Th. 2017. *Quantitative corpus linguistics with R: A practical introduction*. Second edition. New York: Routledge.

Teks biasa (*plain text*)

Unicode UTF-8 encoding





sekumpulan sampel teks (i) digital yang (ii) bersifat **otentik**, (iii) representatif dan berimbang, serta (iv) berukuran besar
(Stefanowitsch 2020: 22-28; Gries 2017: 7)

Stefanowitsch, A. (2020). *Corpus linguistics: A guide to the methodology*. Language Science Press. <http://langsci-press.org/catalog/book/148>
Gries, Stefan Th. 2017. *Quantitative corpus linguistics with R: A practical introduction*. Second edition. New York: Routledge.

Otentisitas korpus:
“real life’ language use”
(McEnery & Wilson 2001: 1)

McEnery, Tony & Andrew Wilson. 2001. *Corpus linguistics: An introduction*. 2. ed., repr. Edinburgh: Edinburgh Univ. Press.

Otentisitas korpus:
**“language that is not, as it were, performed for the linguist based
on what speakers believe constitutes “good” or “proper” language”**
(Stefanowitsch 2020: 23)

Stefanowitsch, A. (2020). *Corpus linguistics: A guide to the methodology*. Language Science Press. <http://langsci-press.org/catalog/book/148>



sekumpulan sampel teks (i) digital yang (ii) bersifat otentik, (iii) representatif dan berimbang, serta (iv) berukuran besar
(Stefanowitsch 2020: 22-28; Gries 2017: 7)

Stefanowitsch, A. (2020). *Corpus linguistics: A guide to the methodology*. Language Science Press. <http://langsci-press.org/catalog/book/148>

Gries, Stefan Th. 2017. *Quantitative corpus linguistics with R: A practical introduction*. Second edition. New York: Routledge.

keterwakilan sampel:
Sampel yang representatif adalah **subbagian (*subset*)** suatu populasi yang identik dengan populasi secara keseluruhan.

(Stefanowitsch 2020: 28)

Stefanowitsch, A. (2020). *Corpus linguistics: A guide to the methodology*. Language Science Press. <http://langsci-press.org/catalog/book/148>

keterwakilan sampel:

Sampel yang representatif adalah subbagian (*subset*) suatu populasi yang identik dengan populasi secara keseluruhan. **Hal yang identik di sini adalah distribusi suatu fenomena yang diamati (pada sampel) dan berusaha digeneralisasi terhadap populasi (Stefanowitsch 2020: 28)**

Stefanowitsch, A. (2020). *Corpus linguistics: A guide to the methodology*. Language Science Press. <http://langsci-press.org/catalog/book/148>

keterwakilan korpus:
distribusi fenomena kebahasaan (mis. kata, struktur gramatikal, dll.) di dalam korpus semestinya proporsional (i.e., identik) dengan distribusi fenomena tersebut di dalam (ragam) bahasa tersebut (i.e., populasi) secara keseluruhan (Stefanowitsch 2020: 28)

Stefanowitsch, A. (2020). *Corpus linguistics: A guide to the methodology*. Language Science Press. <http://langsci-press.org/catalog/book/148>

keterwakilan korpus

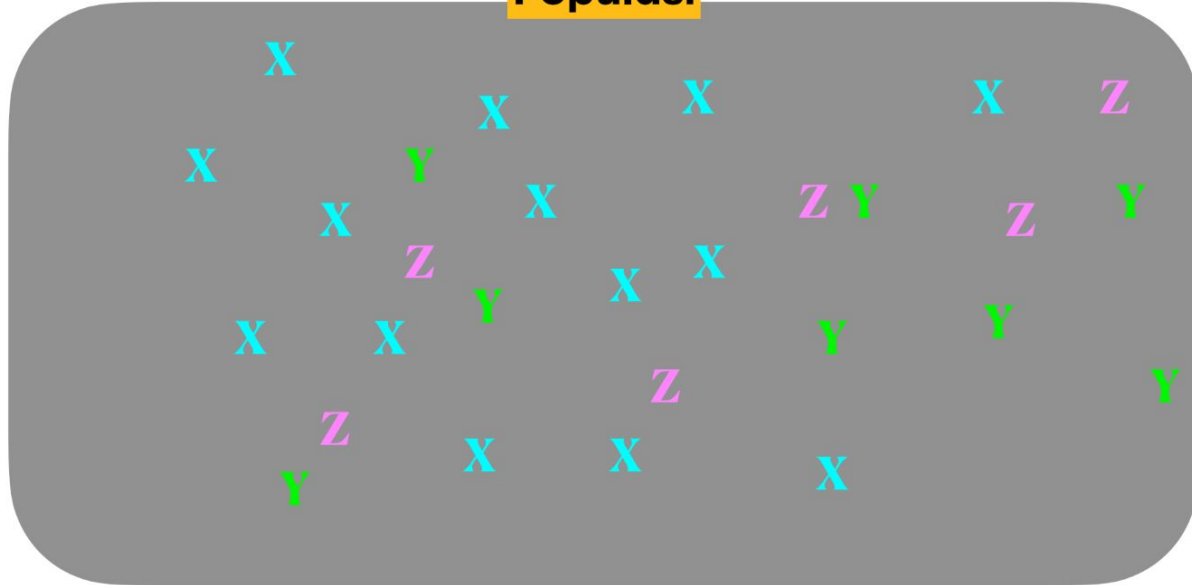
Total = 28 (100%)

X = 14 (50%)

Y = 8 (28.57%)

Z = 6 (21.43%)

Populasi



keterwakilan korpus

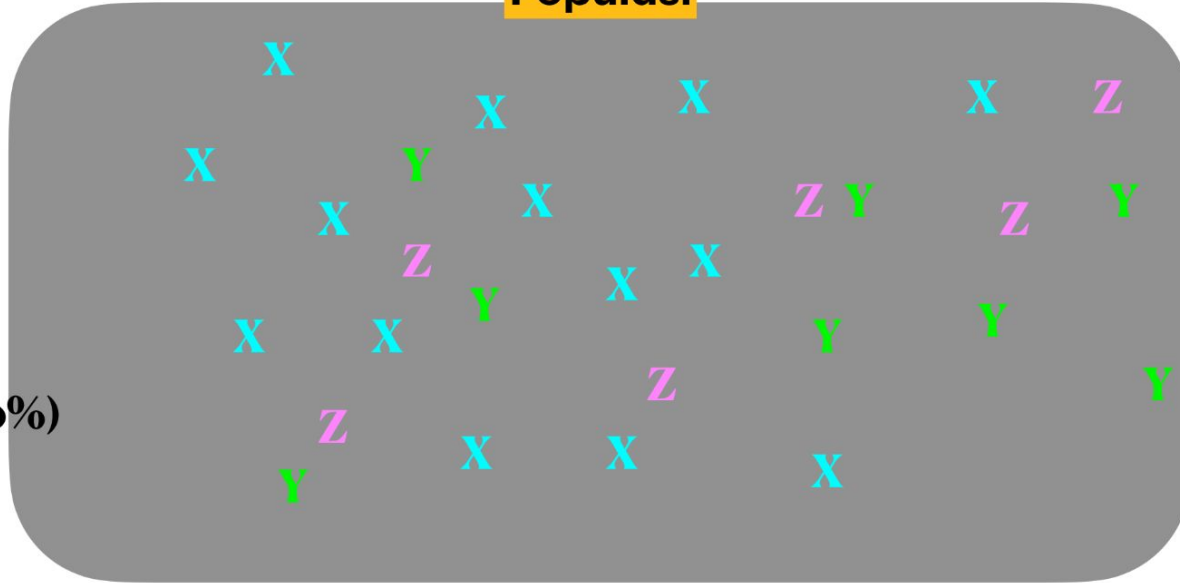
Total = 28 (100%)

X = 14 (50%)

Y = 8 (28.57%)

Z = 6 (21.43%)

Populasi



Sampel = 18 (100%)

X = (50%)

Y = (28.57%)

Z = (21.43%)

keterwakilan korpus

Total = 28 (100%)

X = 14 (50%)

Y = 8 (28.57%)

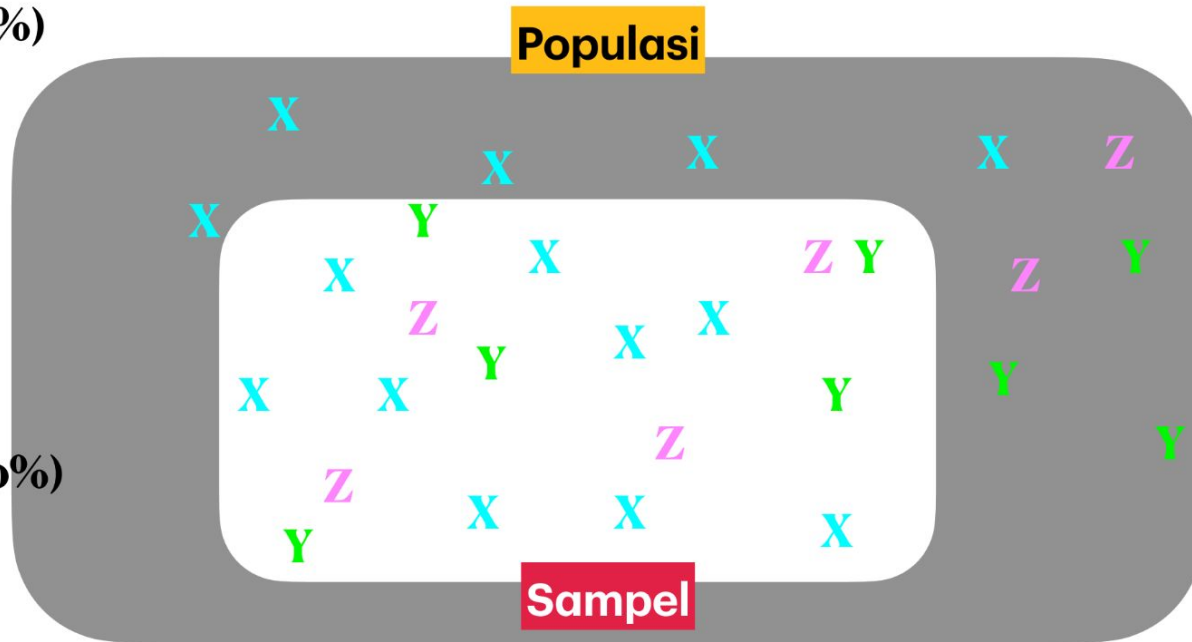
Z = 6 (21.43%)

Sampel = 18 (100%)

X = 9 (50%)

Y = 5 (28.57%)

Z = 4 (21.43%)



keterwakilan korpus

Total = 28 (100%)

X = 14 (50%)

Y = 8 (28.57%)

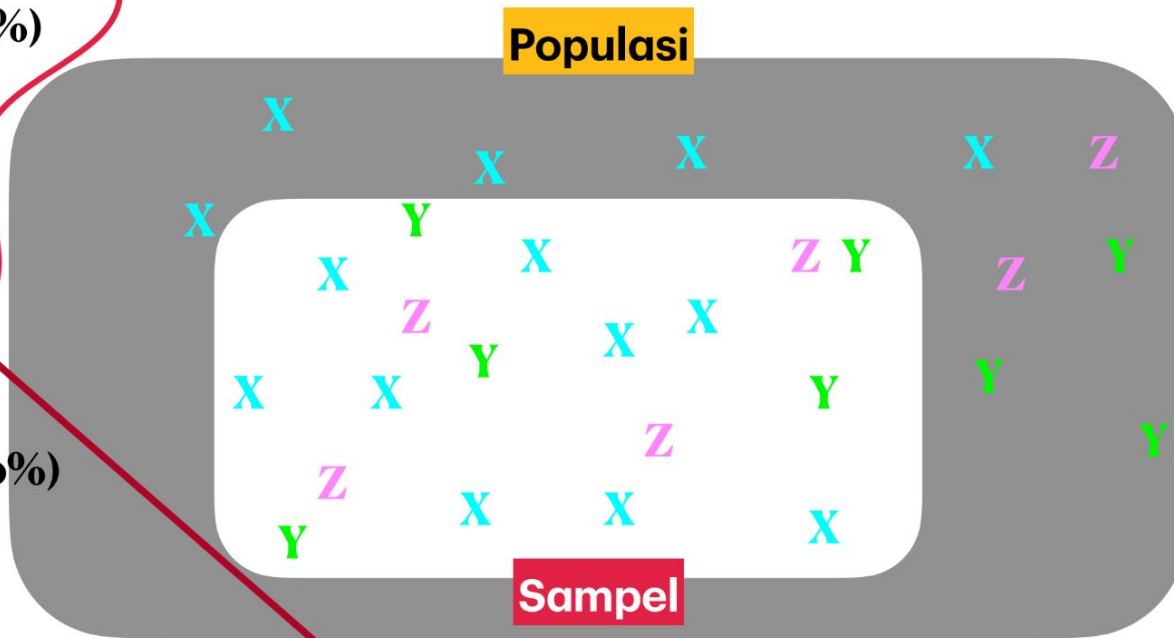
Z = 6 (21.43%)

Sampel = 18 (100%)

X = 9 (50%)

Y = 5 (28.57%)

Z = 4 (21.43%)



Isu: Distribusi tiap-tiap fitur di dalam populasi tidak diketahui pasti

Solusi perancang korpus

Menuju korpus yang representatif

- Beragam **manifestasi pemakaian bahasa** untuk suatu bahasa (populasi) diikutkan di dalam korpus
 - **Ragam bahasa** (perbedaan dalam kriteria demografis, budaya, bahasa)
 - **Genre** (mis. fiksi, sastra, surat kabar, akademik, blog, komentar sosmed)
 - **Register** (fitur bahasa terkait dengan fungsi sosial tertentu)
 - **Formalitas**
 - **Medium** (tulisan, lisan)
 - **Topik** (isi atau kandungan ranah dari teks)

Solusi perancang korpus

Menuju korpus yang representatif

- Beragam **manifestasi pemakaian bahasa** untuk suatu bahasa (populasi) diikuti di dalam korpus
- Korpus Representatif dan Berimbang:
 - mencakup semua kategori teks pemakaian bahasa sebelumnya
 - secara akurat mencerminkan secara kualitatif dan kuantitatif semua ragam pemakaian bahasa pada guyub tutur yang bahasanya (populasi) ingin ditangkap di dalam korpus
 - Idealisme yang tidak mungkin untuk dicapai

Table 1.1 *The LOB Corpus Sampling Frame (after Hofland and Johansson 1982: 2)*

Category mnemonic	Description	Number of text samples in this category
A	Press: reportage	44
B	Press: editorial	27
C	Press: reviews	17
D	Religion	17
E	Skills, trades and hobbies	38
F	Popular lore	44
G	Belles lettres, biography, essays	77
H	Miscellaneous (government documents, foundation reports, industry reports, college, catalogue, industry house organ)	30
J	Learned and scientific writings	80
K	General fiction	29
L	Mystery and detective fiction	24
M	Science fiction	6
N	Adventure and western fiction	29
P	Romance and love story	29
R	Humour	9
Total		500

McEnery, Tony & Andrew Hardie. 2012. *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

Lancaster-Oslo/Bergen (LOB) Corpus

Sampel data untuk masing-masing kategori teks hampir sama (sekitar 2000 kata)
(McEnery & Hardie 2012: 10)



sekumpulan sampel teks (i) digital yang (ii) bersifat otentik, (iii) representatif dan berimbang, serta (iv) **berukuran besar**
(Stefanowitsch 2020: 22-28; Gries 2017: 7)

Stefanowitsch, A. (2020). *Corpus linguistics: A guide to the methodology*. Language Science Press. <http://langsci-press.org/catalog/book/148>
Gries, Stefan Th. 2017. *Quantitative corpus linguistics with R: A practical introduction*. Second edition. New York: Routledge.

Mega Korpus

<https://www.english-corpora.org>

Corpus (online access)	Download	# words	Dialect	Time period	Genre(s)
iWeb: The Intelligent Web-based Corpus		14 billion	6 countries	2017	Web
News on the Web (NOW)		13.3 billion+	20 countries	2010-yesterday	Web: News
Global Web-Based English (GloWbE)		1.9 billion	20 countries	2012-13	Web (incl blogs)
Wikipedia Corpus		1.9 billion	(Various)	2014	Wikipedia
Coronavirus Corpus		1.15 billion+	20 countries	Jan 2020-yesterday	Web: News
Corpus of Contemporary American English (COCA)		1.0 billion	American	1990-2019	Balanced
Corpus of Historical American English (COHA)		475 million	American	1820-2019	Balanced
The TV Corpus		325 million	6 countries	1950-2018	TV shows
The Movie Corpus		200 million	6 countries	1930-2018	Movies
Corpus of American Soap Operas		100 million	American	2001-2012	TV shows
Hansard Corpus		1.6 billion	British	1803-2005	Parliament
Early English Books Online		755 million	British	1470s-1690s	(Various)
Corpus of US Supreme Court Opinions		130 million	American	1790s-present	Legal opinions
TIME Magazine Corpus		100 million	American	1923-2006	Magazine
British National Corpus (BNC) *		100 million	British	1980s-1993	Balanced
Strathy Corpus (Canada)		50 million	Canadian	1970s-2000s	Balanced
CORE Corpus		50 million	6 countries	2014	Web
From Google Books n-grams (compare)					
American English		155 billion	American	1500s-2000s	(Various)
British English		34 billion	British	1500s-2000	(Various)

Mega Korpus



Search in 906 Corpus-Based Monolingual Dictionaries for 291 Languages.

Selected language: Indonesian > Mixed 2013

Enter a word



Search suggestions: tambah · stabil · menderita · sukses · timur

More information about: Indonesian > Mixed 2013

[Change corpus](#)

The corpus **ind_mixed_2013** is a Indonesian mixed corpus based on material from 2013. It contains 74,329,815 sentences and 1,206,281,985 tokens. [Details](#)



DOWNLOADS

Download parts of this corpus.



STATISTICS

More details about this corpus on our corpus and language statistics page.

<https://corpora.uni-leipzig.de>

Download Corpora Indonesian

To download a corpus select a corpus size - given in number of sentences - and download the corresponding data

German English French Arabic Russian Indonesian All Languages

Mixed ?

Year	Country	Downloads
2012		10K 30K 100K 300K 1M 3M

Mixed-tufs4 ?

Year	Country	Downloads
2012		10K 30K 100K 300K 1M 3M

News ?

Year	Country	Downloads
2008		10K 30K 100K 300K 1M 3M
2009		10K 30K 100K 300K 1M 3M
2010		10K 30K 100K 300K 1M 3M
2011		10K 30K 100K 300K 1M 3M
2012		10K 30K 100K 300K 1M 3M
2019		10K 30K 100K 300K 1M 3M
2020		10K 30K 100K 300K 1M 3M

Konsep *korpus* (dan jenis-jenisnya)

Konsep korpus (dan jenis-jenisnya)

- (1) I did get a postcard from him. **Korpus Mentah**
- (2) I_I did_do get_get a_a postcard_postcard from_from him_he._punct **Lematisasi**
- (3) I<PersPron> did<VerbPast> get<VerbInf> a<Det> postcard<NounSing>
from<Prep> him<PersPron>.<punct> **Kelas Kata (Part of Speech tags)**
- (4) [@:]·I·^did·get·a·!p\ostcard·fr/om·him#·-·-·- **Anotasi fonologis**
- (5) <Subject, ·NP>
I<PersPron>
<Predicate, ·VP>
did<Verb> **Pemilahan struktur sintaksis (Syntactic parsing)**
get<Verb>
<DirObject, ·NP>
a<Det>
postcard<NounSing>
<Adverbial, ·PP>
from<Prep>
him<PersPron>.
- (6) *CHI: I did get a postcard from him
%mor: pro|I·v|do&PAST·v|get·det|a·n|postcard·prep|from·
pro|him·.
%lex: get **Anotasi bertingkat**
%syn: trans

Korpus Beranotasi

Gries, Stefan Th. 2017. *Quantitative corpus linguistics with R: A practical introduction*. Second edition. New York: Routledge.

Konsep *korpus* (dan jenis-jenisnya)

Korpus Umum & Korpus Khusus

Korpus Mentah & Korpus Beranotasi

Korpus Diakronis & Korpus Sinkronis

Jenis-jenis korpus (Gries 2017: 9-11)


Korpus Monobahasa & Korpus Paralel

Korpus Statis & Korpus Dinamis

Korpus Pembelajar (*Learner Corpora*)

Gries, Stefan Th. 2017. *Quantitative corpus linguistics with R: A practical introduction*. Second edition. New York: Routledge.

Butir pembahasan

1. ~~Konsep *korpus*~~
2. Target Bahasa 
3. Sumber (data & daya)
4. Aksesibilitas + Infrastruktur
5. Contoh kasus:
 - a. Bahasa Enggano
 - b. Bahasa Bali
 - c. Bahasa-bahasa di Australia (LDaCA)

Bahasa daerah yang seperti apa?

Bahasa daerah “besar”?

- Jumlah penutur besar
- Vitalitas pemakaian tinggi
- Aksesibilitas calon data tersedia luas

Bahasa daerah “besar”?

atau *juga*

- Jumlah penutur besar
- Vitalitas pemakaian tinggi
- Aksesibilitas calon data tersedia luas

Bahasa daerah minoritas?


- Jumlah penutur sedikit
- Vitalitas/keberlangsungan terancam
- Belum diteliti dan dijelaskan sebelumnya
(*under-described/undocumented languages*)
- Aksesibilitas data terbatas
(*underresourced*)

Bahasa daerah minoritas?

(lebih banyak) mendapat perhatian peneliti luar Indonesia

- Jumlah penutur sedikit
- Vitalitas/keberlangsungan terancam
- Belum diteliti dan dijelaskan sebelumnya
(*under-described/undocumented languages*)
- Aksesibilitas data terbatas
(*underresourced*)

Butir pembahasan

- ~~1. Konsep *korpus*~~
- ~~2. Target Bahasa~~
3. Sumber (data & daya) 
4. Aksesibilitas + Infrastruktur
5. Contoh kasus:
 - a. Bahasa Enggano
 - b. Bahasa Bali
 - c. Bahasa-bahasa di Australia (LDaCA)

Sumber (data & daya)

1#

Apakah bahasa target sudah memiliki sumber calon data dalam format digital?

(fitur pertama dari konsep *korpus* sebelumnya)

1.1#

Jika telah dalam format digital,
apakah calon datanya
computer-readable (mis. bisa
dilakukan pencarian kata, dll)
atau masih berupa *image*?
(fitur pertama dari konsep *korpus* sebelumnya)

1#

Apakah bahasa target sudah memiliki sumber calon data dalam format digital?

(fitur pertama dari konsep *korpus* sebelumnya)

Jika “belum”, berarti kita perlu melakukan tahap digitalisasi terlebih dahulu (jika formatnya cetak) - yang merupakan proyek tersendiri + sumber daya

1#

Apakah bahasa target sudah memiliki sumber calon data dalam format digital?

(fitur pertama dari konsep *korpus* sebelumnya)

ATAU mungkin (1) perlu melakukan DOKUMENTASI bahasa terlebih dahulu, dan berkolaborasi dengan komunitas + linguis - yang juga proyek tersendiri

1#

Apakah bahasa target sudah memiliki sumber calon data dalam format digital?

(fitur pertama dari konsep *korpus* sebelumnya)

ATAU mungkin (2) DOKUMENTASI TELAH DILAKUKAN UNTUK BAHASA DAERAH TERSEBUT SEBELUMNYA (dibahas pada studi kasus nanti)

2#

Apakah sumber calon data tersebut mencerminkan beragam jenis teks? modalitas?
(fitur keterwakilan kandungan *korpus*)

2#


Apakah sumber calon data tersebut mencerminkan beragam jenis teks? modalitas?
(fitur keterwakilan kandungan *korpus*)

Jika hanya beberapa ragam teks, mungkin kita perlu bersikap PRAGMATIS - gunakan semaksimal mungkin data yang ada (lih. solusi keterwakilan pada bagian “konsep korpus”)

Ke(tidak)beradaan calon data +
tujuan pembangunan
korpusnya memungkinkan
adanya paket kerja secara
bertahap

(contoh pada studi kasus berikutnya)

Butir pembahasan

- ~~1. Konsep *korpus*~~
- ~~2. Target Bahasa~~
- ~~3. Sumber (data & daya)~~
4. Aksesibilitas + Infrastruktur 
5. Contoh kasus:
 - a. Bahasa Enggano
 - b. Bahasa Bali
 - c. Bahasa-bahasa di Australia (LDaCA)

Aksesibilitas + infrastruktur

Aksesibilitas

Tertutup?

Terbuka?

Aksesibilitas

Tertutup?

Terbuka?

- dapat diunduh? (mis. [PARADISEC](#), [ELAR](#), [LDAcA](#))
- antar muka daring saja? (mis. [english-corpora.org](#), [CQPWeb](#))
- antar muka daring + bisa diunduh (mis. [Leipzig Corpora](#))
- lisensi cc-by? (mis. seperti di Leipzig Corpora)
- atribusi (mis. sitiran) terhadap penyedia/penyusun korpus

Aksesibilitas

FAIR data principle (<https://www.go-fair.org/fair-principles/>)

F(indable) - pengguna bisa menemukan

A(ccessible) - cara mengakses data tsb.

I(nteroperability) - interaksi dengan peranti lain

R(eusable) - penggunaan ulang (untuk tujuan berbeda)

Leipzig Corpora

Contoh korpus dengan FAIR
principle

Leipzig Corpora

Contoh korpus dengan FAIR
principle

FINDABLE



CORPORA COLLECTION
LEIPZIG

Search in 1018 Corpus-Based Monolingual Dictionaries for 290 Languages.

Selected language: [English](#) > [News 2012](#)



Search suggestions: [scandal](#) · [golf](#) · [join](#) · [involving](#) · [known as](#)

More information about: [English](#) > [News 2012](#)

[Change corpus](#)

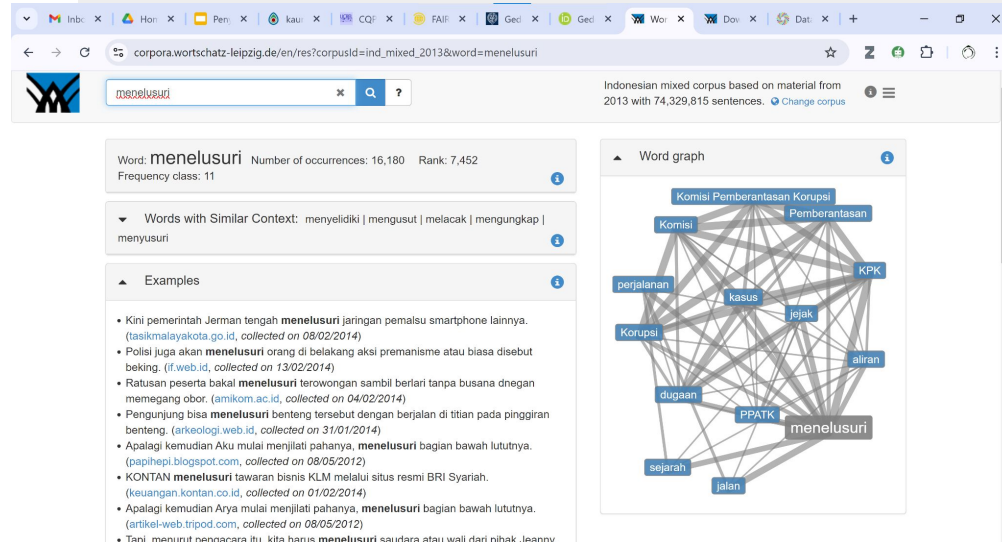
The corpus [eng_news_2012_3M](#) is a English news subcorpus based on material from 2012 (3,000,000 sentences). It contains 3,000,000 sentences and 62,393,073 tokens.

[Details](#)

ACCESSIBLE

Leipzig Corpora

Contoh korpus dengan FAIR principle



The screenshot displays the Leipzig Corpora website interface for the word "menelusuri". The browser address bar shows the URL: corpora.wortschatz-leipzig.de/en/res?corpusid=ind_mixed_2013&word=menelusuri. The search bar contains the word "menelusuri".

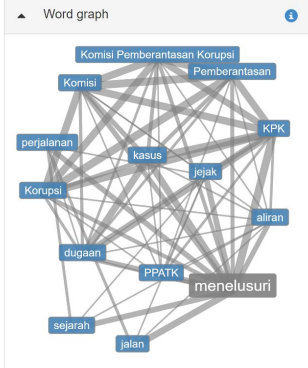
Word: **menelusuri** Number of occurrences: 16,180 Rank: 7,452
Frequency class: 11

Words with Similar Context: menyelidiki | mengusut | melacak | mengungkap | menelusuri

Examples

- Kini pemerintah Jerman tengah **menelusuri** jaringan pemalsu smartphone lainnya. (tasikmalayakota.go.id, collected on 08/02/2014)
- Polisi juga akan **menelusuri** orang di belakang aksi premanisme atau biasa disebut beking. (f.web.id, collected on 13/02/2014)
- Ratusan peserta bakal **menelusuri** terowongan sambil berlari tanpa busana dnegan memegang obor. (amikom.ac.id, collected on 04/02/2014)
- Pengunjung bisa **menelusuri** benteng tersebut dengan berjalan di titian pada pinggir benteng. (arkeologi.web.id, collected on 31/01/2014)
- Apalagi kemudian Aku mulai menjilati pahanya, **menelusuri** bagian bawah lututnya. (papihepi.blogspot.com, collected on 08/05/2012)
- KONTAN **menelusuri** tawaran bisnis KLM melalui situs resmi BRI Syariah. (keuangan.kontan.co.id, collected on 01/02/2014)
- Apalagi kemudian Arya mulai menjilati pahanya, **menelusuri** bagian bawah lututnya. (artikel-web.tripod.com, collected on 08/05/2012)
- Tapi, menurut pencacara itu, kita harus **menelusuri** saudara atau wali dari pihak Jeanny.

Word graph



The word network diagram shows "menelusuri" at the center, connected to various related terms: "KPK", "jejak", "kasus", "Korupsi", "PPATK", "jalan", "sejarah", "dugaan", "perjalanan", "Komisi", "Korupsi", "KPK", "jejak", "aliran".

Cara mengakses daring + fitur infrastruktur terkait analisis (mis. word network, word co-occurrence)

Leipzig Corpora

Contoh korpus dengan FAIR
principle

ACCESSIBLE



The screenshot shows the Wortschatz Leipzig website. At the top, there is a search bar with the text "Search in more than 30 million sentences of German newspaper material:" and a search button. Below the search bar, there are four links: "Download Corpora", "Download SentWS", "Download TinyCC", and "Download ASV Toolbox". The main heading is "Leipzig Corpora Collection Download Page". Below this, there is a paragraph: "The Leipzig Corpora Collection provides different tools and data for download, which are protected by copyright. For more details please refer to our [terms of usage](#)." There is a section titled "Download Corpora" with a description: "The Leipzig Corpora Collection presents corpora in different languages using the same format and comparable sources. All data are available as plain text files and can be imported into a MySQL database by using the provided import script. They are intended both for scientific use by corpus linguists as well as for applications such as knowledge extraction programs. The corpora are identical in format and similar in size and content. They contain randomly selected sentences in the language of the corpus and are available in sizes from 10,000 sentences up to 1 million sentences. The sources are either newspaper texts or texts randomly collected from the web. The texts are split into sentences. Non-sentences and foreign language material was removed. Because word co-occurrence information is useful for many applications, these data are precomputed and included as well. For each word, the most significant words appearing as immediate left or right neighbor or appearing anywhere within the same sentence are given. More information about the format and content of these files can be found [here](#). The corpora are automatically collected from carefully selected public sources without considering in detail the content of the contained text. No responsibility is taken for the content of the data. In particular, the views and opinions expressed in specific parts of the data remain exclusively with the authors." Below this, there is a note: "If you use one of these corpora in your work we kindly ask you to cite [this paper](#) as". At the bottom, there is a citation: "D. Goldhahn, T. Eckart & U. Quasthoff: Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In: *Proceedings of the 8th International Language Resources and Evaluation (LREC'12)*, 2012".

Cara mengakses (termasuk
mengunduh) sangat eksplisit & terbuka

T&C + cara atribusi jika menggunakan

Leipzig Corpora

Contoh korpus dengan FAIR
principle

INTEROPERABLE

“the data need to
interoperate with
applications or workflows
for analysis, storage, and
processing.”

<https://www.go-fair.org/fair-principles/>

INTEROPERABLE

Leipzig Corpora

Contoh korpus dengan FAIR principle

AntConc

File Edit Settings Help

Target Corpus

Name: temp

Files: 1

Tokens: 15102635

ind_mixed_2012_1M-sentences.txt

KWIC Plot File View Cluster N-Gram Collocate Word Keyword Wordcloud

Total Hits: 3089 Page Size 100 hits 1 to 100 of 3089 hits

	File	Left Context	Hit	Right Context
1	ind_...	asi yang kurang solid, tidak memiliki visi, tidak	terbuka,	dan tidak cepat tanggap akan membawa damp
2	ind_...	da ruh awal kemunculannya, sebagai kelomok	terbuka	dan tidak eksklusif, tentunya peran sosial politik
3	ind_...	erugian dengan bus yg cukup bagian dadanya	terbuka	dan tidak memakai bra. 804108 Sebab, lahan y
4	ind_...	donesia (APPSI). 543614 Ada juga yang begitu	terbuka	dan tidak mempersoalkan perbedaan agama. 5
5	ind_...	ih cukup memberi bukti bahwa kami memang	terbuka	dan tidak mempunyai maksud terselubung bag
6	ind_...	endang sajalah telapak kaki dari lawannya itu "	terbuka"	dan tidak terlindung atau tersembunyi. 414938
7	ind_...	gan cara itu akhirnya anda akan bersikap lebih	terbuka	dan tidak terpaku di satu sudut pandang belak
8	ind_...	ngan visi Yesus. 246286 "Perlu diadakan secara	terbuka	dan tidak tertutup tujuannya biar bisa menjadi
9	ind_...	epatlah "perbaiki" apa maunya, dengan saling	terbuka,	dan jujur bersama suami. 188965 Lima menitir
10	ind_...	ra tercinta. 412040- Saya merasa bisa bersikap	terbuka	dan jujur disini - Saya mendapatkan kegembira
11	ind_...	berkata, "Saya berpendapat bahwa kita harus	terbuka	dan jujur mengenai bagaimana metode pende
12	ind_...	in meracai ini untukmu. 932117 Lebih baik 23	Terbuka	dan jujur satu sama lain. 932118 Saya khawatir

Search Query Words Case Regex Results Set All hits Context Size 10 token(s)

terbuka Start Adv Search

Sort Options Sort to right Sort 1 1R Sort 2 2R Sort 3 3R Order by freq

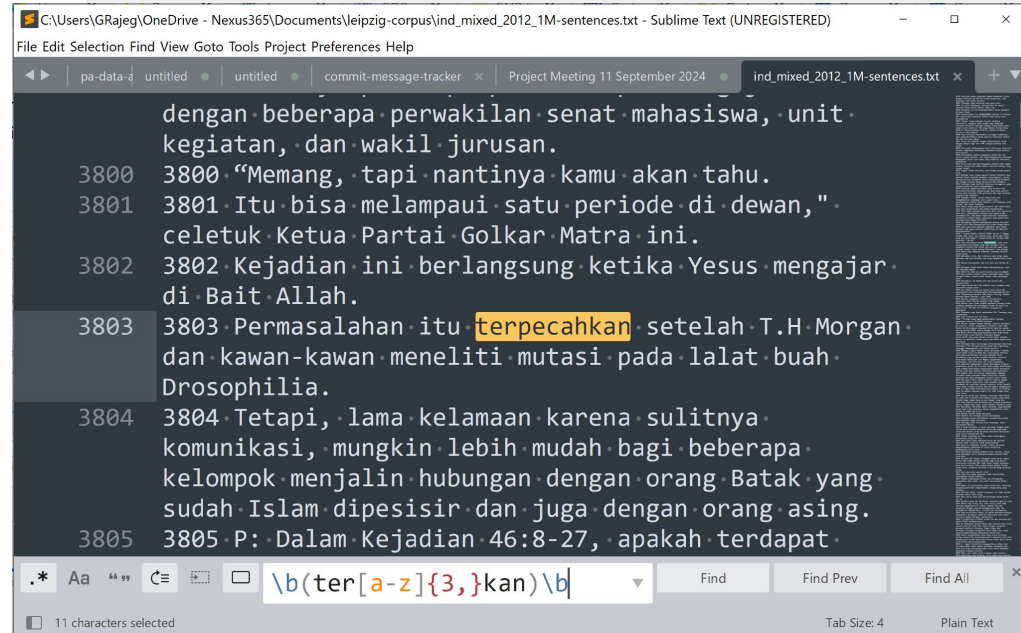
Progress 100%

Time taken (creating KWIC results): 0.522 sec

INTEROPERABLE

Leipzig Corpora

Contoh korpus dengan FAIR principle



```
C:\Users\GRajeg\OneDrive - Nexus365\Documents\leipzig-corpus\ind_mixed_2012_1M-sentences.txt - Sublime Text (UNREGISTERED)
File Edit Selection Find View Goto Tools Project Preferences Help
pa-data | untyped | untyped | commit-message-tracker | Project Meeting 11 September 2024 | ind_mixed_2012_1M-sentences.txt
dengan beberapa perwakilan senat mahasiswa, unit kegiatan, dan wakil jurusan.
3800 3800 "Memang, tapi nantinya kamu akan tahu.
3801 3801 Itu bisa melampaui satu periode di dewan," celetuk Ketua Partai Golkar Matra ini.
3802 3802 Kejadian ini berlangsung ketika Yesus mengajar di Bait Allah.
3803 3803 Permasalahan itu terpecahkan setelah T.H Morgan dan kawan-kawan meneliti mutasi pada lalat buah Drosophila.
3804 3804 Tetapi, lama kelamaan karena sulitnya komunikasi, mungkin lebih mudah bagi beberapa kelompok menjalin hubungan dengan orang Batak yang sudah Islam dipesisir dan juga dengan orang asing.
3805 3805 P: Dalam Kejadian 46:8-27, apakah terdapat
.* Aa "" ☰ ☷ \b(ter[a-z]{3,}kan)\b Find Find Prev Find All
11 characters selected Tab Size: 4 Plain Text
```

Leipzig Corpora

Contoh korpus dengan FAIR principle

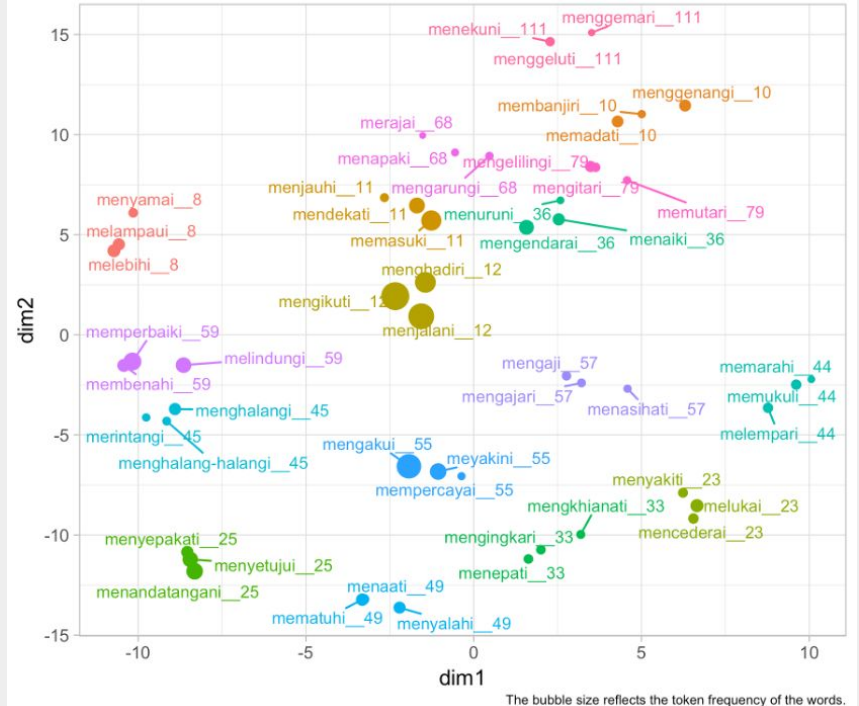
Rajeg, G. P. W., Denistia, K., & Musgrave, S. (2019). Vector Space Models and the usage patterns of Indonesian denominal verbs: A case study of verbs with meN-, meN-/kan, and meN-/i affixes. *NUSA*, 67, 35–75.
<https://doi.org/10.15026/94452>

Rajeg, G. P. W. (2024, October 24). *Penyusunan Korpus Bahasa Daerah* [Invited talk]. *Konsinyasi Penyiapan Data Korpus Bahasa Daerah dan Pemetaan Bahasa*. Badan Pengembangan dan Pembinaan Bahasa, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi melalui Pusat Pengembangan dan Pelindungan Bahasa dan Sastra, Online. <https://doi.org/10.25446/oxford.27867759>

INTEROPERABLE

Semantic vector space representation for 'me- -i' words that fall into a 3-member clustering.

The numbers following each words show the clusters of the words. For instance, 'menyamai', 'melampai', and 'melebihi' belong to the same cluster (i.e. cluster no. 8). The clustering is derived from the 'Partition Around Medoid (PAM)' clustering technique (cf. Levshina, 2015, p. 319).



The bubble size reflects the token frequency of the words.

REUSABLE

Leipzig Corpora

Contoh korpus dengan FAIR principle


The screenshot shows a web browser displaying a Figshare dataset page. The URL is figshare.com/articles/dataset/Dataset_for_i_Vector_space_model_and_the_usage_patterns_of_Indonesian_denominal_verbs_i/8187155?file.... The page lists three files:

- ngramexempl_3gr_melangkah... txt (57.73 kB)
- ngramexempl_3gr_menapak.txt (42.93 kB)
- ngramexempl_3gr_menapaki.txt (91.97 kB)


A message states: "sorry, we can't preview this file ...but you can still download [leipzig_w2v_vector_full.bin](#)". Below the file list, there is a "Switch View" button and a file viewer for "leipzig_w2v_vector_full.bin (72.16 MB)". The dataset title is "Dataset for *Vector space model and the usage patterns of Indonesian denominal verbs*". Action buttons include "Cite", "Download all (114.73 MB)", "Share", "Embed", and "+ Collect". The dataset was posted on 2019-10-12, 04:55, authored by Gede Primahadi Wijaya Rajeg, Karlina Denistia, Simon Musgrave. Usage metrics show 2938 views and 3141 downloads, with the latter circled in red.

<https://doi.org/10.6084/m9.figshare.8187155.v1>

Butir pembahasan

- ~~1. Konsep *korpus*~~
- ~~2. Target Bahasa~~
- ~~3. Sumber (data & daya)~~
- ~~4. Aksesibilitas + Infrastruktur~~
5. Contoh kasus :
 - a. Bahasa Enggano
 - b. Bahasa Bali
 - c. Bahasa-bahasa di Australia (LDaCA)

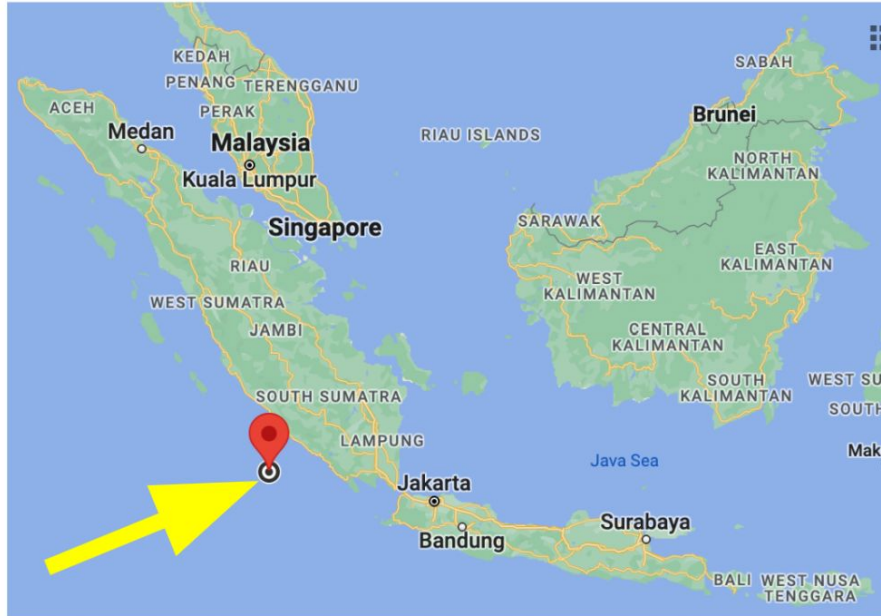
Butir pembahasan

- ~~1. Konsep *korpus*~~
- ~~2. Target Bahasa~~
- ~~3. Sumber (data & daya)~~
- ~~4. Aksesibilitas + Infrastruktur~~
5. Contoh kasus :
 - a. Bahasa Enggano ([sumber data + daya, aksesibilitas](#))
 - b. Bahasa Bali ([sumber data + daya; model gerakan komunitas; infrastruktur awal](#))
 - c. Bahasa-bahasa di Australia (LDaCA) ([sumber data + daya + infrastruktur nasional](#))

Acknowledgment: mèk em nah Enggano team & community



The Enggano language and its speakers



- Austronesian
- ± 1,500 speakers
 - **endangered**
 - increasing shift towards Indonesian

<https://enggano.ling-phil.ox.ac.uk/>

(Hemmings & Arka 2023)

Hemmings, Charlotte & I Wayan Arka. 2023. Evidence of contact with Malay/Indonesian in the Enggano Language. (Ed.) Hiroki Nomoto & Asako Shiohara. *NUSA: Linguistic studies of languages in and around Indonesia* (Language Contact between Malay and Indigenous Languages) 75. 19–51. <https://doi.org/10.15026/0002000126>.

Dokumentasi Bahasa Kontemporer

Topik teks beragam
terkait budaya dan
adat Enggano

Introduction (Wikitongues)

This video is also available on [YouTube](#) and on the [Wikitongues Enggano site](#).

Participants/*Peserta*: Milson Kaitora, Engga Zakaria Kauno

Date of recording/*Tanggal perekaman*: March 2023

Location/*Lokasi*: Oxford

[Video recording with transcription/Rekaman video dengan transkripsi \(LingView\)](#)

Downloadable resources/*Bebagai bahan yang dapat diunduh*:

Audio file/*Berkas audio* (WAV): [Wikitongues.wav](#)

Video file/*Berkas video* (MP4): [Wikitongues.mp4](#)

Adat Perkawinan (Wedding tradition)

Participants/*Peserta*: Piterjun Kaohao, I Wayan Arka

Date of recording/*Tanggal perekaman*: 6 February 2018

Location/*Lokasi*: Enggano Island

[Video recording with transcription/Rekaman video dengan transkripsi \(LingView\)](#)

Downloadable resources/*Bebagai bahan yang dapat diunduh*:

Audio file/*Berkas audio* (WAV): [eno_20180206_adat_perkawinan.wav](#)

Video file/*Berkas video* (MP4): [eno_20180206_adat_perkawinan.mp4](#)

Bakblau

Participants/*Peserta*: Harun Kaarubi, I Wayan Arka

Date of recording/*Tanggal perekaman*: 6 February 2018

Location/*Lokasi*: Enggano Island

[Video recording with transcription/Rekaman video dengan transkripsi \(LingView\)](#)

Downloadable resources/*Bebagai bahan yang dapat diunduh*:

Audio file/*Berkas audio* (WAV): [eno_20180206_bakblau.wav](#)

Video file/*Berkas video* (MP4): [eno_20180206_bakblau.mp4](#)

<https://enggano.ling-phil.ox.ac.uk/static/recordings.html>

Dokumentasi Bahasa Kontemporer

Transkripsi dan
terjemahan melalui
ELAN

The screenshot displays the ELAN 6.7 software interface. At the top, the title bar reads "ELAN 6.7 - eno_20231014_missing-from-FLEEx_04.eaf". Below the menu bar, the "Grid" tab is active, showing a list of annotations for the word "eno-word-MK". The selected annotation is "kabua", with a list of related terms: "kita · kibak · onea' · paraeah · na'pe iate' pen ean · kahinén karié · kahúr karya · selus kahuar úki sore ne'en · kähä · · parúda' · kebea'".

The interface also shows a timeline with a selection range from 00:16:42.825 to 00:16:45.475. Below the timeline, there are several tracks for annotations, including "AK_Word-txt-eno", "AK_Word-gls-id", "AK_Sense-sn-en", "AK_Lexeme-txt-eno", "AK_Participant-note-en", "eno-word-MK", "eno-word-gloss-idn", "note", and "eno-sent-gloss-idn". A list of annotations is visible on the right side of the interface, including "kabua'", "membelah", "add to FLEEx", and "Selus membelah kelapa".

Dokumentasi Bahasa Kontemporer

Terjemahan leksikon
dan anotasi morfologis
di FLE_x

Fieldwork Language
Explorer

Contemporary-Enggano-FLEx-database-20241004 - FieldWorks Language Explorer

File Send/Receive Edit View Data Insert Format Tools Parser Window Help

Enggano

Texts

Interlinear Texts
Concordance
Complex Concordance
Word List Concordance
Word Analyses
Bulk Edit Wordforms
Statistics

Title: SD Teaching Materials - Unit 3

Info Baseline Gloss Analyze Tagging Print View Text Chart

3.1 Word U buh kũda' yahbari' kerupuk emping kũtã i Enggano

Morphemes u buh kũda' y- ah- bari' kerupuk emping kũtã i enggano

Lex. Entries u buh₁ kũda' e- ah- pari' kerupuk ping kũtã i₂ Enggano

Lex. Gloss *bst. an* 1SG want tell NM (Noun marker) ANTIP make cracker cracker nut tree sp. LOC Enggano

Lex. Gloss *Ind* saya mau beritahu Awalana penanda kata benda kerupuk emping melinjo di Enggano

Lex. Gram. Info. pro aux v Noun v:(Voice) v n n (Indo) n prep npro

Word Gloss *Eng* I want tell making cracker cracker melinjo at Enggano

Word Gloss *Ind* saya mau beritahu buat kerupuk emping melinjo di Enggano

Word Cat. pro aux v n n n prep npro

Free *Eng* I want to talk about making melinjo crackers in Enggano.

Ind Saya ingin memberitahu pembuatan emping melinjo di Enggano.

Lit. *Eng*

Ind

3.2 Word Pehde ik kibuh baher

Morphemes peh =de ik ki- buh b- ah- er

Lex. Entries peh =de ik ki-₁ buh₁ bu- ah- er₁

Lex. Gloss *bst. an* origin 3SG.POSS 1PL.INCL KI want BU ANTIP climb a tree

Lex. Gloss *Ind* awal =nya kita mau BU Awalana penanda kata kerja (verba) mau BU memanjat pohon

Lex. Gram. Info. n det pro v:Any aux Attaches to any category v:(Voice) v

Word Gloss *Eng* first we want climb

Word Gloss *Ind* pertamanya kita mau memanjat

Word Cat. adv pro aux v

kũtã mẽ' homanak

kũtã mẽ' ho = m- a- nak

kũtã mẽ' ho = bu- a- nak₁

nut tree sp. REL already (PERF) BU VBLZ (verbaliser) ripe fruit

22/Oct/2024 Queue: (-/-) No Parser Loaded Sorted by Title 53/79

Yang menarik dari
Enggano:

Kekayaaninggalan
material (*legacy
materials*) sejak
pertengahan abad 19

Teks Korpus Bahasa Enggano Mula (1916) dengan terjemahan bahasa Melayu (oleh O. L. Helfrich)

520 BIJDRAGE TOT DE KENNIS VAN HET ENGGANEESCH.

D. VERHALEN (Ekoedaäjo).

I. Ekoedaäjo ekoanoe kiphóna hi ekaka.
Tjeritera tikoes kawin sama orang.

E. Kikia hale ilopo eana aroea epoinamo moö,
M. Adalah doeloe dinegrie itoe doea gadis jang
E. káoea epoeahhadia, kamoephi ipahona hi epo-
M. amat elok roepanja tiada maoe kawin dengan boe-

Helfrich, O. L. (1916). Nadere bijdrage tot de kennis van het engganeesch. *Bijdragen Tot de Taal-, Land- En Volkenkund*, 71(1), 472–555.

<https://enggano.ling-phil.ox.ac.uk/static/previous.html>

PDF-scanned -
digitalisasi calon
sumber data yang
ada

Teks Korpus Bahasa Enggano Mula (1930-an) dengan terjemahan bahasa Jerman (oleh Hans Kähler)

10. ke'anaha kaminā'ūyāhā eka'e'e yahag i'kē'ā ukua, kanīō ekanī'kōō. kapu'udaha eka'e'e e'ana, be kia ka'ekoi⁷⁸⁾ eiri' upae i'isōnīā. kabudodo hēmō's e'apō ukua ka'ō; kakunā'āhā⁷⁹⁾ yopu'udaha itōpō⁸⁰⁾, yapakō'a epuedia ipu'udahadia. kamōshōhā epae e'ana kabupua, yahag yabakiu⁸¹⁾ i'kō ukāi ukēpū⁸²⁾. ke'anaha kahii mū'ēnūhā⁸³⁾ eka'e'e ipu'udahadia nē'ēnī, kabu'uaha nō'si'īē eka'e'e eiri': "pakōshō⁸⁴⁾ epue'ey⁸⁵⁾! upōa kia, be bupua kia i'icō'ōy!"

11. ke'anaha kahii ba'icōia epae e'ana; i'kii apakō'shō'ā⁸⁶⁾ kia kikia i'kō ukāi ukēpū. e'uhag ukō'e'e: "ke adō uia? 'ua kipōa 'a'ā ki'kēhēkū i'kō ukāi ukēpū!" ke'anaha kahii ba'ia epae e'ana. dipuakaha, dahii babai'ā dai iuba. kanō'sāhāmā'ā edī'ōbuda ki. dababai, ke'anaha kīkii bupōaha ehūā ukanī'kōō kahai'. kahii bahagēda'ā eka'e'e, i'kii abgēda'ā ehūā ukanī'kōō. kahii bu'uada'ā epae e'ana: "yara'ā budōō kōda, be 'ōpu'uda, yabakōu⁸⁷⁾ epuebu! 'ōbudōō hēmō's ka'ō!" kanō'sāhā eiri' upae eiri'e.

12. kahii mīnā'ūyā eka'e'e e'ana i'kē'ā ukanī'kōō. kahii pu'udada'ā. be edī'ōbu upahēnēhā epae e'ana mōhōikū i'kōmōnīā'ā kia mūnī'kōō⁸⁸⁾. kabupua epae e'ana, i'kii bahag, kahag⁸⁹⁾ iuba. ya-kōkōnā⁹⁰⁾ 'ōu'ūpūnā mō's karanāmū⁹¹⁾ ukupadō. e'uhagādia: "mēō upupua?", e'uhag ukōpū e'ana i'icō ukōshōpīōnīā e'ana. e'uhag ukōshōpīōnīā uadōda'ā i'isōnīā: "u'uada⁹²⁾ eka'e'e kikarar⁹³⁾ 'ua!" "mēō i'ōbuda nō'sāhā i'isōmō?" "he kia kahāpī mūnōō 'ua!", kanō'sāhā edī'ua upae eiri'e. "ape anō'sāhā edī'ōbudia i'isōmō, 'ōbai ite'ē! 'upakū 'ō'ā i'kē'ē ukupadō eiri'e!" "he nō'sāhā!", e'uhag upae e'ana bayō'ōi eiri' u'ūpūnīā. kabahagā yabakūaha itita.

13. ka e'anaha kabai'ā eka'e'e hēmō's ka'icōi kia nē'ēnī. yakōkōnāhā 'ōukāpū e'ana kanīō e'abakōkō⁹⁴⁾. e'uhag ukōpū e'ana: "mēō u'ō'ōbua'ā 'ō'ō? 'ō'ō kahāpī mūnōō ekōshōpīō'sū", e'uhag ukōpū e'ana pauaha'ā⁹⁵⁾ ekōshōpīōnīā yara'ā kua inōō kia eka'e'e. ke'anaha kapahō⁹⁶⁾ kia eka'e'e. kabu'uaha nō'si'īē i'icō ukōshōpīōnīā: "he abōhal 'ōbahag iubabu, uobakōi⁹⁷⁾ e'ei-bu⁹⁸⁾ eka'ō, ita'ōyā abai kia nā'ānī ainōōō kia eka'ō e'ana!" kanō'sāhā edī'ua ukōpū nē'ēnī i'icō ukōshōpīōnīā.

14. ke e'anaha kahii ...

Kähler, H. (1955). Ein Text von der Insel Enggano (Westküste von Sumatra). *Afrika Und Übersee*, 39, 89–94.

PDF-scanned - digitalisasi calon sumber data yang ada

<https://enggano.ling-phil.ox.ac.uk/static/previous.html>

iegen. Zu der Zeit trug der epākū genannte Baum Früchte, den die Leute z waren. — Deshalb gingen sie zu dritt mit ihrem Kinde nach der Pflanzung. Di der Pflanzung, die bereits überwuchert war. Auch ihr Mann machte sich auf den Schlinge (Vögel) bei jenem epākū-Baum. Er fing eine Wildtaube, die er mit nahm, die der Aufenthaltsort seiner Frau mit ihrem Kinde war. Aber ihr Kind ikeit hatte es überfallen, und sein Körper war matt.

2. Sein Vater machte sich auf den Weg und kam dort an. (Da) sprach er zu eine Wildtaube als Nahrung für uns hier!" "Ja!", erwiderte seine Frau als Ar ihres Mannes. Alsdann briet sie die Wildtaube. "Nun, sie ist gar!" Darauf spr dermassen zu ihrem Manne: "Lass uns nur essen! Los, wecke unser Kind, I 3. Deshalb weckte er sein Kind, (aber) es vermochte nichts zu spüren. "Wi dern) lass uns vorweg essen! Nachher, wenn es erwacht, wird es wieder ess antwortete ihr Mann, Und dann assen sie.

4. Nachdem sie mit dem Essen fertig waren, nahmen sie ihre Arbeit wieder Nacht. "Nun, wir sind fertig mit der Arbeit, und wir gehen nach Hause, denn t Los, wecke unser Kind!" Er machte sich auf den Weg, und dann weckte er de Es vermochte jedoch nichts zu spüren, weil es wie ein Toter schlief. Deshalb s Wassollen wir mit ihm anfangen? Denn (es kann die Erde nicht fühlen =) ist h tun folgendermassen, vielleicht kann es das dann spüren!"

5. Und dann holten sie Ameisen, die (es) bissen. Aber es spürte nichts. "I mit ihm anfangen? Wie ist es, wir haben alle Ameisen genommen, und sie hab sen. Nun, so sei es, lass uns wieder etwas anderes suchen. Lass es uns noc einem Tausendfüssler!" Sie holten einen Tausendfüssler, und sie liessen ih Aber es vermochte nichts zu spüren. "Los, höre auf mit ihm! Es schläft ih

Teks Korpus Bahasa Enggano Mula (1930-an) dengan terjemahan bahasa Jerman (oleh Hans Kähler)

Kähler 1955
Enggano and German transcription by Barnaby Burleigh, September 2019
Enggano transcription checked by Mary Dalrymple, February 2020
English translation by Barnaby Burleigh, April 2020
Indonesian translation by I Komang Sumaryana Putra, May 2020
Transcription system: common transcription



I. ekúda'ayo ukaka halEE 'akorupada.
I. Erzählung von Leuten in alter Zeit, die zu dritt mit ihrem Kinde waren
I. Tale of People in the old Days who were together (three) with their Child
I. Kisah Orang pada Zaman dahulu (berjumlah tiga) bersama dengan Anak mereka

1. ki kaha:E ipiada, kaha:E ba'apia.
1. Sie machten sich auf den Weg in ihre Pflanzung, (sie) wollten gehen, um einen Garten anzulegen.
1. They started towards their plantation, (they) wanted to go in order to make a garden.
1. Mereka mulai menuju ke perkebunan mereka, (mereka) pergi untuk membuat kebun.

kabia e'ana ea'ahā:ūāmāhā ukuo hēmō'ō kanīũ epūkā hēmō'ō edixoo ba'ubuoda'a ukaka baka:ühua.
Zu der Zeit trug der epīkâ genannte Baum Früchte, den die Leute zu benutzen gewohnt waren.
At that time the tree called epīka, which the people were used to using, bore fruit.
Pada saat itu pohon yang disebut epīka, yang biasa digunakan orang-orang, sedang berbuah.

Computer-readable, plain-text (fitur *digital* pada konsep *korpus*)

<https://enggano.ling-phil.ox.ac.uk/static/previous.html>

Selanjutnya juga diolah dengan FLEx

Enggano-Kahler-2023 - FieldWorks Language Explorer

File Send/Receive Edit View Data Insert Format Tools Parser Window Help

English

Texts & Words **Texts** **Text**

Interlinear Texts
Concordance
Complex Concordance
Word List Concordance
Word Analyses
Bulk Edit Wordforms
Statistics

Show All

1955
1957
1958
1960a
1960b
1961
1962
1964
1975 Bootsrennen
1975 Damonen-Vorstellungen
1975 Dorfleben
1975 Ermordung
1975 Heiratsvorschriften
1975 Kinderverlobnis
1975 Krieg
1975 Mitgift
1975 Rechtsprechung
1975 Tod
1975 Unterweisung

Title Eno-Kah
Eng 1955

Info Baseline Gloss Analyze Tagging Print View Text Chart

Ger Zu der Zeit trug der epika genannte Baum Früchte, den die Leute zu benutzen gewohnt waren.

1.5 Word kE'anaha dibaha:Eha ipia

Morphemes	kE =	'ana	= ha	di-	b-	aha:E	= ha	i-	pia
Lex. Entries	kE =	'ana	= ha	ki- ₂	bu-	aha:E	= ha	i- ₁	pia
Lex. Gloss	but/then/and	DEM.MEDIAL	EMPH	3PL.SUBJ	INF	go/walk	EMPH	LOC	garden
Lex. Gram. Info.	conn	dem	adv	v:(VAgrPrfx)	v:(VPrefopt)	v1 (Class1)	adv	n:LocArt/(ArtOpt)	subs
Word Gloss	***			***				***	
Word Cat.	dem			Vlex				subs	

e'ana ki 'akorupara

e- 'ana ki '0- 'akodu -pada

e-₁ 'ana ki e-₂ 'akodu -pada

DIR DEM.MEDIAL 3PL.PRO DIR.SG three together

n:(DirLoc)/Art/(ArtOpt) dem pro Nhum:DirHumArt num num>Nhum

*** 3PL.PRO ***

dem pro Nhum

Free Eng Thus they went to the plantation as three with their child.
Ind Jadi mereka pergi ke perkebunan bertiga bersama dengan anak mereka.
Ger Deshalb gingen sie zu dritt mit ihrem Kinde nach der Pflanzung.

Kami memiliki korpus
Enggano Mula &
Kontemporer untuk, mis.,
kajian diakronis perubahan
bahasa Enggano

FLEx bisa diekspor ke XML (interoperability)

Enggano-Kahler-2023 - FieldWorks Language Explorer

File Send/Receive Edit View Data Insert Format Tools Parser Window Help

English

Texts & Words **Texts** **Text**

Interlinear Texts
Concordance
Complex Concordance
Word List Concordance
Word Analyses
Bulk Edit Wordforms
Statistics

Show All

1955
1957
1958
1960a
1960b
1961
1962
1964
1975 Bootsrennen
1975 Damonen-Vorstellungen
1975 Dorfleben
1975 Ermordung
1975 Heiratsvorschriften
1975 Kinderverlobnis
1975 Krieg
1975 Mitgift
1975 Rechtsprechung
1975 Tod
1975 Unterweisung

Title Eno-Kah
Eng 1955

Info Baseline Gloss Analyze Tagging Print View Text Chart

Ger Zu der Zeit trug der epika genannte Baum Früchte, den die Leute zu benutzen gewohnt waren.

1.5 Word	kE'anaha			dibaha:Eha				ipia
Morphemes	kE = 'ana = ha			di- b- aha:E = ha				i- pia
Lex. Entries	kE = 'ana = ha			ki-2 bu- aha:E = ha				i-1 pia
Lex. Gloss	but/then/and DEM.MEDIAL	EMPH	3PL.SUBJ	INF	go/walk	EMPH	LOC	garden
Lex. Gram. Info.	conn dem	adv	v:(VAgrPrfx)	v:(VPrefopt)	v1 (Class1)	adv	n:LocArt/(ArtOpt)	subs
Word Gloss	***			***			***	
Word Cat.	dem			Vlex				subs

e'ana ki 'akorupara

e- 'ana ki '0- 'akodu -pada

e-1 'ana ki e-2 'akodu -pada

DIR DEM.MEDIAL 3PL.PRO DIR.SG three together

n:(DirLoc)/Art/(ArtOpt) dem pro Nhum:DirHumArt num num>Nhum

*** 3PL.PRO ***

dem pro Nhum

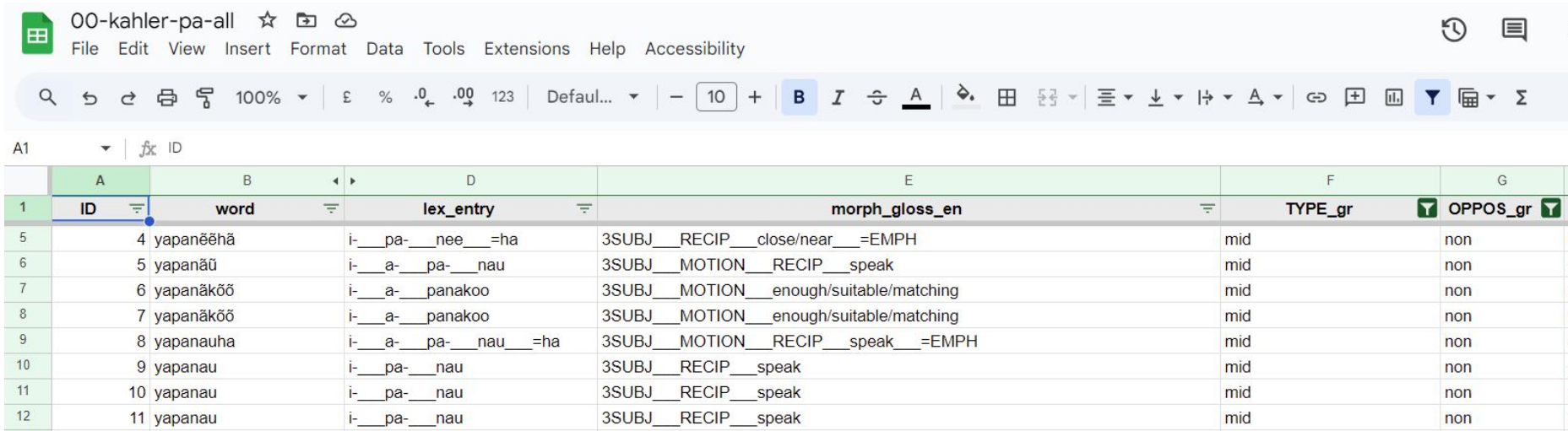
Free Eng Thus they went to the plantation as three with their child.
Ind Jadi mereka pergi ke perkebunan bertiga bersama dengan anak mereka.
Ger Deshalb gingen sie zu dritt mit ihrem Kinde nach der Pflanzung.

XML (text file) bisa diolah lebih lanjut scr. komputasional

```
<phrase guid="3ce41d83-6ab2-4016-9f7c-0880e73c02d7">
  <item type="txt" lang="eno-ID-x-Kahler-1939">kabia·e'ana·ea'ahã:ũãmãhã·ukuo·hëmõ'õ·kanĩũ·
  epũkã·hëmõ'õ·edixoo·ba'ubuoda'a·ukaka·baka:ùhũa. </item>
  <item type="segnum" lang="en">1.4</item>
  <words>
    <word guid="88d86541-ce59-4fc3-95b9-4ded2eeb6a8a">
      <item type="txt" lang="eno-ID-x-Kahler-1939">kabia</item>
      <morphemes>
        <morph type="prefix" guid="d7f713db-e8cf-11d3-9764-00c04f186933">
          <item type="txt" lang="eno-ID-x-Kahler-1939">ka-</item>
          <item type="cf" lang="eno-ID-x-Kahler-1939">ka-</item>
          <item type="hn" lang="eno-ID-x-Kahler-1939">1</item>
          <item type="gls" lang="en">3SUBJ</item>
          <item type="msa" lang="en">v:(VAgrPrfx)</item>
        </morph>
        <morph type="prefix" guid="d7f713db-e8cf-11d3-9764-00c04f186933">
          <item type="txt" lang="eno-ID-x-Kahler-1939">b-</item>
          <item type="cf" lang="eno-ID-x-Kahler-1939">bu-</item>
          <item type="gls" lang="en">INF</item>
          <item type="msa" lang="en">v:(VPrefopt)</item>
        </morph>
      </morphemes>
    </word>
  </words>
</phrase>
```

Menampilkan data kata dengan morfem/awalan tertentu untuk kajian morfologi, etc. dalam format *spreadsheet* untuk dianalisis lebih lanjut (qual. & quant.)

interoperability (FLEX -> XML text file + coding -> spreadsheet/table) & reusable



	A	B	D	E	F	G
1	ID	word	lex_entry	morph_gloss_en	TYPE_gr	OPPOS_gr
5	4	yapanēēhā	i-__pa-__nee__=ha	3SUBJ__RECIP__close/near__=EMPH	mid	non
6	5	yapanāū	i-__a-__pa-__nau	3SUBJ__MOTION__RECIP__speak	mid	non
7	6	yapanākōō	i-__a-__panakoo	3SUBJ__MOTION__enough/suitable/matching	mid	non
8	7	yapanākōō	i-__a-__panakoo	3SUBJ__MOTION__enough/suitable/matching	mid	non
9	8	yapanauha	i-__a-__pa-__nau__=ha	3SUBJ__MOTION__RECIP__speak__=EMPH	mid	non
10	9	yapanau	i-__pa-__nau	3SUBJ__RECIP__speak	mid	non
11	10	yapanau	i-__pa-__nau	3SUBJ__RECIP__speak	mid	non
12	11	yapanau	i-__pa-__nau	3SUBJ__RECIP__speak	mid	non

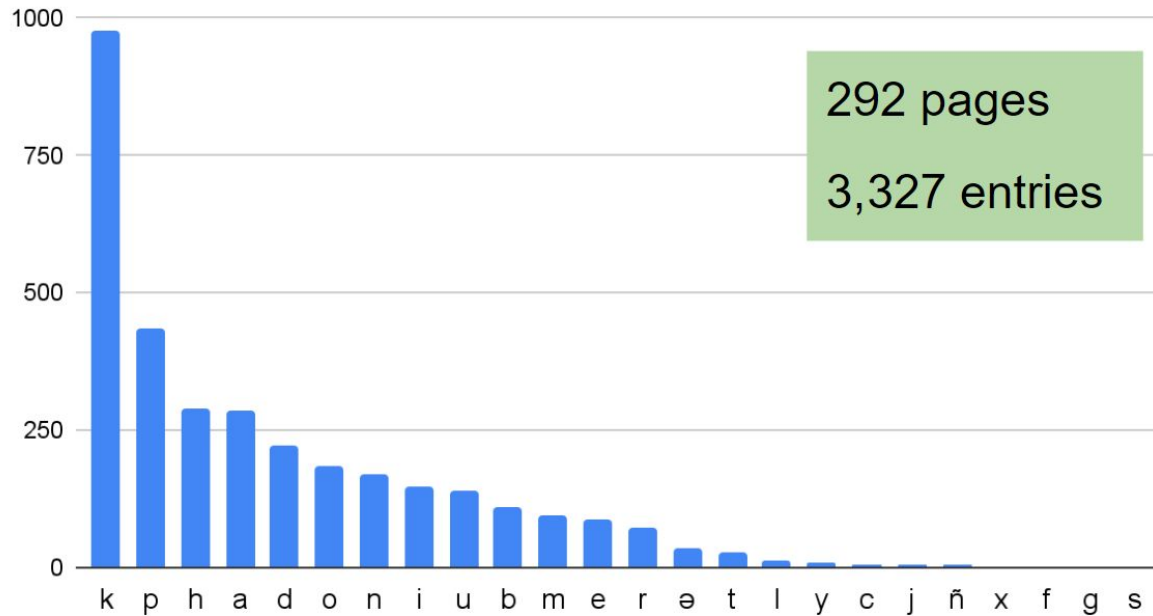
Kamus Enggano - Jerman (cetak -> digitalisasi) (*work-in-progress*)

Rajeg, Gede Primahadi Wijaya; Paramartha, Cokorda Rai Adi; Arka, I Wayan; Dalrymple, Mary (2023). Enggano-German dictionary turns digital: Challenges and opportunities in retro-digitising historical materials of an endangered language. University of Oxford. Presentation.

<https://doi.org/10.25446/oxford.25217018.v1>

Enggano-German (Kähler 1987)

Number of stems/headwords per alphabet (manual count)



Enggano-German (Kähler 1987)



Arts and
Humanities
Research Council



nāpū ^{III} (=II?):	kaʔanāpūāʔā = kanāpūāʔā	tropfen, tröpfeln
	ebō kaʔanāpūāʔā	das Wasser tropft
	hō mānāpūāʔā	tröpfelte bereits
	kīpanānāpū	tropfen lassen
nāpū ^{IV}	kināpū	herausschneiden, -lösen (G68)
	kamānāpū	wird herausgeschnitten
nāpūnū:	kanāpūnū	verfault, morsch
°nasi	(< ML nasi)	gekochter Reis
nāū	> panāū	
nāʔūmānā		morgen (§ 29b)
enāʔūō	(arch,G42) = edopo	Land, Erde

(Kähler 1987: 200)

tē	(DIA) = tō, dō	Art und Weise (§ 31a)
kitē	= kidō	sein wie
tebe:	itebe = iēbe (DIA)	an der Oberseite, oben (§ 26b)
teʔei:	kaʔiteʔei	jmd schlechte Speisen anbieten
teo	kiteo ekital	überreden
təhəda	(DIA) = dəhəda	
tixoi	= didixoi, didiki: kitixoi	etwas umwickeln
tikī	kahatikī	(mit etwas Biegsamem) schlagen
°etirī	(< ML sirih)	Betelbissen
tō	= dō, tē (DIA)	wie (beim Vergleich; § 31a)
	kitoiya < kitō eiya	wie das Wesen sein

(Kähler 1987: 276)

First step with OCR

°etaku	(◁ MKB/ML sagu)	Sago	(Kähler 1987: 276)
°takui	(◁ ML sekoi)	Gerste	
tau?uo	= dau?uo:	kitau?uo	einschließen, einsperren
	ekitau?uo		("d Eingeschlossene" =) Witwe(r) während
			der offiziellen Trauerzeit
	ekitau?ua		Einschließensort
°etawaha	(◁ ML sawah)	nasses Reisfeld	
tē	(DIA) = tō, dō	Art und Weise (§ 31a)	
	kitē = kidō	sein wie	
tebe:	itebe = iēbe (DIA)	an der Oberseite, oben (§ 26b)	
te?ei:	ka?ite?ei	jmd schlechte Speisen anbieten	
teo	kiteo ekitai	überreden	
tohōda	(DIA) = dahōda		
tixoi	= didixoi, didiki:	kitixoi	etwas umwickeln
tiki	kahatikī	(mit etwas Biegsamem)	schlagen
°etiri	(◁ ML sirih)	Betelbissen	
tō	= dō, tē (DIA)	wie (beim Vergleich; § 31a)	
	kitoiya ' kitō eiya	wie das Wesen sein	

OCR output via the tesseract R package

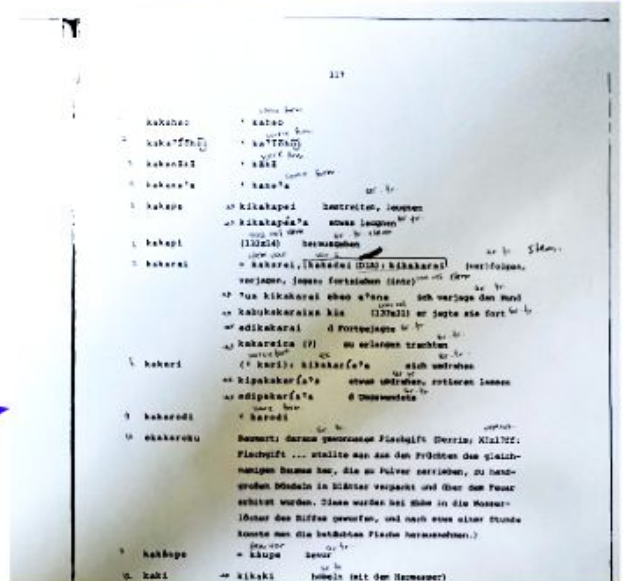
```
9 "etaku (◁ MKB/ML sagu) Sago
10 "takui (◁ ML sekoi) Gerste
11 7 7 * " a iJ 7
12 tau ?uo = dau ?uo: kitau 2u0 einschließen, einsperren
13 Ne
14 ekitau?uo ("d Eingeschlossene" =) Witwe(r)) w&hrend
15 der offiziellen Trauerzeit
16 Pe
17 ekitay Pua Einschließensort
18 "etawaha (◁ ML sawah) nasses Reisfeld
19 te (DIA) = to, dd Art und Weise (§ 31a)
20 kite = kido sein wie
21 tebe: itebe = idebe (DIA) an der Oberseite, oben (§ 26b)
22 te?ei: ka?ite?ei jmd schlechte Speisen anbieten
23 teo kiteo ekitai tiberreden
24 tohōda (DIA) = dshōda
25 tixoi = didixoi, didiki: kitixoi etwas umwickeln
26 tiki kahatikī (mit etwas Biegsamem) schlagen
27 "etiri (◁ ML sirih) Betelbissen
28 to = do, té (DIA) wie (beim Vergleich; § 31a)
29 kitoiya *' kits eiya wie das Wesen sein
```

Reverse-Engineering

Deconstructing and extracting (manually!) the individual components of an entry in the dictionary

104

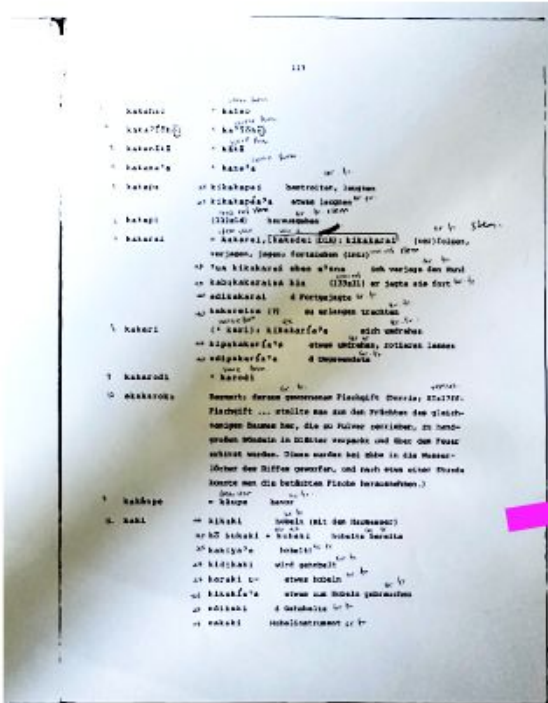
-ka?a	unsere (in), von mir und euch: Possessivsuffix der 1.Pl in (§ 11)
eubāka?a	unsere (in) Haus
eka?a?aho	= eka?aho, eka?a?oho Baumart (ML kasai, Pometia pinnata Forst.)
Ka?a?ĪI	Ort auf Enggano
ka?abaha	(ka + abaha) und nachdem (Konjunktion)
ka?abu	kika?abu vergessen
	hō buka?abu vergaß bereits
	kika?abui jmd, etwas vergessen
	dibuka?abui sie vergaßen etwas
	eka?abua Vergessensort
	eaka?abu = eka?abu das Vergessene
	eka?abuīa?a d Vergessene
	eka?abuīa?au edobu mein Vergessenes waren die Sachen
	eka?abuīyo d zu Vergessende



Manual annotation/identification of the entry components

Each page was checked for accuracy

Reverse-Engineering



KAMUS ENGGANO

Menu

Users

Input Data Kamus

Questionnaire

Logout

Manajemen Stem

Kamus Bahasa Enggano

Menu / Manajemen Stem

Data Stem

Tambah Stem

10 entries per page

Search...

Alphabet	Page	Stem	Stem Form	Create by	Create Date	Opsi
K	157	1	ekoba?a	Putu Dea Indah Kartini	2023-05-11 02:38:30	✎ 🔍 🗑
K	157	2	ekobaruku	Putu Dea Indah Kartini	2023-05-11 02:40:47	✎ 🔍 🗑
K	157	3	ekobituu	Putu Dea Indah Kartini	2023-05-11 02:43:39	✎ 🔍 🗑
K	157	4	kobu	Putu Dea Indah Kartini	2023-05-11 02:45:49	✎ 🔍 🗑
K	157	5	kode	Putu Dea Indah Kartini	2023-05-11 02:58:06	✎ 🔍 🗑
K	157	6	ekodekode uaparaa	Putu Dea Indah Kartini	2023-05-11 02:59:31	✎ 🔍 🗑
K	157	7	kodi	Putu Dea Indah Kartini	2023-05-11 03:28:23	✎ 🔍 🗑
K	157	8	kadio	Putu Dea Indah Kartini	2023-05-11 03:31:40	✎ 🔍 🗑

Typing the components into the relevant slot in the entry database

Author/ <i>Penulis</i>	Stay in Enggano/ <i>Periode tinggal di Enggano</i>	Publication/ <i>Publikasi</i>	Number of Enggano words/ <i>Jumlah kata Enggano</i>
Djoeragan Boewang	several times between 1840 and 1855, including a visit of one year	Verslag orntrent het Eiland Enggano [Report on Enggano Island/ <i>Laporan di Pulau Enggano</i>], <i>Tijdschrift voor Indische Taal-, Land- en Volkenkunde</i> , 1854	22
D. J. Brouwer	early 1850s	Woordenlijst van het weinige door de Inlanders op Engano, te ... Barhao ons medegedeelde, der Enganeesche taal vertolkt door den Chinees Tsi-ting aan den Lieut(enant) D.J.S. Brouwer [Glossary of several words of the Enganeese language communicated to us by the inhabitants of Engano in Barhao and communicated by the Chinese [interpreter] Tsi-ting to the lieutenant D.J.S. Brouwer]. The list is referenced as item Or. 3386U in the Inventory of the Oriental Manuscripts of the Library of the University of Leiden .	104
Carl Benjamin Hermann Baron von Rosenberg	10-24 September 1852	Beschrijving van Enggano en van deszelfs bewoners [Description of Enggano and its inhabitants/ <i>Deskripsi Enggano dan penduduknya</i>], <i>Tijdschrift voor Indische Taal-, Land- en Volkenkunde</i> , 1855	154
Johannes van der Straaten and Pieter Severijn	10 June-2 July 1854	Verslag van een in 1854 bewerkstelligd onderzoek op het eiland Enggano [Report of an investigation carried out in 1854 on the island of Enggano/ <i>Laporan penyelidikan yang dilakukan pada tahun 1854 di pulau Enggano</i>], <i>Tijdschrift voor Indische Taal-, Land- en Volkenkunde</i> , 1855	201
J. Walland	6-20 May 1863	Het Eiland Enggano [Enggano Island/Pulau Enggano], <i>Tijdschrift voor Indische Taal-, Land- en Volkenkunde</i> , 1864	250
R. Francis	1865/66 and 1868/70	Enganeesche woordenlijst [Enggano wordlist/ <i>Daftar kata Enggano</i>], <i>Notulen der Bataviaasch Genootschap</i> , 1874 and 1877	91
Jean Abraham Chrétien Oudemans	-	Woordenlijst van de talen van Enggano, Mentawai en Nias [Vocabulary of the languages of Enggano, Mentawai, and Nias/ <i>Kosakata bahasa Enggano, Mentawai, dan Nias</i>], <i>Tijdschrift voor Indische Taal-, Land- en Volkenkunde</i> , 1879	150
A. C. Oudemans	-	Engano (bewesten Sumatra), Zijne Geschiedenis, Bewoners en Voortbrengselen [Engano (west of Sumatra), Its History, Inhabitants and Products/ <i>Engano (Sumatera barat), Sejarah, Penduduk dan Produknya</i>], <i>Tijdschrift van het Koninklijk Nederlandsch Aardrijkskundig Genootschap</i> , 1889. This seems to be a reproduction of the Francis lists.	160

Brouwer (c. early 1850)

(Thanks to Daniel Krauß for deciphering this cryptic yet valuable hand-writing)

Woordenlijst van het weinige
 door de Inlanders op Engano, te Barhao
 ons medegedeelde der Enganeesche taal
 Soektoek door den Chinese Tji-bong
 van den Lande D. J. Brouwer

Soektoek medegedeeld door de Inland, de juiste uitpraak
 ongekend.

Weg	Tob	Luiven	Lechoc
Maan	Mona	Papayaa	Kata
nacht	hohob-lobob	Sarkas	abi-abi
Zon	Rahad	Kasse	wafob
Herren	apulhoa	muis	to-ab
Egen	Suwob	musset	Taka
Strom	Kiee	man	amama
Wunder	Kahoc	Wrauw	amini
Walden	Kahoko-Mona	Kiee	awobob
Wardel	Lopob	Jader	amama
Wasser	hohob-lobob	Lustee	Kahomun
Seur	abi-abi	Wroder	Johé-awobob
Wuer	Bilob		honni
Berg	Kahon	Shuru	Papohida
Eiland	hohob	Kuste	Sela

Schip	Apela-ben-Seari
Leem	oedje
Teus	paauw
Kampong	paaba
Boeg	Kahoa
Frucht	Koero-Kahoa
Woods	Kahoko
Kist	kanwohoe-ambobo
Kanon	Konta-Kahoa
Teus	Kawé-Kawé
Frucht	mba-orka
Wierandvoet	oko
Soektoek	hannob
Olis	hann
Wier	awobob
Wier	Kahoc
Wier	oko
Wier	Sela
Wier	Wiekona
Wier	alima-awobob
Wier	apay-afab
Wier	alima-afab
Wier	Kahoc

104 items

Von Rosenberg (1855)

154 items

22

Woordenlijst.

Staan	Jaramatrego	Duiven	mi-komabeun
Gaan	^{Edie} Edie	stelen	paiveta
Loopen	Jarapoerpoera	leugen	pauro-puro
Zitten	ellaké-god	kaeyen	achyie
Komen	Tamarsoij	neekten	passa-aea
Blijven	Dunacklae	binden	paani-i
Geven	Pahè	twilen	gersello-mello
opnen	Peparich-i	zamenmen	poeread lewa
ontrenten	pemanabo	danen	lewa
draagen	pagekli	vischen	kakaleka
missen	popasij	maekten	fatjed-wau
Zien	popasij	huizen	nakafe
maeken	pepovele-i-tyok-i	uolgeu	kakabapij
Zaekken		namen	nej
spreeken	dipaganan	wit	gashijka
eten	Faginono	rood	so-dajee
drincken	fajino-wat	blauw	hi-loes
Koken	lewa	erwt	kakiaki
	poeread-kowi	zeel	himono
			kanoe

Woordenlijst.

Jaramatrego.
 Edie-édie.
 Jarapoerpoera.
 Matrego.
 Tawawej.
 Panoekoe.
 Pahè.
 Pemanabo.
 Pàgehlie.
 Popo-ej.

Francis (1870, in Oudemans 1889)

99 items

NEDERLANDSCH.	FRANCIS.	VAN DER STRAATEN EN SEVERIJN.	VON ROSENBERG.	BOEWANG.
zwemmen	kaäke	pakiea	poeroe lewo lewo	kahai
1 = 1	kahaie	egaij	daheij, dahei	adoea
2 = 2	adoea	adoloe	adoloe	akoloe
3 = 3	akoloe	koloe	agoloe	ampapa
4 = 4	afa	afa	aopa	aliema
5 = 5	aliema	lieba	aliema	akie akiema
6 = 6	kaikiene	akeno	akiaknio, akiakia	aliema adoe, aliemei a-
7 = 5 + 2	aliema é adoea	lieba doea	adoloe	aliema adoea
8 = 4 + 4 of 5 + 3	afa é afa	afa afa	aliema agoloe, aliemei	apa joepa
9 = 5 + 4	aliema é afa	lieba afa	agoloe	
10 = 10 of 4 + 4 + 2	kahafoeloe	tapoeloe	aliema aopa, aliemei	
11 = 10 + 1	kahafoeloe é kahaie		aopa	apa apa adoea
12 = 10 + 2	kahafoeloe é adoea		tahapoeloe	
13 = 10 + 3	kahafoeloe é akoloe			
14 = 10 + 4	kahafoeloe é afa		tahapoeloe alima	
15 = 10 + 5	kahafoeloe é aliema		kahei taka	
20 = 1. 20	kahaie takka		kahei taka tahapoeloe	
30 = 1. 20 + 10	kahaie takka é kahafoeloe		adoeei taka	
40 = 2. 20	adoea takka		adoeei taka tahapoeloe	
50 = 2. 20 + 10	adoea takka é kahafoeloe		akoloe taka	
60 = 3. 20	akoloe takka		akoloe taka tahapoeloe	
70 = 3. 20 + 10	akoloe takka é kahafoeloe		aopeia taka	
80 = 4. 20	afa takka		aopeia taka tahapoeloe	
90 = 4. 20 + 10	afa takka é kahafoeloe		alimeiei taka	
100 = 5. 20	aliema takka			adoea taka tapoeloe (50)
200 = 10. 20	kahafoeloe takka			aliema takka (100)

Comparative
list with the
earlier
Enggano
word lists

EnoLEX: Bank Data Leksikal Diakronis bahasa Enggano (open access)

Main Concept Search Global Search Sources Links ▾

EnoLEX: A diachronic lexical database for the Enggano language



Arts and
Humanities
Research Council

This research is funded by the Arts and Humanities Research Council (AHRC) Grant ID [AH/S011064/1](#) and [AH/W007290/1](#).

Overview

EnoLEX collates lexical data from [legacy materials and contemporary fieldwork data](#) about the Enggano language, ranging from simple/short and extensive word lists, anthropological and ethnographic writings, a dictionary, thesis, and contemporary

<https://enggano.shinyapps.io/enolex/>

How to cite EnoLEX

Krauße, Daniel, Gede Primahadi Wijaya Rajeg, Cokorda Pramatha, Erik Zoebel, Charlotte Hemmings, I Wayan Arka, Mary Dalrymple (2024). *EnoLEX: A Diachronic Lexical Database for the Enggano Language*. Available online at <https://enggano.shinyapps.io/enolex/>

Rajeg, Gede Primahadi Wijaya, Daniel Krauße, and Cokorda Rai Adi Pramatha (2024). [EnoLEX: A Diachronic Lexical Database for the Enggano language](#). In *Proceedings of AsiaLex 2024 (The Asian Association for Lexicography 2024 Hybrid Conference)*. Toyo University, Tokyo: Japan.



Kamus berbasis korpus Kontemporer Bahasa Enggano

The screenshot displays the FieldWorks Language Explorer interface for the Enggano-3-gloss-fixes project. The main window shows a list of dictionary entries under the heading "Main Dictionary Entries". The entries are as follows:


- abaha** v *Ind* pergi *Eng* go *Kahtnu abaha tk* *Dulu kalau kita mau pergi In the past if we went (ah,)*
- abaha** v *Ind* kalau bangun *Eng* when get up *Na'an abaha ki, kamuno* *Nanti, kalau dia bangun, dia akan makan lagi! Later, when it awakens, it will eat again! Yu'ueh !ŋe', na'an abaha ke, kabayo'ode ik* *Ia tidur di sini; nanti kalau bangun, ia akan mengikuti kita! It sleeps here; when it wakes, then it will follow us! (ahar,)*
- abaha** v *Ind* minum *Eng* drink *Seringkan abaha tk me' kopt ehear kan... de ka musang kin'e'ah kan* *Seringkan kalau minum kita kopi disitukan seperti kotoran musang begitukan We often drink coffee there like civet's poop, right? (It,)*
- aba kaha'** *num Ind sembilan Eng nine*
- aba'ia** v *Ind* keluar *Eng* go out *Kapakom na' aba'ia kur tekora* *Nanti kita bertemu keluar dari sekolah We'll meet later coming out of school (a'ia)*
- aba'par** v *Ind* jadi *Eng* become *Aba'par ke* *Dia jadi seperti It becomes like (par)*
- abapih** v *Ind* salah *Eng* wrong *Maha abapih be u e' ke'par yur ya'hao* *Kalau saya salah bercerita karena saya ini bukan kepala suku Maha akun art kanap e' maha abapih* *Entah benar atau salah U ean u e' abapih maha kin'en* *Saya itu saya ini kalau salah atau bagaimana I, I, if I made a mistake or what (apih)*
- abari'ie** v *Ind* buat *Eng* make *Abari'ie ki kab ean* *Kalau mau dia membuat jaring itu If they wanted to make a (fishing) net (pari')*
- aba'u** v *Ind* bagus *Eng* good *Jadi ap'hta' aba'u* *Jadi kalau bisa bagus So if (we) can, we do it well (a'u)*
- abe'** v *Ind* berdiri *Eng* stand *Abe' hat* *Berdirilah! Stand!*
- abe'a** v *berdiri stand*
- kababe'** v *berdiri stand*
- kabe'** v *berdiri stand*

The interface includes a menu bar (File, Send/Receive, Edit, View, Data, Insert, Format, Tools, Parser, Window, Help), a toolbar with various icons, and a left sidebar with navigation options like Lexicon, Texts & Words, Grammar, Notebook, and Lists. The status bar at the bottom shows the date (26/Nov/2023 24/Aug/2024) and a message: "Queue: (-/-) No Parser Loaded".

Ringkasan

- Enggano bahasa daerah yang terancam
- Hanya lebih dikuasai kalangan tetua dibandingkan anak-anak
- Dokumentasi Kontemporer
 - Revitalisasi
 - Pembangunan korpus data untuk: bahan ajar + kamus kontemporer
- Dokumentasi Historis atas *legacy material*
 - Di bahasa yang bukan Indonesia (Jerman + Belanda)
 - Topik penelitian tersendiri
 - Sumber data aksesibel dalam bentuk cetak dan PDF
 - Digitalisasi menjadi computer-readable
 - Database leksikal diakronis
 - Korpus digital bahasa Enggano Mula (*Old Enggano*)
- Data Korpus (transkripsi di ELAN + FLEEx) akan dibagikan terbuka

Butir pembahasan

- ~~1. Konsep *korpus*~~
- ~~2. Target Bahasa~~
- ~~3. Sumber (data & daya)~~
- ~~4. Aksesibilitas + Infrastruktur~~
- ~~5. Contoh kasus:~~
 - ~~a. Bahasa Enggano (sumber data + daya, aksesibilitas)~~
 - b. Bahasa Bali (sumber data + daya; model gerakan komunitas; infrastruktur awal) 
 - c. Bahasa-bahasa di Australia (LDaCA) (sumber data + daya + infrastruktur nasional)

Bahasa Bali

- <https://basabali.org/>
- Komunitas Bahasa Bali Wiki
 - Wikitionary Bahasa Bali (<https://en.wiktionary.org/wiki/Bali>)
 - Bahasa Bali Wiki (kamus daring dengan konsep crowdsourcing)
- Penyuluh bahasa Bali
 - Konservasi dan digitalisasi lontar
 - Transliterasi lontar

BasaBali Wiki Virtual Dictionary - Open Access

BASAbali Wiki
BASAibu Wiki

English ▾ Create account Log in

Search Basa Bali ▾ >> English ▾ 🔍

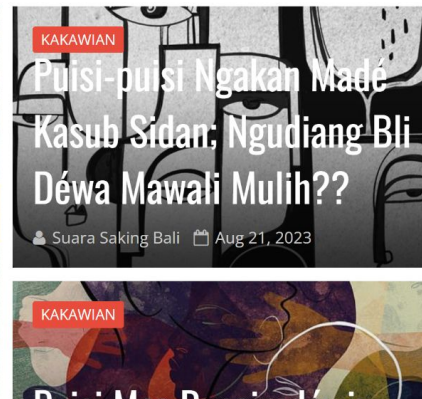
#sampahupacara #sampahsisabanten #sampahbanten

How can we reduce waste in religious activities?










Respond! How to participate? See wikithon entries

Check out Past Wikithons Suggest a Question Use the Dictionary Find out about BASAbali Wiki and BASAibu Community Spaces









Sumber Data Daring berupa Karya Sastra dll. *Suara Saking Bali* - Open Access



Mengumpulkan manual mandiri :(

Name	Date modified	Type	Size
 .Rproj.user	24/10/2017 17:12	File folder	
 2006	28/05/2015 03:00	File folder	
 2007	28/05/2015 03:01	File folder	
 2008	28/05/2015 03:01	File folder	
 2009	28/05/2015 03:01	File folder	
 2010	28/05/2015 03:01	File folder	
 2011	28/05/2015 03:01	File folder	
 2012	28/05/2015 03:01	File folder	
 2013	28/05/2015 03:01	File folder	

Mengumpulkan manual mandiri :(

Name	Date modified	Type	Size
 AGUST, 7 2011-Kasus KDRT ring Buléléng...	31/12/2011 10:00	TXT File	4 KB
 AGUST, 7 2011-Patut Yatna Makarya Prog...	07/12/2011 08:25	TXT File	3 KB
 AGUST, 14 2011-AWIG-AWIG	07/12/2011 08:28	TXT File	3 KB
 AGUST, 14 2011-Kautaman Miwah Kaagu...	07/12/2011 08:29	TXT File	5 KB
 AGUST, 14 2011-Lascarya Dasar Kalepasan	07/12/2011 08:31	TXT File	2 KB
 AGUST, 14 2011-Lumu-lumut Watulumba...	07/12/2011 08:31	TXT File	3 KB
 AGUST, 14 2011-Moksa	07/12/2011 08:32	TXT File	3 KB
 AGUST, 14 2011-Nyabran Rahina Bali Say...	07/12/2011 08:33	TXT File	3 KB

SCOPIC Project on Balinese (I Wayan Arka & Desak Eka Pratiwi)



PARADISEC Catalog

Gede Primahadi Wijaya Rajeg | Sign out

Home Dashboard Collections **Items** Contact

Return To Results Next item

Item details

Item ID	SocCog-ban06	(Collection Details)
Title	Balinese, Gianyar 3	
Description	Irmayanti, Winda, Slamin speaking in Balinese Gianyar dialect	
Origination date	2016-09-16	
Origination date free form		
Archive link	https://catalog.paradisec.org.au/repository/SocCog/ban06	
URL		
Collector	Desak Putu Eka Pratiwi	Find similar
Countries	Indonesia - ID <i>To view related information on a country, click its name</i>	

Content Files (10)

View file contents

Filename ▲▼	Type ▲▼	File size ▲▼	Duration ▲▼	File access
SocCog-ban06-gianyar3_task_1.eaf	application/eaf+xml	301 KB		View
SocCog-ban06-gianyar3_task_1.mp4	video/mp4	567 MB	00:08:54.861	View
SocCog-ban06-gianyar3_task_1.mxf	application/mxf	11.7 GB		View
SocCog-ban06-gianyar3_task_2.eaf	application/eaf+xml	197 KB		View
SocCog-ban06-gianyar3_task_3.eaf	application/eaf+xml	210 KB		View
SocCog-ban06-gianyar3_task_3.mp4	video/mp4	259 MB	00:04:21.461	View
SocCog-ban06-gianyar3_task_3.mxf	application/mxf	5.85 GB		View

- Berbasis stimuli gambar
- Fieldwork-based
- Transkripsi ELAn dibagikan terbuka -> ekspor plain text

SCOPIC Project on Balinese (I Wayan Arka & Desak Eka Pratiwi)

KWIC_results .XLSX

File Edit View Insert Format Data Tools Help Accessibility

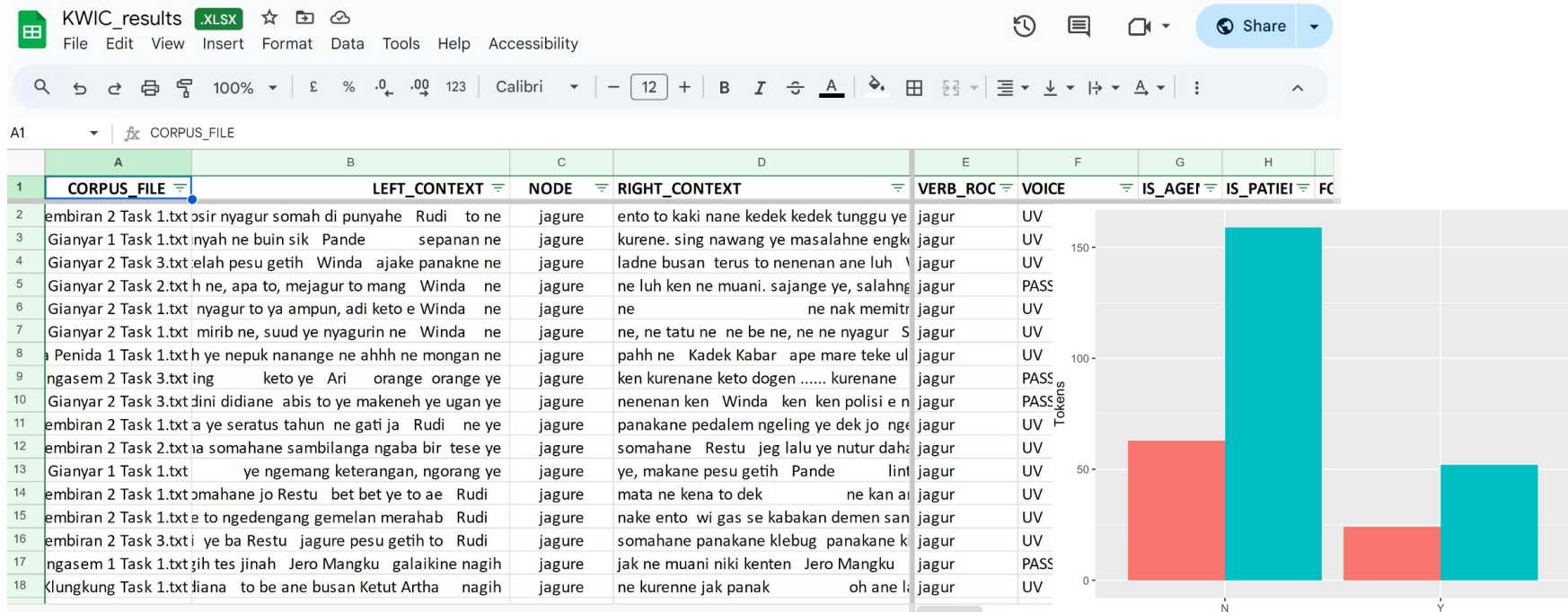
100% Calibri 12

A1 CORPUS_FILE

	A	B	C	D	E	F	G	H	I
1	CORPUS_FILE	LEFT_CONTEXT	NODE	RIGHT_CONTEXT	VERB_ROC	VOICE	IS_AGEI	IS_PATIEI	FC
2	embiran 2 Task 1.txt	osir nyagur somah di punyahe Rudi to ne	jagure	ento to kaki nane kedek kedek tunggu ye	jagur	UV	N	Y	ag
3	Gianyar 1 Task 1.txt	nyah ne buin sik Pande sepanan ne	jagure	kurene. sing nawang ye masalahne engki	jagur	UV	N	Y	ag
4	Gianyar 2 Task 3.txt	elah pesu getih Winda ajake panakne ne	jagure	ladne busan terus to nenenan ane luh	jagur	UV	N	Y	ag
5	Gianyar 2 Task 2.txt	h ne, apa to, mejagur to mang Winda ne	jagure	ne luh ken ne muani. sajange ye, salahng	jagur	PASS	Y	Y	ag
6	Gianyar 2 Task 1.txt	nyagur to ya ampun, adi keto e Winda ne	jagure	ne ne nak memitr	jagur	UV	N	Y	ag
7	Gianyar 2 Task 1.txt	mirib ne, suud ye nyagurin ne Winda ne	jagure	ne, ne tatu ne ne be ne, ne ne nyagur S	jagur	UV	N	Y	ag
8	Penida 1 Task 1.txt	h ye nepuk nanange ne ahhh ne mongan ne	jagure	pahh ne Kadek Kabar ape mare teke ul	jagur	UV	N	Y	ag
9	ngasem 2 Task 3.txt	ing keto ye Ari orange orange ye	jagure	ken kurenane keto dogen kurenane	jagur	PASS	Y	Y	ag
10	Gianyar 2 Task 3.txt	dini didiane abis to ye makeneh ye ugan ye	jagure	nenenan ken Winda ken ken polisi e n	jagur	PASS	Y	Y	ag
11	embiran 2 Task 1.txt	a ye seratus tahun ne gati ja Rudi ne ye	jagure	panakane pedalem ngeling ye dek jo nge	jagur	UV	N	Y	ag
12	embiran 2 Task 2.txt	ia somahane sambilanga ngaba bir tese ye	jagure	somahane Restu jeg lalu ye nutur dahan	jagur	UV	N	Y	ag
13	Gianyar 1 Task 1.txt	ye ngemang keterangan, ngorang ye	jagure	ye, makane pesu getih Pande lint	jagur	UV	N	Y	ag
14	embiran 2 Task 1.txt	omahane jo Restu bet bet ye to ae Rudi	jagure	mata ne kena to dek ne kan a	jagur	UV	N	Y	ag
15	embiran 2 Task 1.txt	e to ngedegang gemelan merahab Rudi	jagure	nake ento wi gas se kabakan demen san	jagur	UV	N	Y	ag
16	embiran 2 Task 3.txt	i ye ba Restu jagure pesu getih to Rudi	jagure	somahane panakane klebug panakane k	jagur	UV	N	Y	ag
17	ngasem 1 Task 1.txt	jih tes jinah Jero Mangku galaikine nagih	jagure	jak ne muani niki kenten Jero Mangku	jagur	PASS	Y	N	ag
18	Klungkung Task 1.txt	hiana to be ane busan Ketut Artha nagih	jagure	ne kurenne jak panak oh ane la	jagur	UV	N	Y	ag


FAIR data principle of *regional corpus*: **interoperable** (bisa diunduh, dibuka di AntConc -> analisis di MS Excel), findable, accessible, reusable

SCOPIC Project on Balinese (I Wayan Arka & Desak Eka Pratiwi)



FAIR data principle of *regional corpus*: **interoperable** (bisa diunduh, dibuka di AntConc -> analisis di MS Excel), findable, accessible, reusable

Butir pembahasan

- ~~1. Konsep *korpus*~~
- ~~2. Target Bahasa~~
- ~~3. Sumber (data & daya)~~
- ~~4. Aksesibilitas + Infrastruktur~~
- ~~5. Contoh kasus:~~
 - ~~a. Bahasa Enggano (sumber data + daya, aksesibilitas)~~
 - ~~b. Bahasa Bali (sumber data + daya; model gerakan komunitas; infrastruktur awal)~~
 - c. Bahasa-bahasa di Australia (LDaCA) (sumber data + daya + infrastruktur nasional) 

LDaCA (Language Data Commons of Australia)



Making nationally significant language data available for academic and non-academic use
Providing a model for ensuring continued access with appropriate community control

<https://www.ldaca.edu.au/>

<https://ardc.edu.au/project/language-data-commons-of-australia/>

LDaCA (Language Data Commons of Australia)

The screenshot shows the LDaCA website interface. At the top, the browser address bar displays `data.ldaca.edu.au/search`. The website header includes the LDaCA logo, navigation links for 'Home', 'Collections', 'Notebooks', 'Browse' (highlighted), 'Login', and 'Help'. On the left side, there is a search bar with the placeholder text 'Search...' and a magnifying glass icon. Below the search bar is a button labeled 'Advanced Search beta'. Underneath is a 'Filters' section. At the bottom left, there is a dropdown menu labeled 'Main Collections'. The main content area features a large heading 'International Corpus of English (ICE-AUS)' in blue. Below this heading, the text reads: 'Type: Dataset RepositoryCollection', 'Language: English', and 'The Australian component of the International Corpus of English (ICE-AUS) is an approximately one million word corpus of transcribed spoken and written Australian English from 1992-1996. It consists of 500 samples of...'. Further down, it states 'Collections: 12, Objects: 558, Files: 558' and includes a 'See more' link. On the right side of the main content area, there are several icons: a book, a microphone, a pencil, and a document. At the bottom right of the screenshot, the URL `https://data.ldaca.edu.au/search` is displayed.

Sejumlah korpus dapat diunduh dalam format plain-text (FAIR)

LDaCA (Language Data Commons of Australia)


A NEW JAPAN

That's the only certainty. It is high irony that what seems to have saved the corruption-soiled Liberal Democratic Party from outright defeat in Sunday's Japanese general election is the collapse of the socialist vote. While in these post-Cold War times the drain of support from the Left is understandable, and perhaps inevitable, it is a fact that had the socialist vote held, the anti-LDP coalition would have won. Prime Minister Kiichi Miyazawa said at his post-election news conference yesterday: The LDP has won as the leading No. 1 party, so it has the duty to continue the nation's government. Mr Miyazawa is unlikely to survive as Prime Minister beyond the holding of a special parliamentary session to be called within a month. What his country undoubtedly faces is a period of some uncertainty at least so far as formal government is concerned and most likely rampant and unseemly horse-trading for power. Sunday's result is not the worst that might have occurred. But it goes close to it. Japan is likely to have even after next month's parliamentary meeting its weakest government in decades.

<https://data.ldaca.edu.au/search>

ICE corpus Australia yang dapat dibuka di notepad -> AntConc, etc.

Butir pembahasan

- ~~1. Konsep *korpus*~~
- ~~2. Target Bahasa~~
- ~~3. Sumber (data & daya)~~
- ~~4. Aksesibilitas + Infrastruktur~~
- ~~5. Contoh kasus:~~
 - ~~a. Bahasa Enggano (sumber data + daya, aksesibilitas)~~
 - ~~b. Bahasa Bali (sumber data + daya; model gerakan komunitas; infrastruktur awal)~~
 - ~~c. Bahasa-bahasa di Australia (LDaGA) (sumber data + daya + infrastruktur nasional)~~ 

Penutup

- Penyusunan korpus bahasa daerah memerlukan pertimbangan terkait keberadaan calon data & aksesibilitas untuk masyarakat umum, dan komunitas yang bahasanya digunakan sebagai korpus
- Bingkai kerja jangka panjang yang bisa berskala nasional multi-institusi ataupun kolaboratif-internasional (mis. LDaCA di Australia, Enggano project)
- FAIR data principle



Penyusunan Korpus Bahasa Daerah

Gede Primahadi Wijaya Rajeg

University of Oxford, UK; Centre for Interdisciplinary Research on the Humanities and Social Sciences (CIRHSS) & *CompLexico* research group, Universitas Udayana

<https://orcid.org/0000-0002-2047-8621>

Konsinyasi Penyjapan Data Korpus Bahasa Daerah dan Pemetaan Bahasa (24 Oktober 2024)

Badan Pengembangan dan Pembinaan Bahasa, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi melalui Pusat Pengembangan dan Pelindungan Bahasa dan Sastra