# 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.

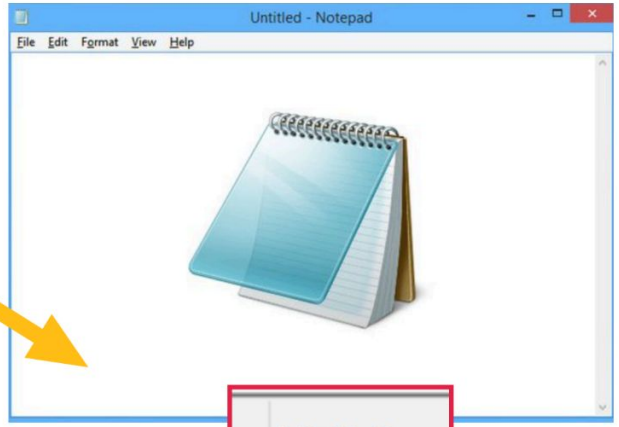

**sekumpulan sampel teks (i) digital yang (ii) bersifat otentik, (iii) representatif dan berimbang, serta (iv) berukuran besar**

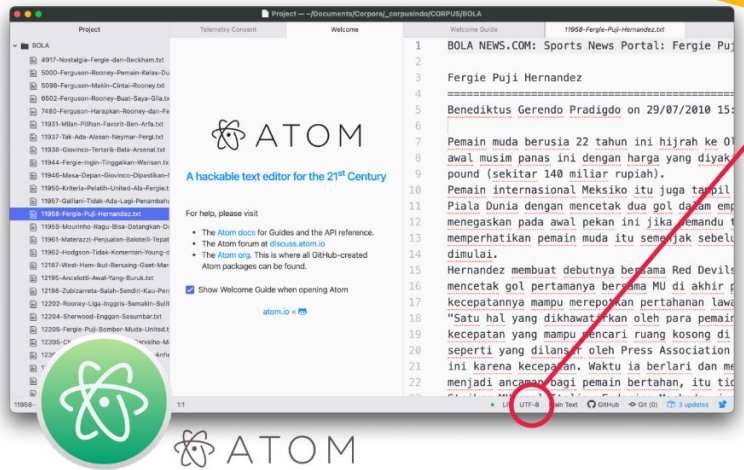
**(Stefanowitsch 2020: 22-28; Gries 2017: 7)**

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.

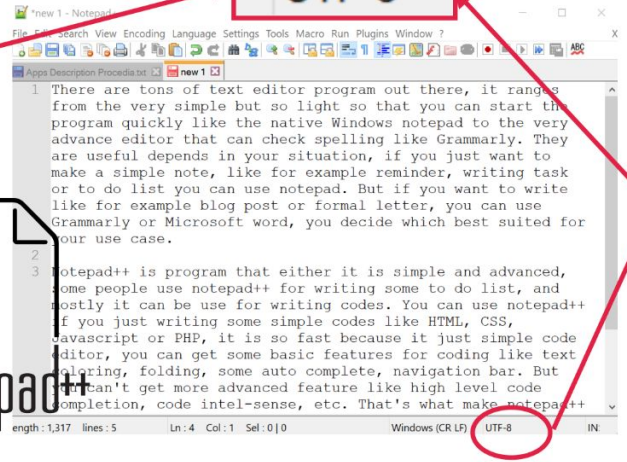
# Teks biasa (*plain text*)

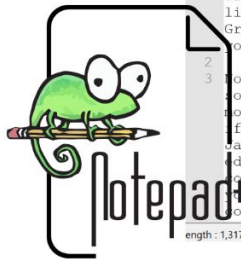
## Unicode UTF-8 encoding


UTF-8




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


**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



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

## 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

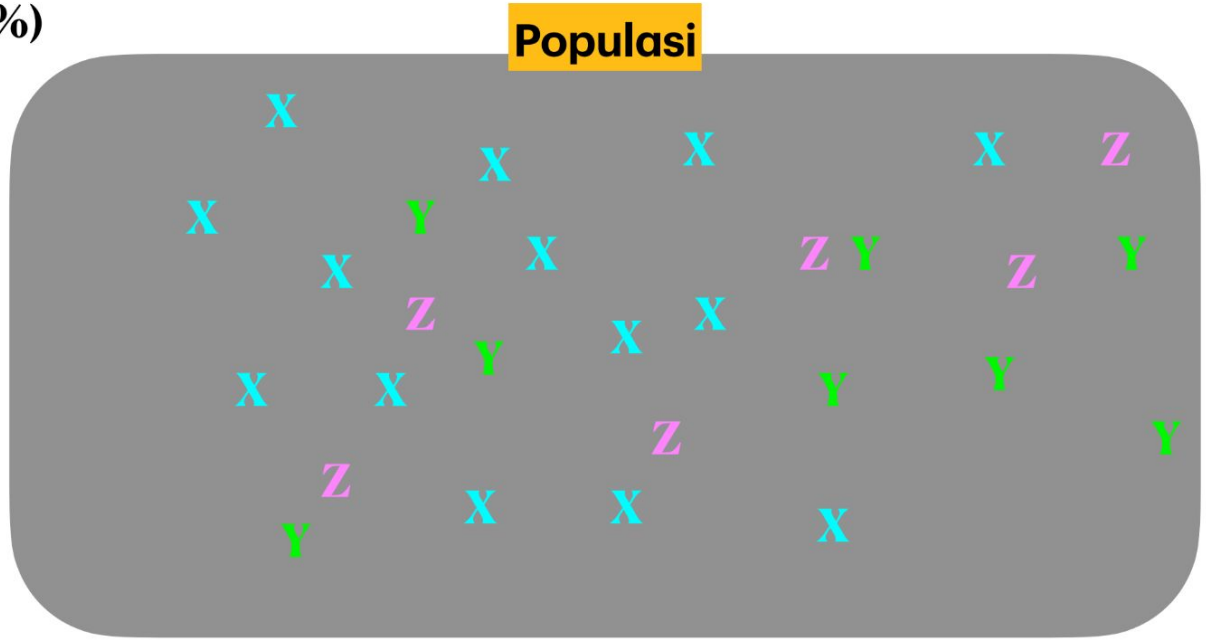
# keterwakilan korpus

**Total = 28 (100%)**

**X = 14 (50%)**

**Y = 8 (28.57%)**

**Z = 6 (21.43%)**




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

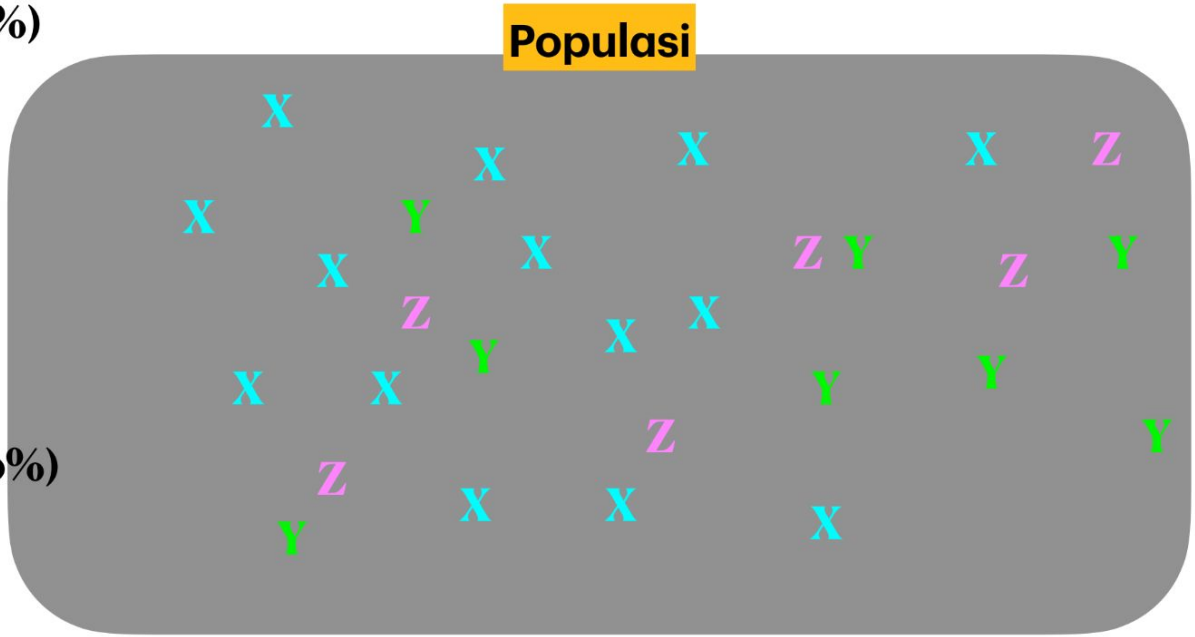
# keterwakilan korpus

**Total = 28 (100%)**

**X = 14 (50%)**

**Y = 8 (28.57%)**

**Z = 6 (21.43%)**




**Sampel = 18 (100%)**

**X = (50%)**

**Y = (28.57%)**

**Z = (21.43%)**

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

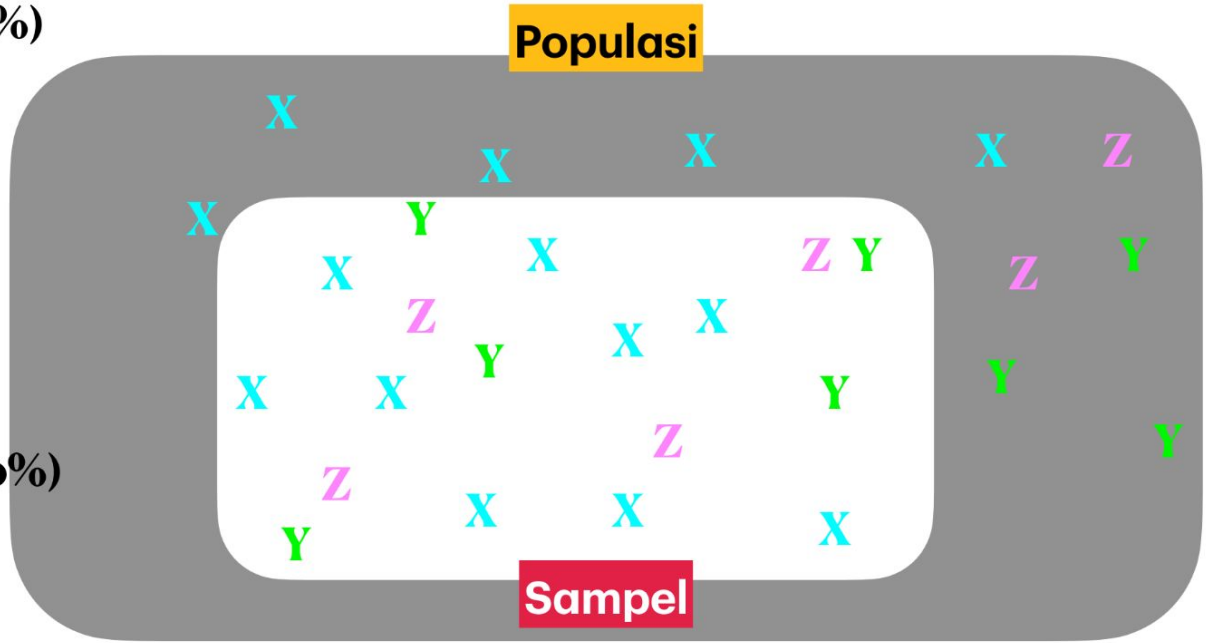
# keterwakilan korpus

**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%)**

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

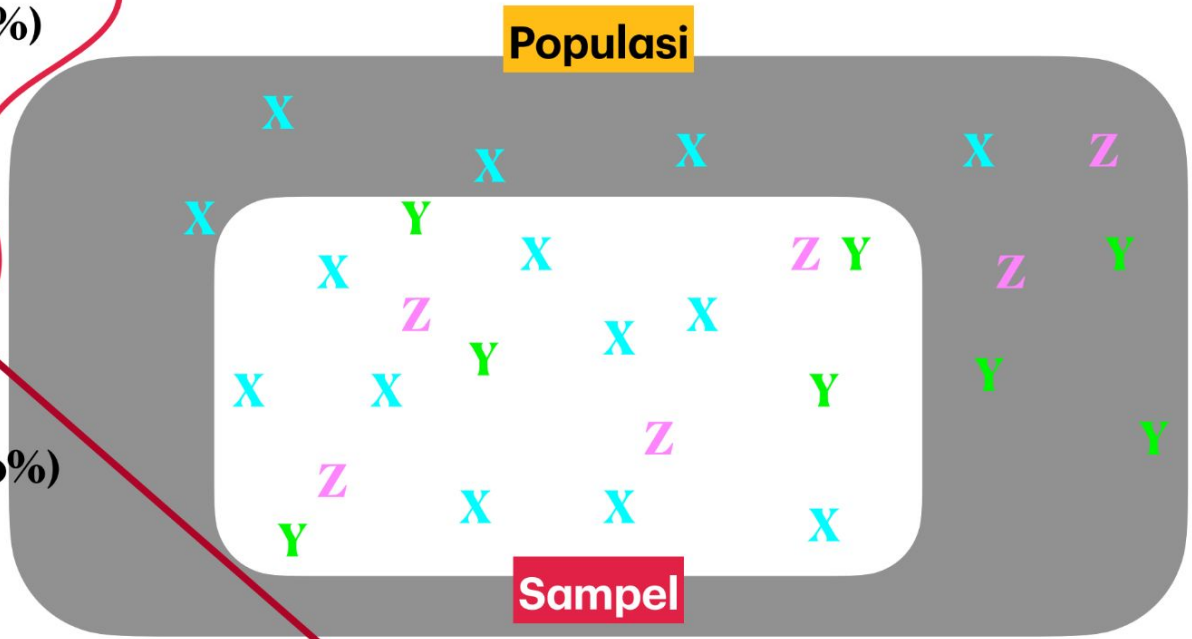
# keterwakilan korpus

**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

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

# 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, suratkabar, akademik, blog, komentar sosmed)
  - Register (fitur bahasa terkait dengan fungsi sosial tertentu)
  - Formalitas
  - Medium (tulis, lisan)
  - Topik (isi atau kandungan ranah dari teks)

# Solusi perancang korpus

## Menuju korpus yang representatif

- Beragam manifestasi pemakaian bahasa untuk suatu bahasa (populasi) diikutkan 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

# Lancaster-Oslo/Bergen (LOB) Corpus

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 |

**Sampel data untuk masing-masing kategori teks hampir sama (sekitar 2000 kata)**

(McEnery & Hardie 2012: 10)

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


**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

https://corpora.uni-leipzig.de

CORPORA COLLECTION
LEIPZIG UNIVERSITY

*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.

## 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>
    get<Verb>
    <DirObject, ·NP>
        a<Det>
        postcard<NounSing>
    <Adverbial, ·PP>
        from<Prep>
        him<PersPron>.
```
Pemilahan struktur sintaksis (*Syntactic parsing*)

(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
%syn:   trans
```
Anotasi bertingkat

**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)



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

# Bahasa daerah yang seperti apa?

# Bahasa daerah “besar”?

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

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

# Bahasa daerah “besar”?

atau ***juga***

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

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

# 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)



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

# 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

# FINDABLE

# Leipzig Corpora

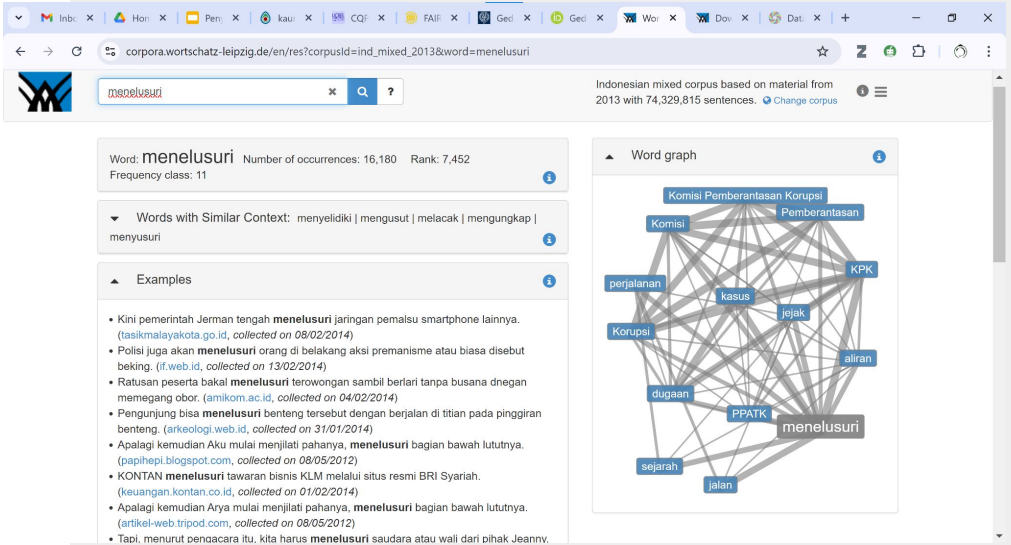
Contoh korpus dengan FAIR principle


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

# ACCESSIBLE

# Leipzig Corpora

Contoh korpus dengan FAIR principle

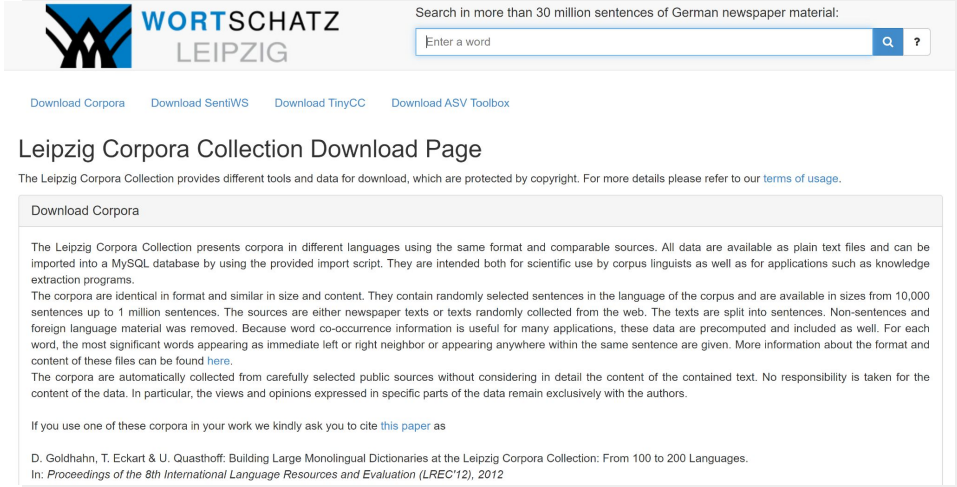



Cara mengakses daring + fitur infrastruktur terkait analisis (mis. word network, word co-occurrence)

# ACCESSIBLE

# Leipzig Corpora

Contoh korpus dengan FAIR principle




Cara mengakses (termasuk mengunduh) sangat eksplisit & terbuka

T&C + cara atribusi jika menggunakan

# Leipzig Corpora

Contoh korpus dengan FAIR principle

## INTEROPERABLE

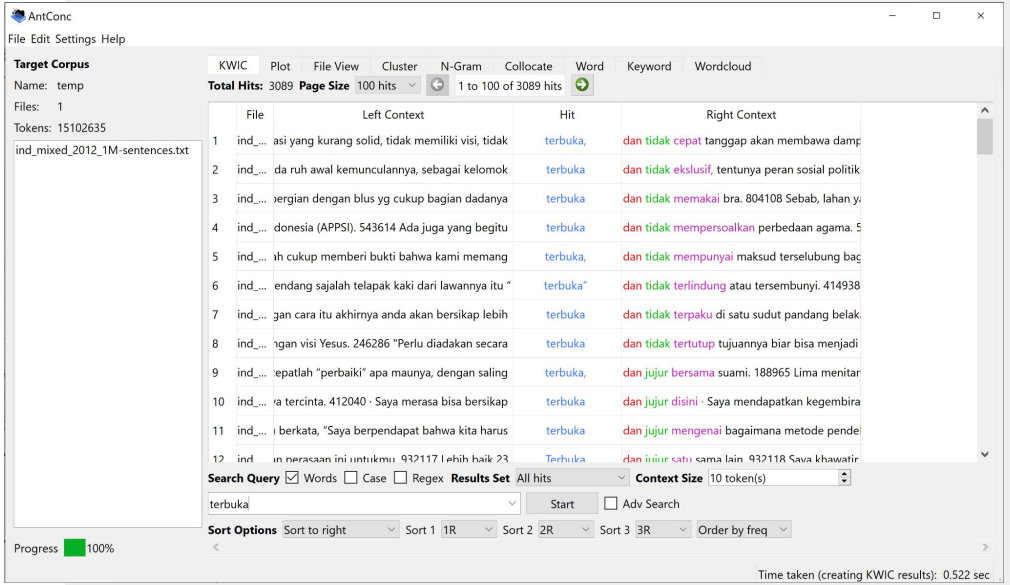
“the data need to interoperate with applications or workflows for analysis, storage, and processing.”

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

# INTEROPERABLE

# Leipzig Corpora

Contoh korpus dengan FAIR principle

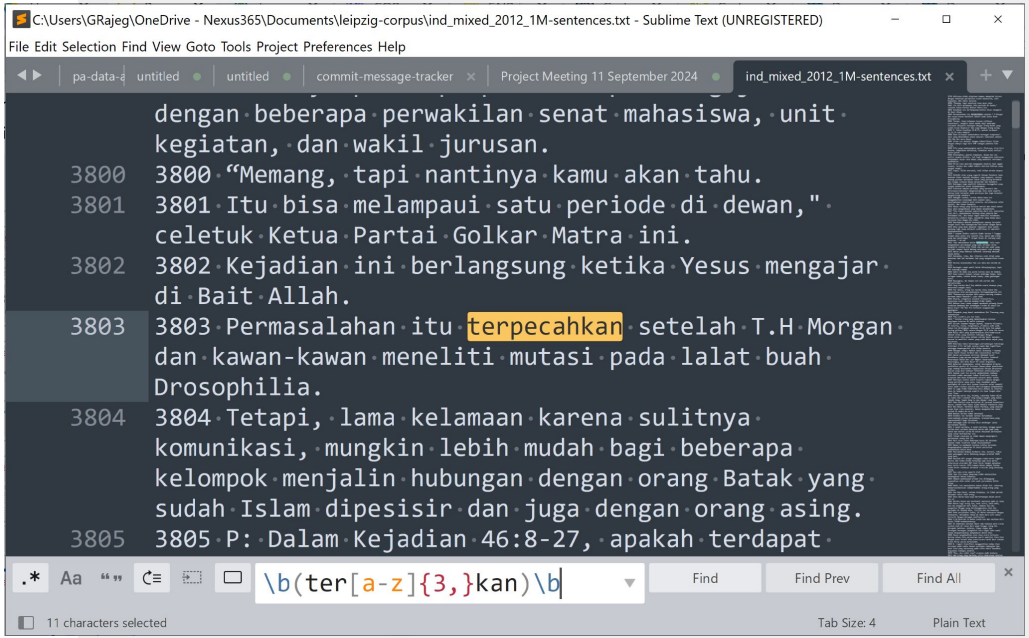



Rajeg, G. P. W. (2024, October 24). *Penyusunan Korpus Bahasa Daerah* [Invited talk]. *Konsinyasi Penyiapan Data Korpus Bahasa Daerah dan Pemetaan Bahasa*. Badan Pengembangan dan Pembinaan Bahasa, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi melalui Pusat Pengembangan dan Pelindungan Bahasa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

# INTEROPERABLE

# Leipzig Corpora

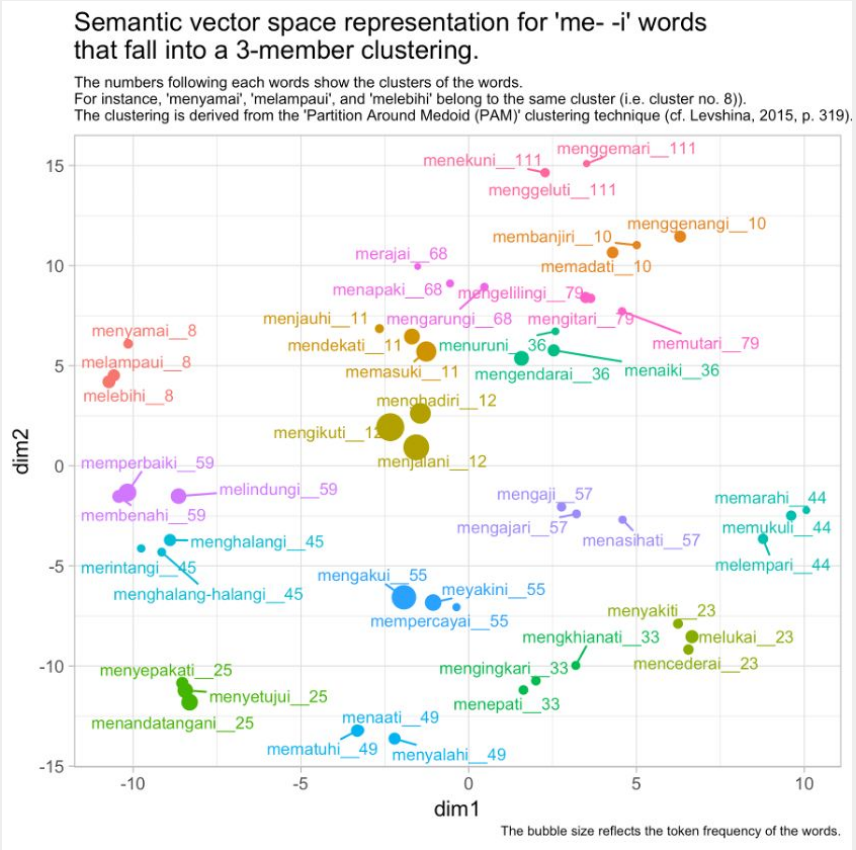
Contoh korpus dengan FAIR principle



# INTEROPERABLE

# Leipzig Corpora

Contoh korpus dengan FAIR principle



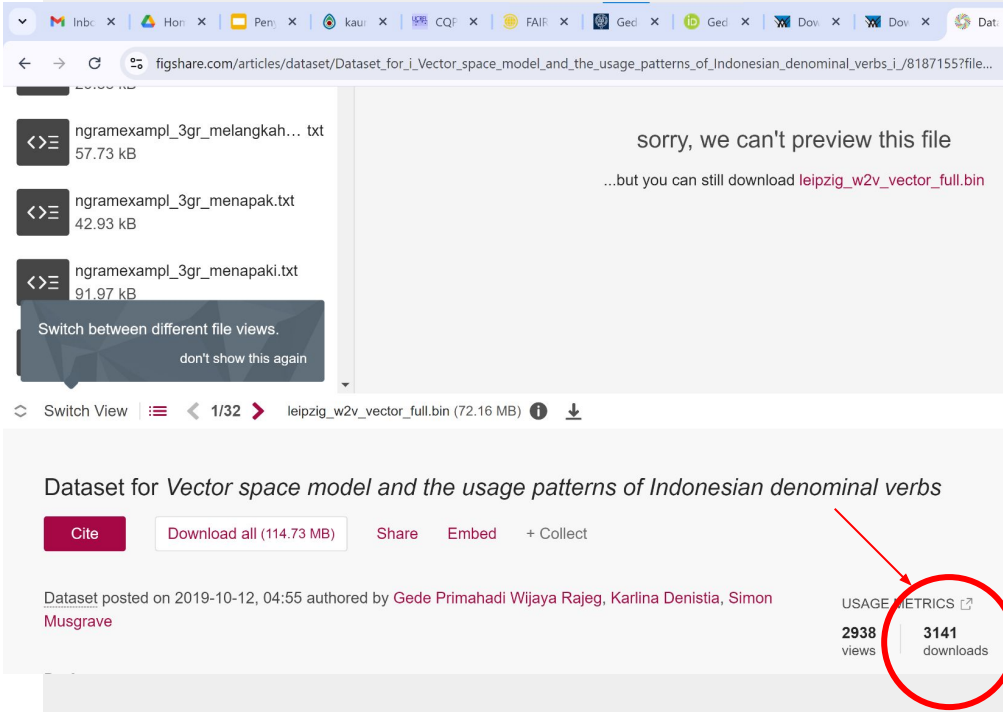


Rajeg, G. P. W., Denistia, K., & Musgrave, S. (2019). Vector Space Models and the usage patterns of Indonesian denominal verbs: A case study of verbs with meN-, meN-/-kan, and meN-/-i affixes. *NUSA*, *67*, 35–75. https://doi.org/10.15026/94452

# REUSABLE

# Leipzig Corpora

Contoh korpus dengan FAIR principle




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

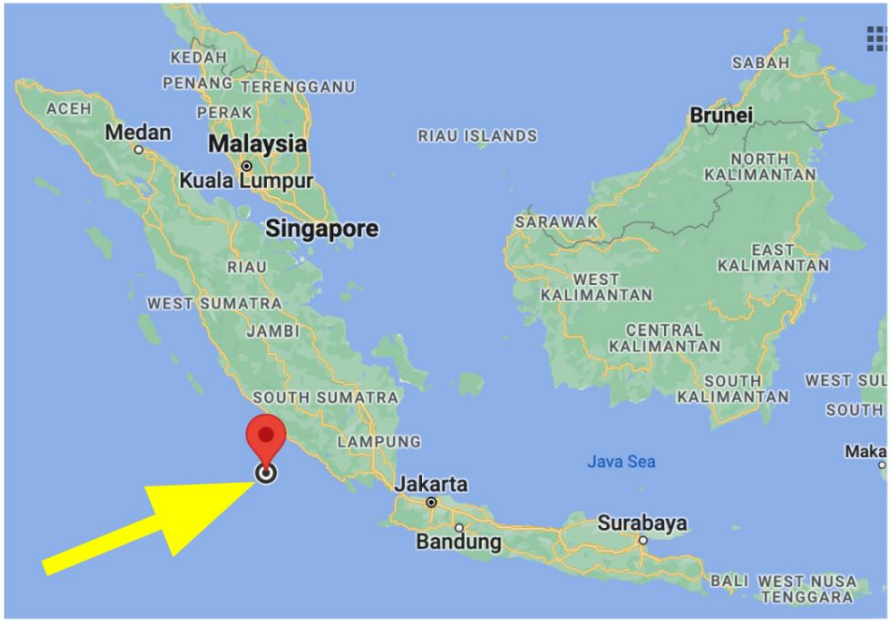
1. ~~Konsep *korpus*~~
2. ~~Target Bahasa~~
3. ~~Sumber (data & daya)~~
4. ~~Aksesibilitas + Infrastruktur~~
5. Contoh kasus 📌:
    a. Bahasa Enggano (sumber data + daya, aksesibilitas)
    b. Bahasa Bali (sumber data + daya; model gerakan komunitas; infrastruktur awal)
    c. Bahasa-bahasa di Australia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

# Acknowledgment: mėk ėm nah

## Enggano team & community


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

# 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

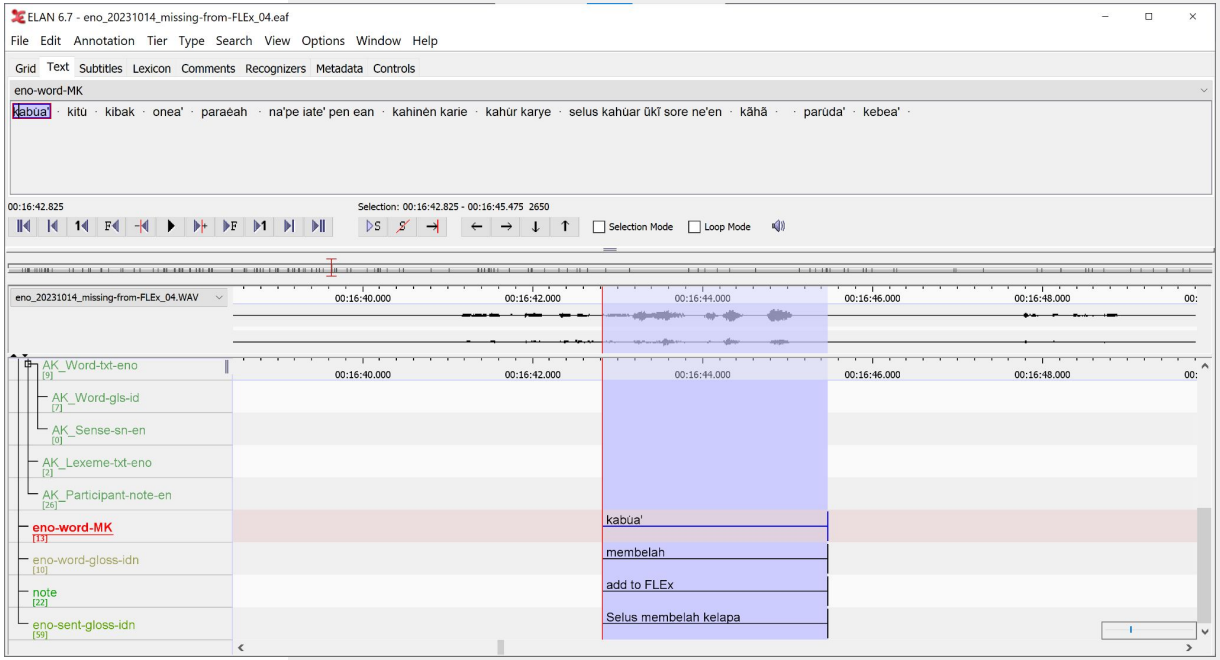
**Bakblau**
Participants/*Peserta*: Harun Kaarubi, I Wayan Arka
Date of recording/*Tanggal perekaman*: 6 February 2018
Location/*Lokasi*: Enggano Island
Video recording with transcription/*Rekaman video dengan transkripsi* (LingView)
Downloadable resources/*Bebagai bahan yang dapat diunduh*:
Audio file/*Berkas 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

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

# Dokumentasi Bahasa Kontemporer

Transkripsi dan terjemahan melalui ELAN



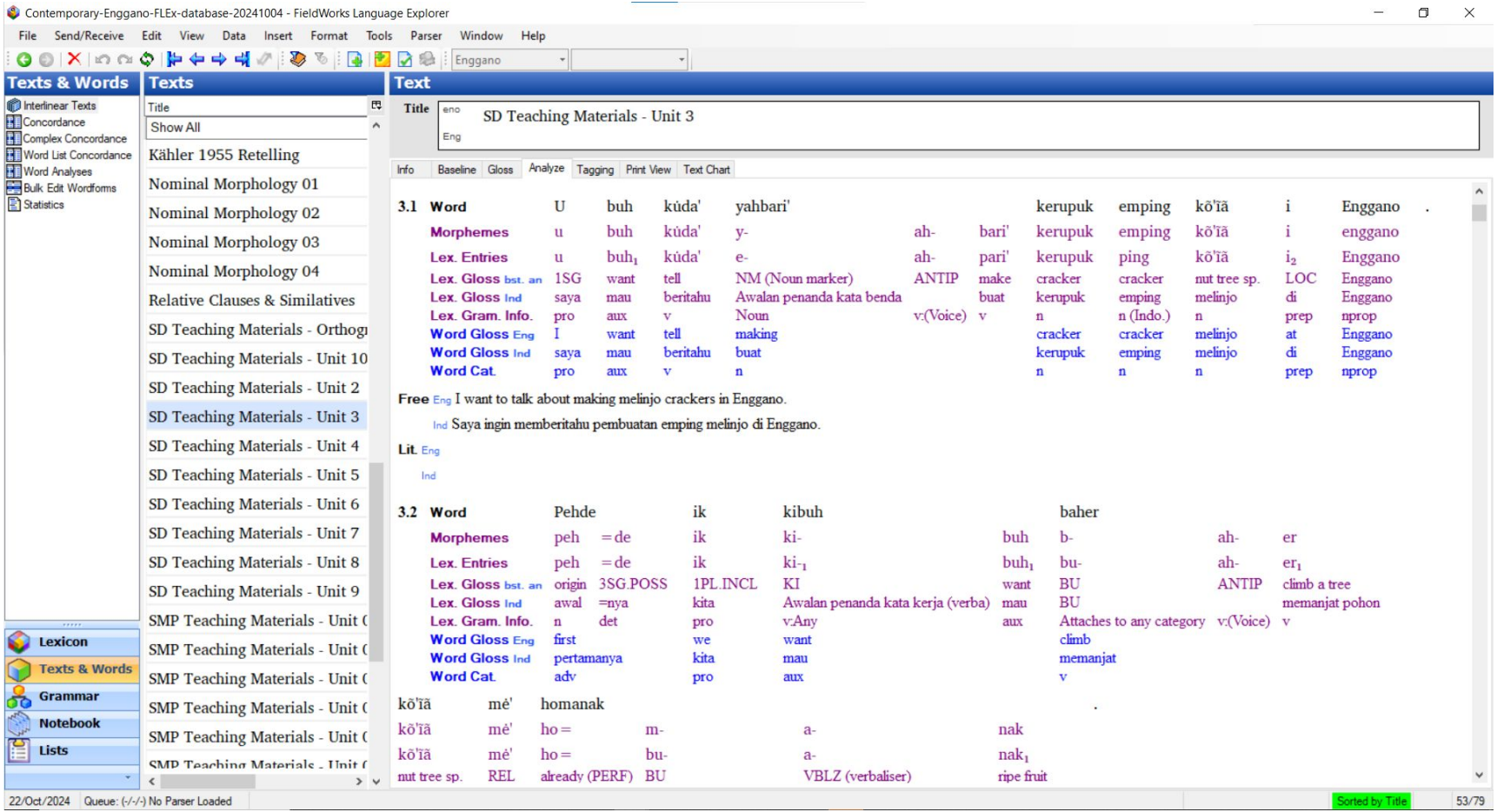



Rajeg, G. P. W. (2024, October 24). *Penyusunan Korpus Bahasa Daerah* [Invited talk]. *Konsinyasi Penyiapan Data Korpus Bahasa Daerah dan Pemetaan Bahasa*. Badan Pengembangan dan Pembinaan Bahasa, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi melalui Pusat Pengembangan dan Pelindungan Bahasa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

# Dokumentasi Bahasa Kontemporer

## Terjemahan leksikon dan anotasi morfologis di FLEx

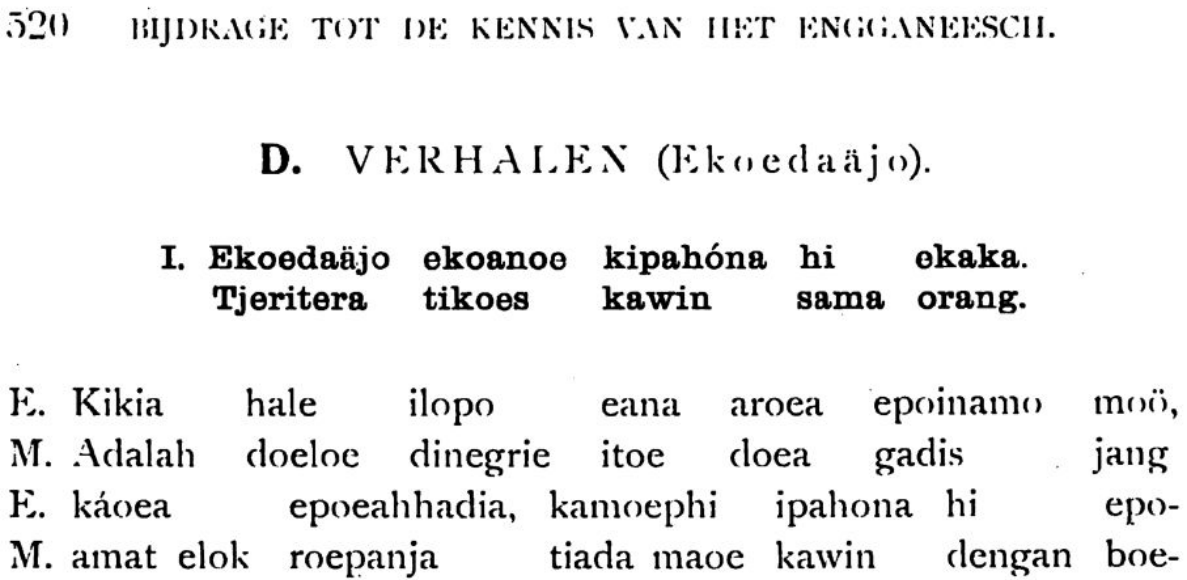
## Fieldwork Language Explorer


Contemporary-Enggano-FLEx-database-20241004 - FieldWorks Language Explorer
File Send/Receive Edit View Data Insert Format Tools Parser Window Help
Enggano
Texts & Words
Interlinear Texts
Concordance
Complex Concordance
Word List Concordance
Word Analyses
Bulk Edit Wordforms
Statistics
Lexicon
Texts & Words
Grammar
Notebook
Lists
Texts
Title
Show All
Kähler 1955 Retelling
Nominal Morphology 01
Nominal Morphology 02
Nominal Morphology 03
Nominal Morphology 04
Relative Clauses & Similatives
SD Teaching Materials - Orthogr
SD Teaching Materials - Unit 10
SD Teaching Materials - Unit 2
SD Teaching Materials - Unit 3
SD Teaching Materials - Unit 4
SD Teaching Materials - Unit 5
SD Teaching Materials - Unit 6
SD Teaching Materials - Unit 7
SD Teaching Materials - Unit 8
SD Teaching Materials - Unit 9
SMP Teaching Materials - Unit (
SMP Teaching Materials - Unit (
SMP Teaching Materials - Unit (
SMP Teaching Materials - Unit (
SMP Teaching Materials - Unit (
SMP Teaching Materials - Unit (
Text
Title
eno
Eng
SD Teaching Materials - Unit 3
Info Baseline Gloss Analyze Tagging Print View Text Chart
3.1 Word U buh kúda' yahbari' kerupuk emping kō'īā i Enggano .
Morphemes u buh kúda' y- ah- bari' kerupuk emping kō'īā i enggano
Lex. Entries u buh1 kúda' e- ah- pari' kerupuk ping kō'īā 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 Awalan penanda kata benda buat kerupuk emping melinjo di Enggano
Lex. Gram. Info. pro aux v Noun v:(Voice) v n n (Indo.) prep nprop
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 n prep nprop
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 Awalan 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
kō'īā mé' homanak .
kō'īā mé' ho= m- a- nak
kō'īā mé' 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:

# Kekayaan tinggalan 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).

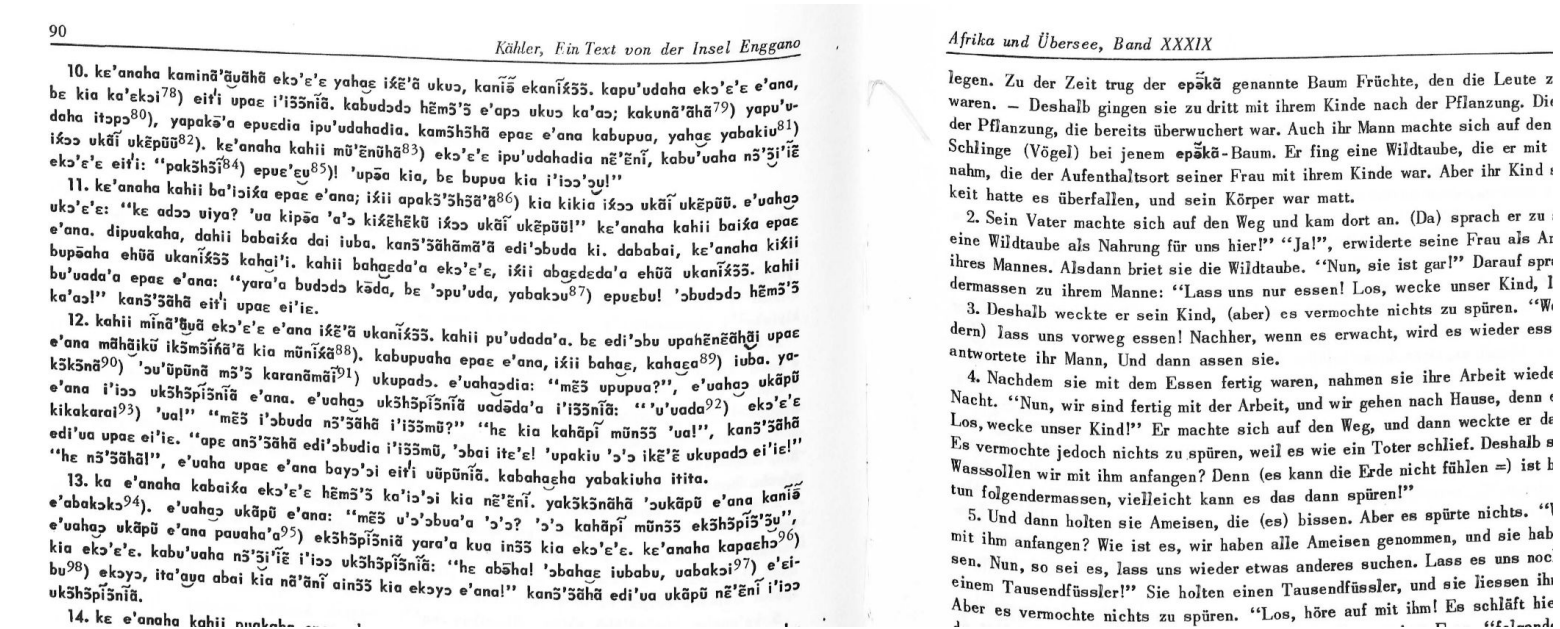
**I. Ekoedaäjo ekoanoe kipahóna hi ekaka.**
**Tjeritera tikoes kawin sama orang.**

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-

PDF-scanned - digitalisasi calon sumber data yang ada

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

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



90 *Kähler, Ein Text von der Insel Enggano*

10. kɛ'anaha kaminã'ãụãhã ekɔ'ɛ'ɛ yahaɛ iɤ̃ɛ̃'ã ukuɔ, kanĩ̃ẽ ekanĩɤ̃ɔ̃ɔ̃. kapu'udaha ekɔ'ɛ'ɛ e'ana, bɛ kia ka'ɛkɔi[78]) eit'i upaɛ i'iɔ̃nĩ̃ã. kabudɔdɔ hẽmɔ̃'ɔ̃ e'apɔ ukuɔ ka'aɔ; kakunã'ãhã[79]) yapu'udaha itɔpɔ[80]), yapakɔ̃'a epuɛdia ipu'udahadia. kamɔ̃hɔ̃hã epaɛ e'ana kabupua, yahaɛ yabakiu[81]) iɤ̃ɔɔ ukãĩ ukẽpũũ[82]). kɛ'anaha kahii mũ'ẽnũhã[83]) ekɔ'ɛ'ɛ ipu'udahadia nẽ'ẽnĩ, kabu'uaha nɔ̃'ɔ̃i'ĩẽ ekɔ'ɛ'ɛ eit'i: "pakɔ̃hɔ̃ĩ[84]) epuɛ'ɛụ[85])! 'upɔ̃a kia, bɛ bupua kia i'iɔɔ'ɔụ!"

11. kɛ'anaha kahii ba'iɔiɤ̃a epaɛ e'ana; iɤ̃ii apakɔ̃'ɔ̃hɔ̃ã'ã[86]) kia kikia iɤ̃ɔɔ ukãĩ ukẽpũũ. e'uahaɔ ukɔ'ɛ'ɛ: "kɛ adɔɔ uiya? 'ua kipɔ̃a 'a'ɔ kiɤ̃ẽhẽkũ iɤ̃ɔɔ ukãĩ ukẽpũũ!" kɛ'anaha kahii baiɤ̃a epaɛ e'ana. dipuakaha, dahii babaiɤ̃a dai iuba. kanɔ̃'ɔ̃ãhãmã'ã edi'ɔbuda ki. dababai, kɛ'anaha kiɤ̃ii bupɔ̃aha ehũã ukanĩɤ̃ɔ̃ɔ̃ kahaị'i. kahii bahaɛda'a ekɔ'ɛ'ɛ, iɤ̃ii abaɛdɛda'a ehũã ukanĩɤ̃ɔ̃ɔ̃. kahii bu'uada'a epaɛ e'ana: "yara'a budɔdɔ kãda, bɛ 'ɔpu'uda, yabakɔu[87]) epuɛbu! 'ɔbudɔdɔ hẽmɔ̃'ɔ̃ ka'aɔ!" kanɔ̃'ɔ̃ãhã eit'i upaɛ ei'iɛ.

12. kahii mĩnã'ãụã ekɔ'ɛ'ɛ e'ana iɤ̃ɛ̃'ã ukanĩɤ̃ɔ̃ɔ̃. kahii pu'udada'a. bɛ edi'ɔbu upahẽnẽãhãi upaɛ e'ana mãhãikũ ikɔmɔ̃ĩɲã'ã kia mũnĩɤ̃ã[88]). kabupuaha epaɛ e'ana, iɤ̃ii bahaɛ, kahaɛa[89]) iuba. yakɔ̃kɔnã[90]) 'ɔu'ũpũnã mɔ̃'ɔ̃ karanãmãĩ[91]) ukupadɔ. e'uahaɔdia: "mẽɔ̃ upupua?", e'uahaɔ ukãpũ e'ana i'iɔɔ ukɔ̃hɔ̃pĩɔ̃nĩã e'ana. e'uahaɔ ukɔ̃hɔ̃pĩɔ̃nĩã uadɔ̃da'a i'iɔ̃nĩã: "'u'uada[92]) ekɔ'ɛ'ɛ kikakarai[93]) 'ua!" "mẽɔ̃ i'ɔbuda nɔ̃'ɔ̃ãhã i'iɔ̃mũ?" "hɛ kia kahãpĩ mũnɔ̃ɔ̃ 'ua!", kanɔ̃'ɔ̃ãhã edi'ua upaɛ ei'iɛ. "apɛ anɔ̃'ɔ̃ãhã edi'ɔbudia i'iɔ̃mũ, 'ɔbai ite'ɛ! 'upakiu 'ɔ'ɔ ikẽ'ɛ ukupadɔ ei'iɛ!" "hɛ nɔ̃'ɔ̃ãhãl", e'uaha upaɛ e'ana bayɔ'ɔi eit'i uũpũnĩã. kabahaɛha yabakiuha itita.

13. kɑ e'anaha kabaiɤ̃a ekɔ'ɛ'ɛ hẽmɔ̃'ɔ̃ kɑ'iɔ'ɔi kia nẽ'ẽnĩ. yakɔ̃kɔnãhã 'ɔukãpũ e'ana kanĩ̃ẽ e'abakɔkɔ[94]). e'uahaɔ ukãpũ e'ana: "mẽɔ̃ u'ɔ'ɔbua'a 'ɔ'ɔ? 'ɔ'ɔ kahãpĩ mũnɔ̃ɔ̃ ekɔ̃hɔ̃pĩɔ̃'ɔụ", e'uahaɔ ukãpũ e'ana pauaha'a[95]) ekɔ̃hɔ̃pĩɔ̃nĩã yara'a kua inɔ̃ɔ̃ kia ekɔ'ɛ'ɛ. kɛ'anaha kapaɛhɔ[96]) kia ekɔ'ɛ'ɛ. kabu'uaha nɔ̃'ɔ̃i'ĩẽ i'iɔɔ ukɔ̃hɔ̃pĩɔ̃nĩã: "hɛ abɔ̃ha! 'ɔbahaɛ iubabu, uabakɔi[97]) e'eibu[98]) ekɔyɔ, ita'aụa abai kia nã'ãnĩ ainɔ̃ɔ̃ kia ekɔyɔ e'ana!" kanɔ̃'ɔ̃ãhã edi'ua ukãpũ nẽ'ẽnĩ i'iɔɔ ukɔ̃hɔ̃pĩɔ̃nĩã.

14. kɛ e'anaha kahii [illegible]

*Afrika und Übersee, Band XXXIX*

legen. Zu der Zeit trug der epɔ̃kã 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 epɔ̃kã-Baum. Er fing eine Wildtaube, die er mit nahm, die der Aufenthaltsort seiner Frau mit ihrem Kinde war. Aber ihr Kind keit 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 A 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. "W 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 wiede Nacht. "Nun, wir sind fertig mit der Arbeit, und wir gehen nach Hause, denn Los, wecke unser Kind!" Er machte sich auf den Weg, und dann weckte er d 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. " 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 hie

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

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

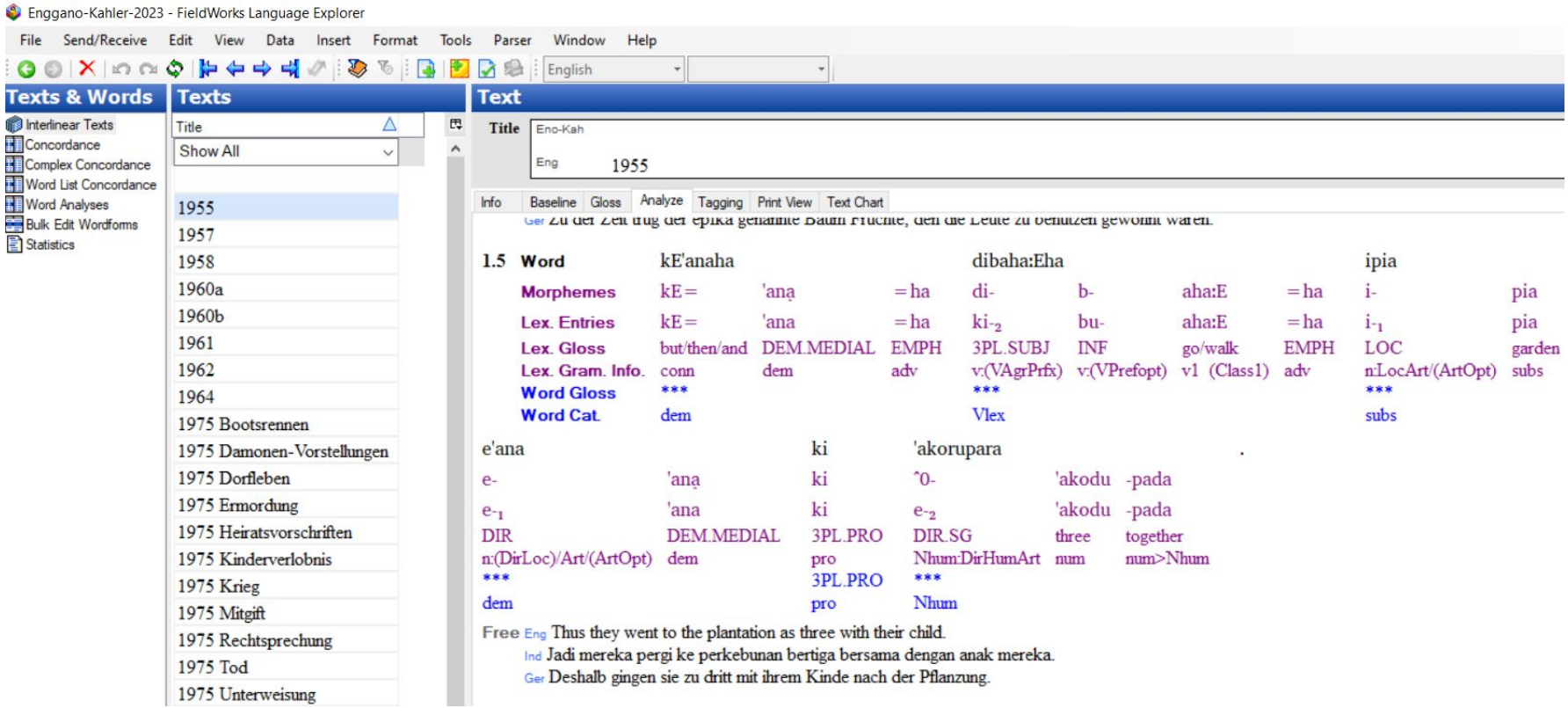
hēm 1/17

```
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. ekùda'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'ahã:ũãmãhã ukuo hẽmõ'õ kanĩŭ epŭkã hẽmõ'õ edixoo ba'ubuoda'a ukaka baka:ùhùa.
Zu der Zeit trug der epÎkâ genannte Baum Früchte, den die Leute zu benutzen gewohnt waren.
At that time the tree called epîka, which the people were used to using, bore fruit.
Pada saat itu pohon yang disebut epîka, yang biasa digunakan orang-orang, sedang berbuah.
```

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

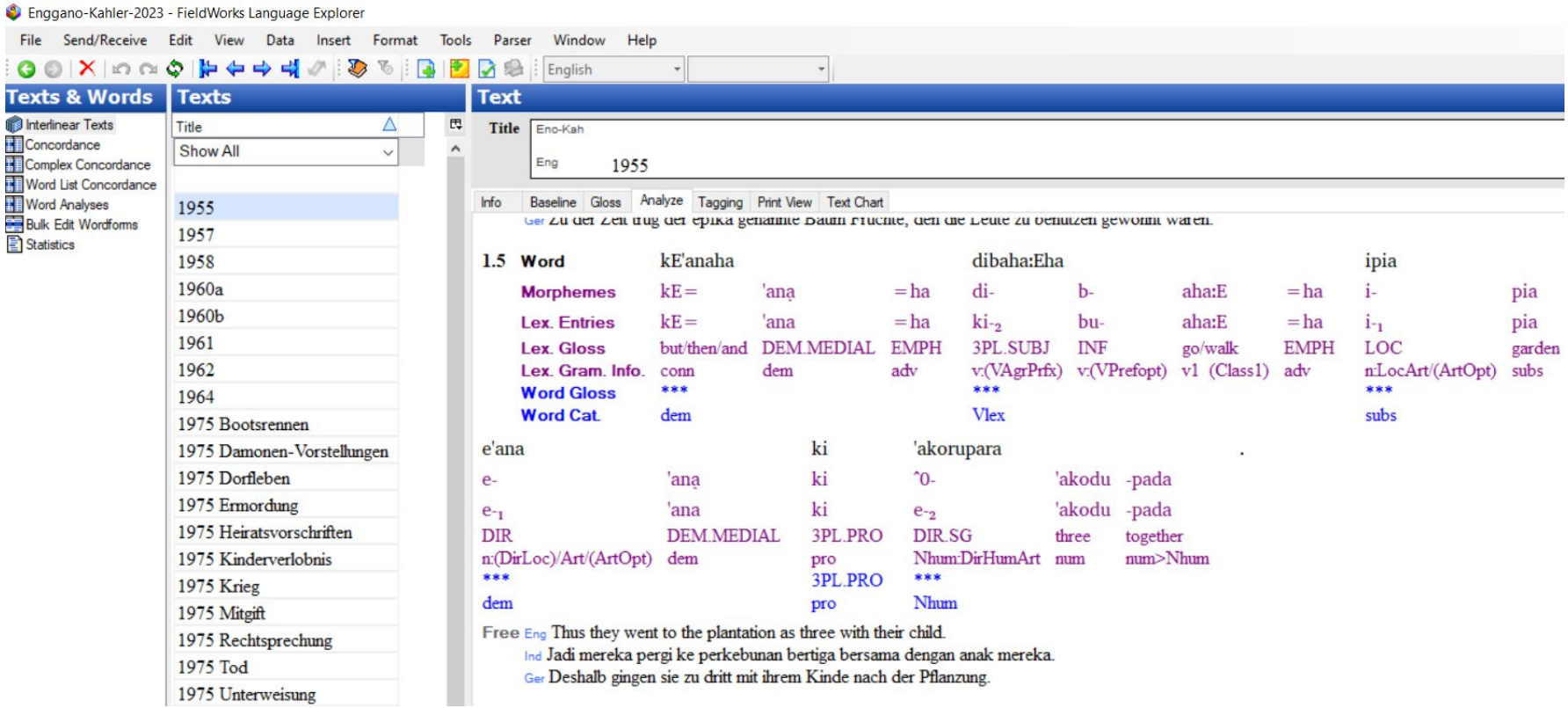
https://enggano.ling-phil.ox.ac.uk/static/previous.html

# Selanjutnya juga diolah dengan FLEx



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

# FLEx bisa diekspor ke XML (interoperability)






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

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

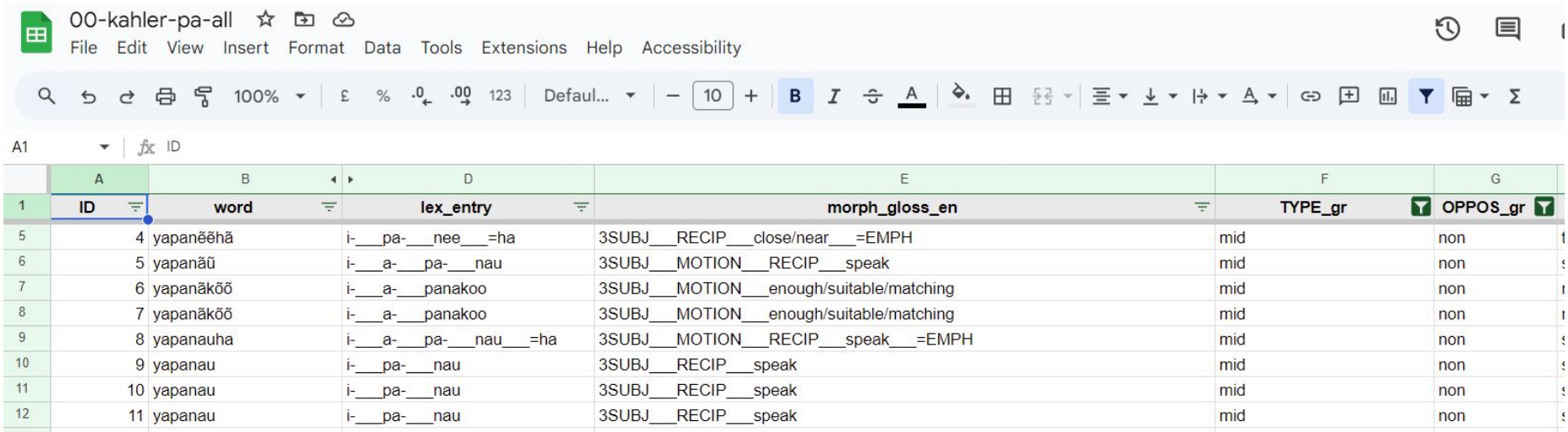
```
<phrase guid="3ce41d83-6ab2-4016-9f7c-0880e73c02d7">
  <item type="txt" lang="eno-ID-x-Kahler-1939">kabia e'ana ea'ahã:ũãmãhã ukuo hẽmõ'õ kanĩũ
  epũkã hẽmõ'õ edixoo ba'ubuoda'a ukaka baka:ùhùa. </item>
  <item type="segnum" lang="en">1.4</item>
  <words>
    <word guid="88d86541-ce59-4fc3-95b9-4ded2eeb6a8a">
      <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>
```



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

# 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 | yapanẽẽhã | i-___pa-___nee___=ha | 3SUBJ___RECIP___close/near___=EMPH | mid | non |
| 6 | 5 | yapanãũ | i-___a-___pa-___nau | 3SUBJ___MOTION___RECIP___speak | mid | non |
| 7 | 6 | yapanãkõõ | i-___a-___panakoo | 3SUBJ___MOTION___enough/suitable/matching | mid | non |
| 8 | 7 | yapanãkõõ | 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 |

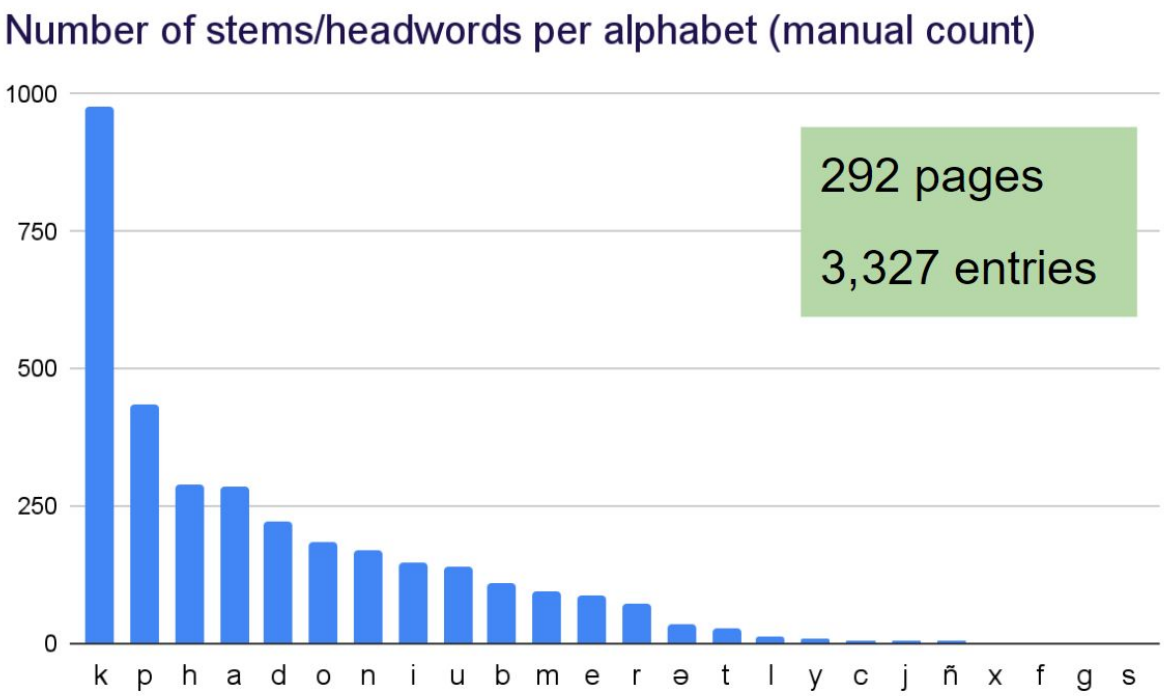


Rajeg, G. P. W. (2024, October 24). *Penyusunan Korpus Bahasa Daerah* [Invited talk]. *Konsinyasi Penyiapan Data Korpus Bahasa Daerah dan Pemetaan Bahasa*. Badan Pengembangan dan Pembinaan Bahasa, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi melalui Pusat Pengembangan dan Pelindungan Bahasa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

# 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)







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

# Enggano-German (Kähler 1987)

| | | |
|---|---|---|
| nãpũ III (=II?): | ka?anãpũã?ã = kanãpũã?ã | tropfen, trõpfeln |
| | ebõ ka?anãpũã?ã | das Wasser tropft |
| | hõ mãnãpũã?ã | tröpfelte bereits |
| | kipanãnãpũ | tropfen lassen |
| nãpũ IV | kinãpũ | herausschneiden, -lösen (G68) |
| | kamãnãpũ | wird herausgeschnitten |
| nãpũnũ: | kanãpũnũ | verfault, morsch |
| °nasi | (‹ ML nasi) | gekochter Reis |
| nãũ | › panãũ | |
| nã?ũmãnã | morgen (§ 29b) | |
| enãũõ | (arch,G42) = edopo | Land, Erde |

(Kähler 1987: 200)

| | | |
|---|---|---|
| tẽ | (DIA) = tõ, dõ | Art und Weise (§ 31a) |
| | kitẽ = kidõ | sein wie |
| tebe: | itebe = iẽebe (DIA) | an der Oberseite, oben (§ 26b) |
| te?ei: | ka?ite?ei | jmd schlechte Speisen anbieten |
| teo | kiteo ekitai | überreden |
| təhəda | (DIA) = dəhəda | |
| tixoi | = didixoi, didiki: | kitixoi etwas umwickeln |
| tikĩ | kahatikĩ | (mit etwas Biegsamem) schlagen |
| °etirĩ | (‹ ML sirih) | Betelbissen |
| tõ | = dõ, tẽ (DIA) | wie (beim Vergleich; § 31a) |
| | kitoiya ‹ kitõ eiya | wie das Wesen sein |

(Kähler 1987: 276)

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

# First step with OCR

(Kähler 1987: 276)

°etaku (‹ MKB/ML sagu) Sago

°takui (‹ ML sekoi) Gerste

tau̯?uo = dau̯?uo: kitau̯?uo einschließen, einsperren
ekitau̯?uo ("d Eingeschlossene" =) Witwe(r) während der offiziellen Trauerzeit
ekitau̯?ua Einschließensort

°etawaha (‹ ML sawah) nasses Reisfeld

tē (DIA) = tō, dō Art und Weise (§ 31a)
kitē = kidō sein wie

tebe: itebe = ißebe (DIA) an der Oberseite, oben (§ 26b)

te?ei: ka?ite?ei jmd schlechte Speisen anbieten

teo kiteo ekitai 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

OCR output via the `tesseract` R package

```
"etaku (¢ MKB/ML sagu) Sago
"takui (< ML sekoi) Gerste
7 7 * “ a iJ 7
tau ?uo = dau ?uo: kitau 2u0 einschlieBen, einsperren
Ne
ekitau?uo ("d Eingeschlossene" =) Witwe(r}) w&hrend
der offiziellen Trauerzeit
Pe
ekitay Pua Einschliefensort
"etawaha (< ML sawah) nasses Reisfeld
te (DIA) = to, dd Art und Weise (§ 3la)
kite = kido sein wie
tebe: itebe = idebe (DIA) an der Oberseite, oben (§ 26b)
te?ei: ka?ite’ei jmd schlechte Speisen anbieten
teo kiteo ekitai tiberreden
tohoeda (DIA) = dshsda
tixoi = didixoi, didiki: kitixoi etwas umwickeln
tiki kahatiki (mit etwas Biegsamem) schlagen
“etiri (< ML sirih) Betelbissen
to = do, té (DIA) wie (beim Vergleich; § 3la)
kitoiya *‘ kits eiya wie das Wesen sein
```

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

# Reverse-Engineering

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

lo4

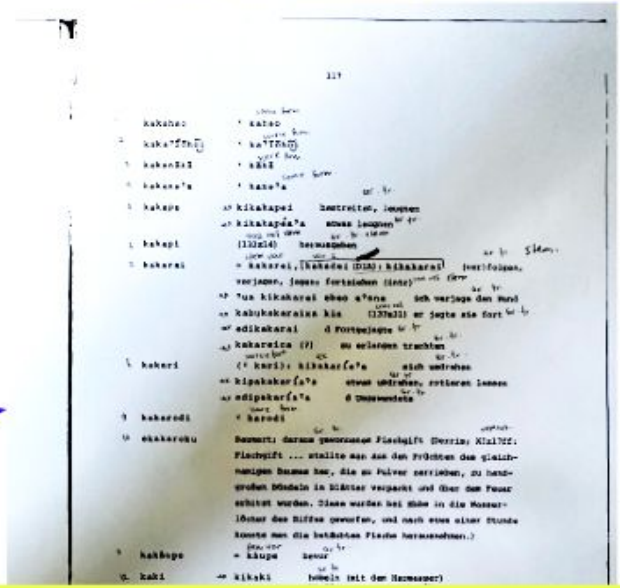
-ka?a — unser (in), von mir und euch: Possessivsuffix der 1.Pl in (§ 11)
eubáka?a unser (in) Haus

eka?a?aho = eka?aho, eka?a?oho Baumart (ML kasai, Pometia pinnata Forst.)

Ka?a?ăĭ Ort auf Enggano

ka?abeha (ka + abeha) und nachdem (Konjunktion)

ka?abu
kika?abu vergessen
hŏ buka?abu vergaß bereits
kika?abui jmd, etwas vergessen
dibuka?abui sie vergaßen etwas
eka?abua Vergessensort
eaka?abu = eka?abu das Vergessen
eka?abuía?a d Vergessene
eka?abuía?au edobu mein Vergessenes waren die Sachen
eka?abuiyo d zu Vergessende



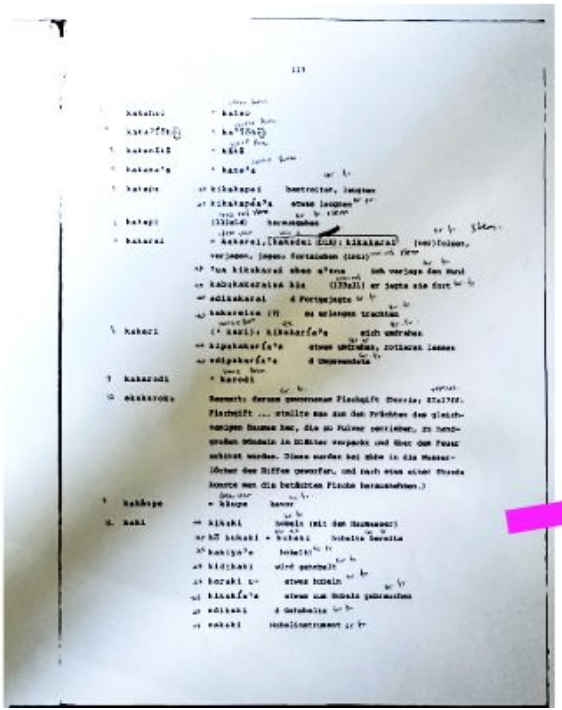
**Manual annotation/identification of the entry components**

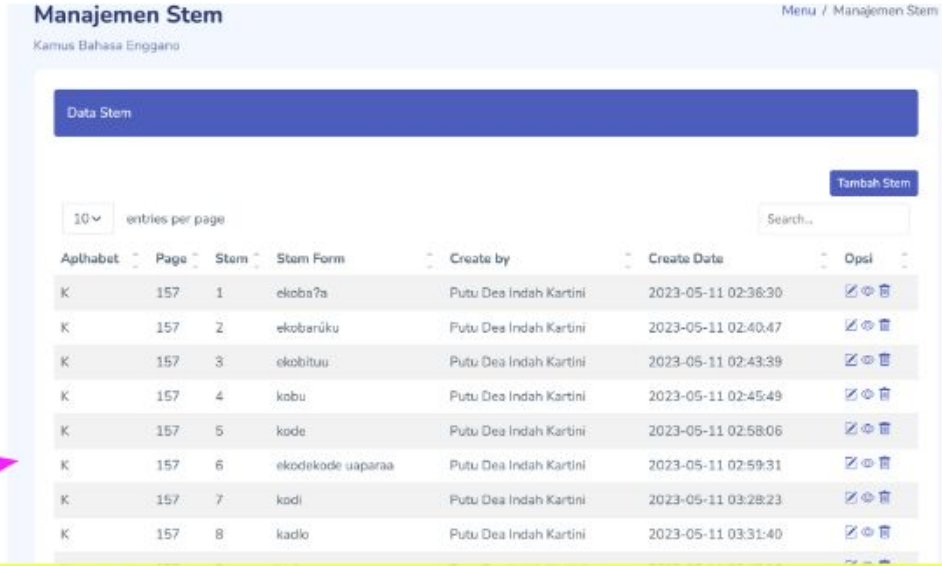
**Each page was checked for accuracy**



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

# Reverse-Engineering




**Typing the components into the relevant slot in the entry database**

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

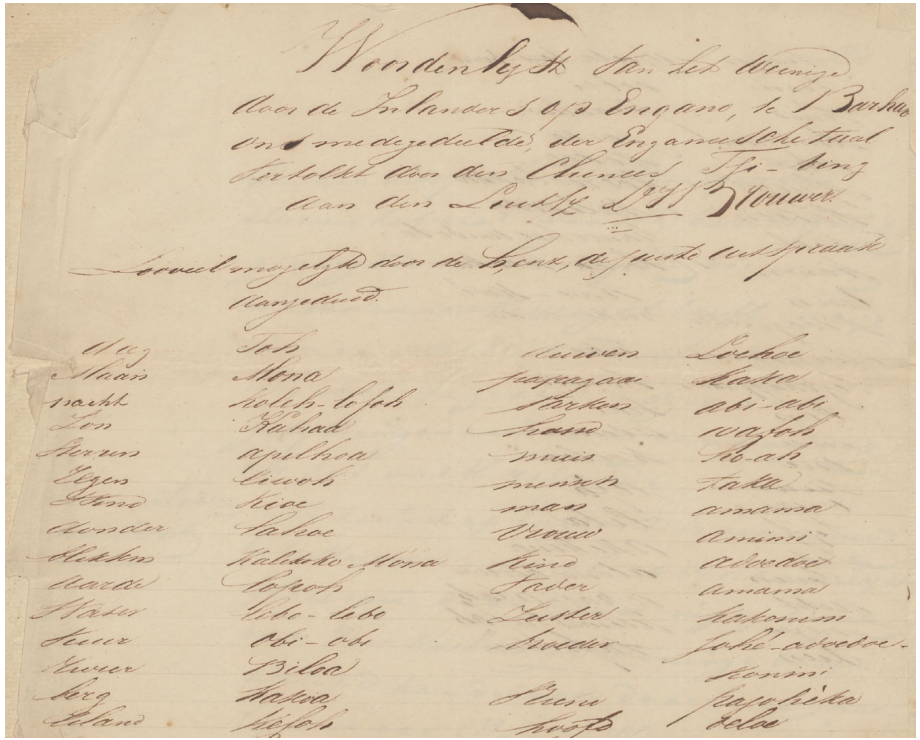
# Daftar Kata

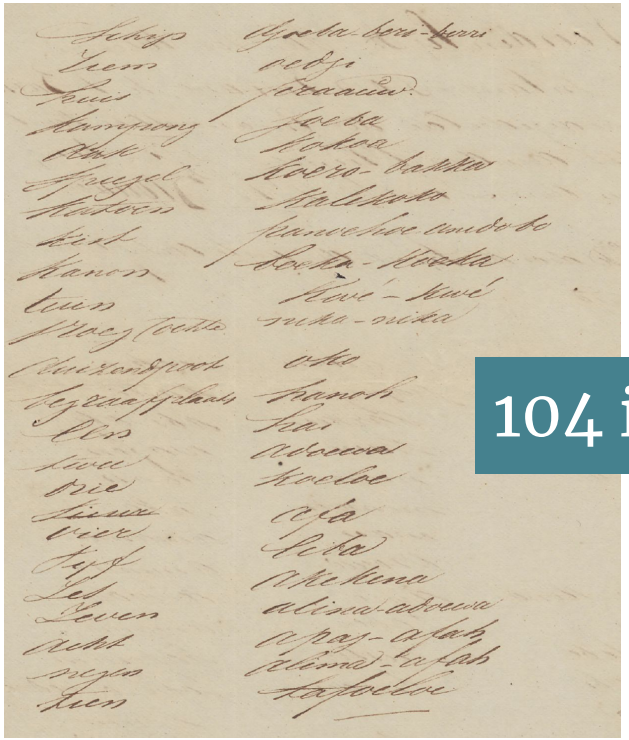
https://enggano.ling-phil.ox.ac.uk/static/previous.html

| **Author/*Penulis*** | **Stay in Enggano/*Periode tinggal di Enggano*** | **Publication/*Publikasi*** | **Number of Enggano words/*Jumlah kata Enggano*** |
|---|---|---|---|
| Djoeragan Boewang | several times between 1840 and 1855, including a visit of one year | Verslag orntrent het Eiland Enggano [Report on Enggano Island/*Laporan di Pulau Enggano*], *Tijdschrift voor Indische Taal-, Land- en Volkenkunde*, 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 Enganese 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/*Deskripsi Enggano dan penduduknya*], *Tijdschrift voor Indische Taal-, Land- en Volkenkunde*, 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/*Laporan penyelidikan yang dilakukan pada tahun 1854 di pulau Enggano*], *Tijdschrift voor Indische Taal-, Land- en Volkenkunde*, 1855 | 201 |
| J. Walland | 6-20 May 1863 | Het Eiland Enggano [Enggano Island/*Pulau Enggano*], *Tijdschrift voor Indische Taal-, Land- en Volkenkunde*, 1864 | 250 |
| R. Francis | 1865/66 and 1868/70 | Engganeesche woordenlijst [Enggano wordlist/*Daftar kata Enggano*], *Notulen der Bataviaasch Genootschap*, 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/*Kosakata bahasa Enggano, Mentawai, dan Nias*], *Tijdschrift voor Indische Taal-, Land- en Volkenkunde*, 1879 | 150 |
| A. C. Oudemans | - | Engano (bewesten Sumatra), Zijne Geschiedenis, Bewoners en Voortbrengselen [Engano (west of Sumatra), Its History, Inhabitants and Products/*Engano (Sumatera barat), Sejarah, Penduduk dan Produknya*], *Tijdschrift van het Koninklijk Nederlandsch Aardrijkskundig Genootschap*, 1889. This seems to be a reproduction of the Francis lists. | 160 |

# Brouwer (c. early 1850)

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

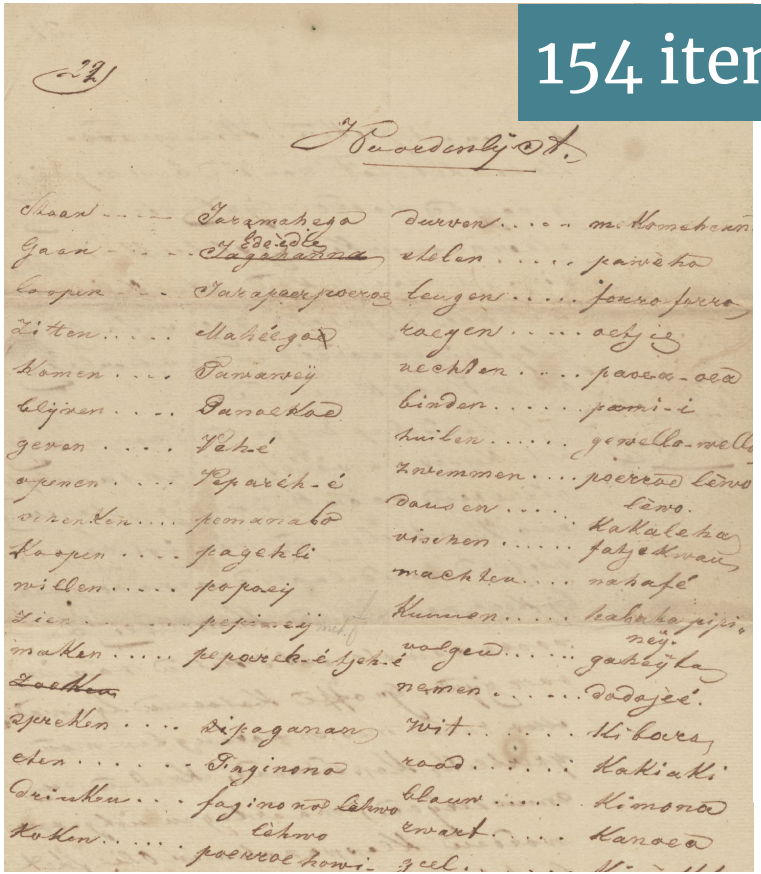




104 items

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

# Von Rosenberg (1855)



154 items

**Woordenlijst.**

| | |
|---|---|
| Staan | Jaramatrego. |
| Gaan | Edie-édie. |
| Loopen | Jarapoerpoera. |
| Zitten | Matrego. |
| Komen | Tawaweij. |
| Blijven | Panoekoe. |
| Geven | Pahè. |
| Schenken | Pemanabo. |
| Koopen | Pàgehlie. |
| Willen | Popo-ej. |

Von Rosenberg, Hermann. 1855. Beschrijving van Engano en van deszelfs bewoners. *Tijdschrift voor Indische Taal-, Land-en Volkenkunde* 3. 370–386.

# Francis (1870, in Oudemans 1889)

99 items

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

Comparative list with the earlier Enggano word lists

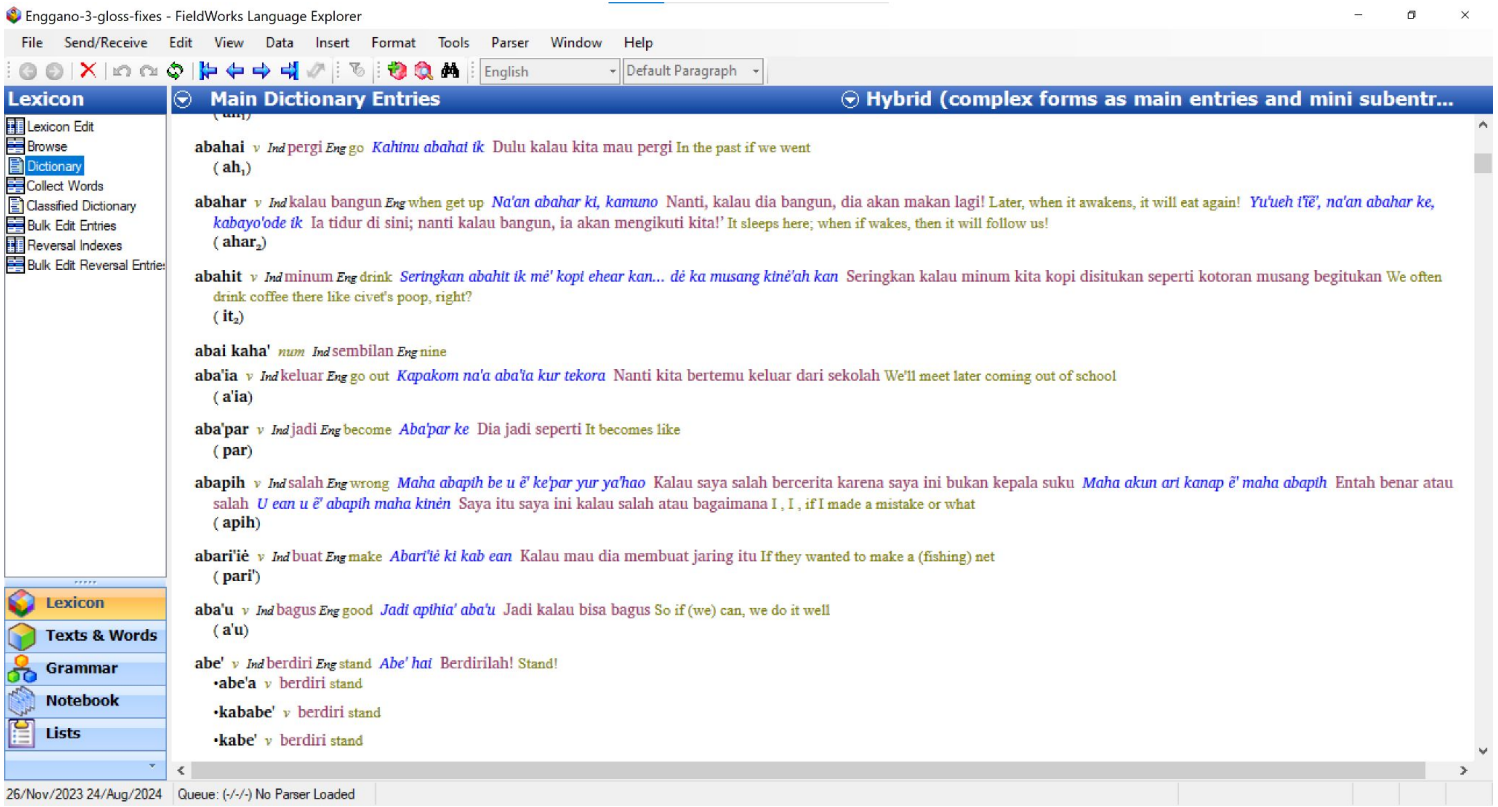
Oudemans, Jean Abraham Chrétien. 1889. Engano (bewesten Sumatra), zijne geschiedenis, bewoners en voortbrengselen. *Tijdschrift van het Koninklijk Nederlandsch Aardrijkskundig Genootschap* 6(1). 109–164.

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


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

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

# Kamus berbasis korpus Kontemporer Bahasa Enggano




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

# 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 + FLEx) akan dibagikan terbuka

# Butir pembahasan

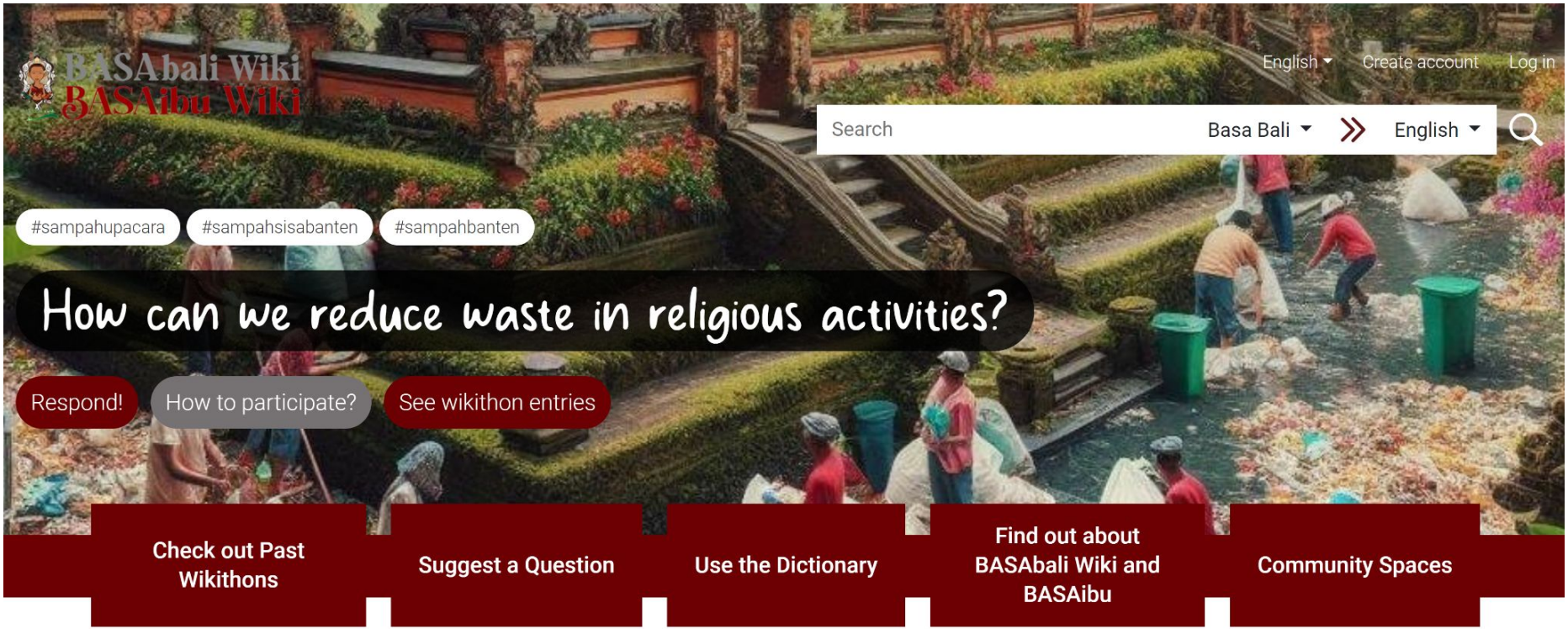
1. ~~Konsep *korpus*~~
2. ~~Target Bahasa~~
3. ~~Sumber (data & daya)~~
4. ~~Aksesibilitas + Infrastruktur~~
5. ~~Contoh kasus:~~
    a. ~~Bahasa Enggano (sumber data + daya, aksesibilitas)~~
    b. Bahasa Bali (sumber data + daya; model gerakan komunitas; infrastruktur awal) 📌
    c. Bahasa-bahasa di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

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

# Bahasa Bali

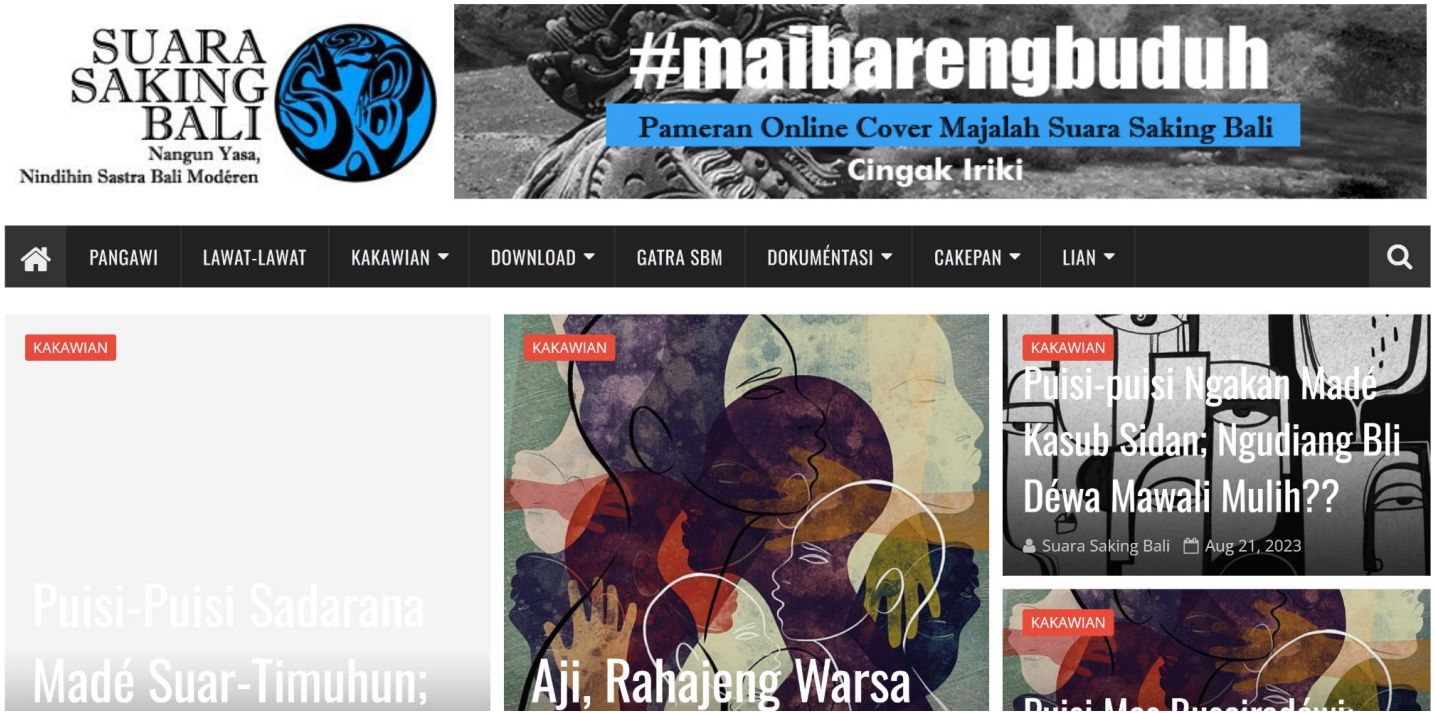
- https://basabali.org/
- Komunitas Bahasa Bali Wiki
  - Wikitionary Bahasa Bali (https://en.wiktionary.org/wiki/Bali)
  - Bahasa Bali Wiki (kamus daring dengan konsep crowdsourcing)
- Penyuluh bahasa Bali
  - Konservasi dan digitalisasi lontar
  - Transliterasi lontar

# BasaBali Wiki Virtual Dictionary - Open Access



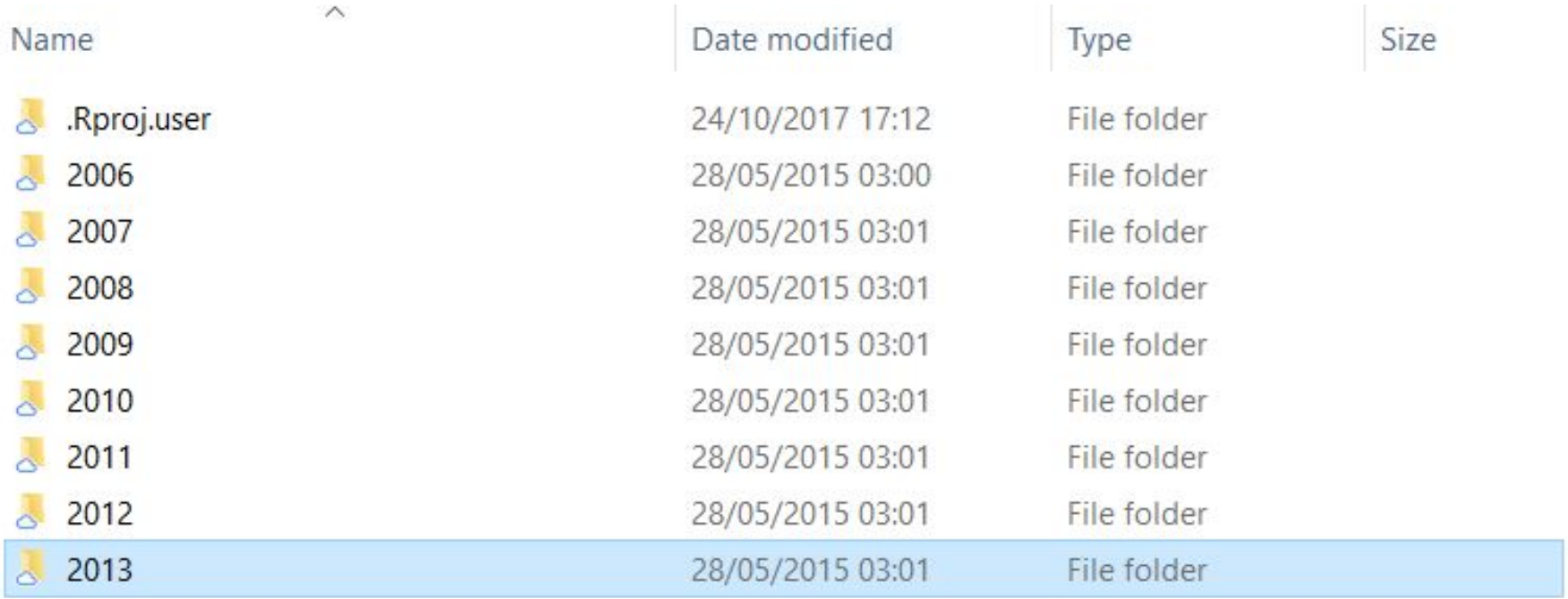

Rajeg, G. P. W. (2024, October 24). *Penyusunan Korpus Bahasa Daerah* [Invited talk]. *Konsinyasi Penyiapan Data Korpus Bahasa Daerah dan Pemetaan Bahasa*. Badan Pengembangan dan Pembinaan Bahasa, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi melalui Pusat Pengembangan dan Pelindungan Bahasa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

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




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

# 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 Bulélèng... | 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

## 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<br>*To view related information on a country, click its name* |

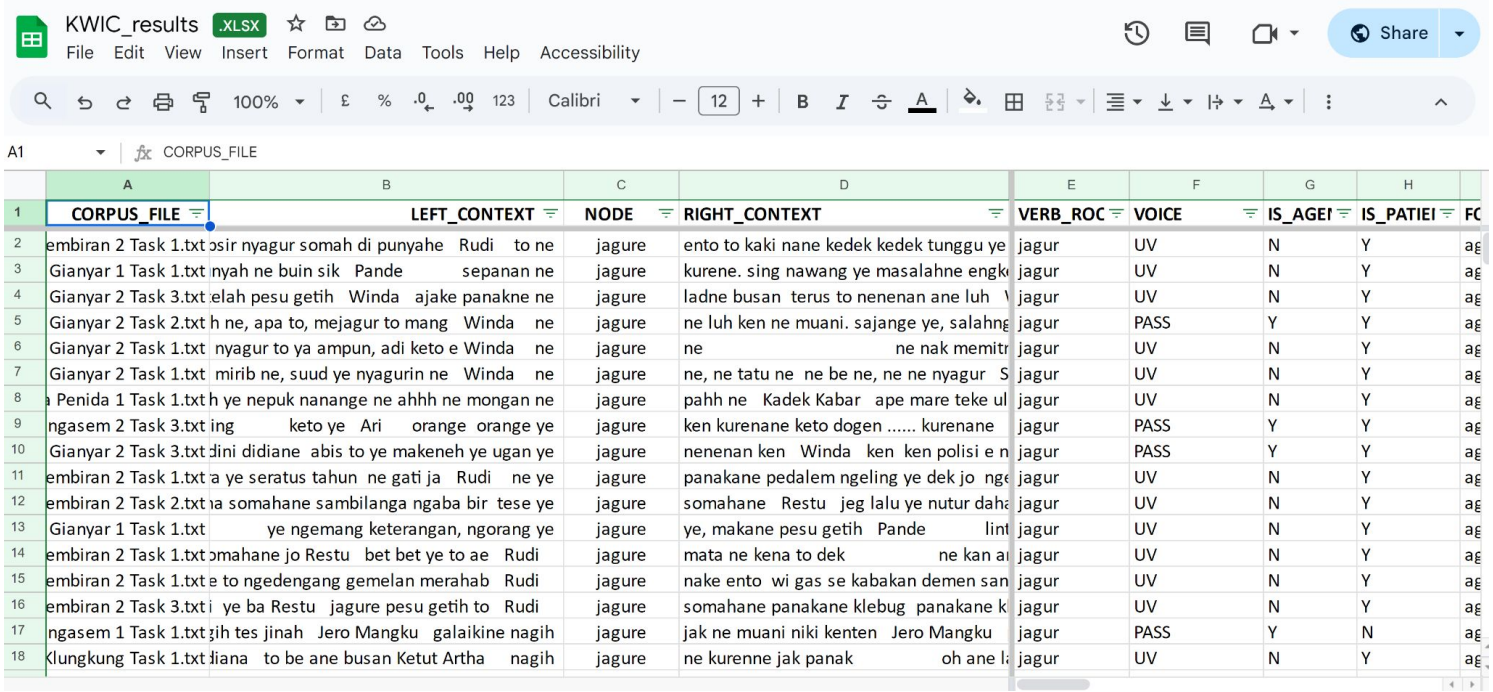
**Content Files (10)**

**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)




| | A | B | C | D | E | F | G | H | |
|---|---|---|---|---|---|---|---|---|---|
| 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 engk | 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 | a 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 | ıa somahane sambilanga ngaba bir tese ye | jagure | somahane Restu jeg lalu ye nutur daha | 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 ngedengang 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 | ɀih tes jinah Jero Mangku galaikine nagih | jagure | jak ne muani niki kenten Jero Mangku | jagur | PASS | Y | N | ag |
| 18 | Klungkung Task 1.txt | liana to be ane busan Ketut Artha nagih | jagure | ne kurenne jak panak oh ane l | 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)

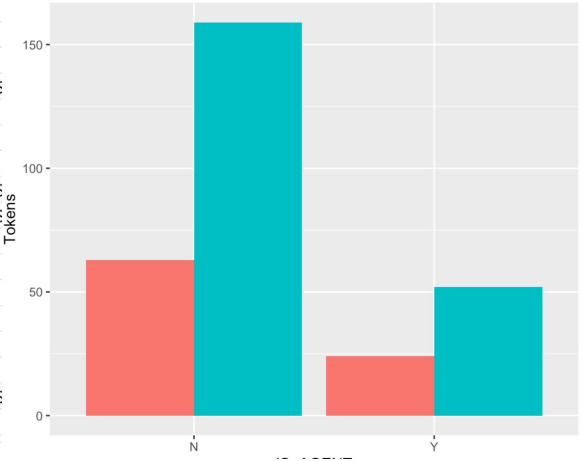
KWIC_results .XLSX

File Edit View Insert Format Data Tools Help Accessibility

Share

A1 CORPUS_FILE

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | CORPUS_FILE | LEFT_CONTEXT | NODE | RIGHT_CONTEXT | VERB_ROC | VOICE | IS_AGEI | IS_PATIEI |
| 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 | | |
| 3 | Gianyar 1 Task 1.txt | nyah ne buin sik Pande sepanan ne | jagure | kurene. sing nawang ye masalahne engk | jagur | UV | | |
| 4 | Gianyar 2 Task 3.txt | elah pesu getih Winda ajake panakne ne | jagure | ladne busan terus to nenenan ane luh | jagur | UV | | |
| 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 | | |
| 6 | Gianyar 1 Task 1.txt | nyagur to ya ampun, adi keto e Winda ne | jagure | ne ne nak memitr | jagur | UV | | |
| 7 | Gianyar 1 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 | | |
| 8 | a 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 | | |
| 9 | ngasem 2 Task 3.txt | ing keto ye Ari orange orange ye | jagure | ken kurenane keto dogen ..... kurenane | jagur | PASS | | |
| 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 | | |
| 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 | | |
| 12 | embiran 2 Task 2.txt | ıa somahane sambilanga ngaba bir tese ye | jagure | somahane Restu jeg lalu ye nutur daha | jagur | UV | | |
| 13 | Gianyar 1 Task 1.txt | ye ngemang keterangan, ngorang ye | jagure | ye, makane pesu getih Pande lint | jagur | UV | | |
| 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 | | |
| 15 | embiran 2 Task 1.txt | e to ngedengang gemelan merahab Rudi | jagure | nake ento wi gas se kabakan demen san | jagur | UV | | |
| 16 | embiran 2 Task 3.txt | i ye ba Restu jagure pesu getih to Rudi | jagure | somahane panakane klebug panakane k | jagur | UV | | |
| 17 | ngasem 1 Task 1.txt | gih tes jinah Jero Mangku galaikine nagih | jagure | jak ne muani niki kenten Jero Mangku | jagur | PASS | | |
| 18 | Klungkung Task 1.txt | liana to be ane busan Ketut Artha nagih | jagure | ne kurenne jak panak oh ane l | jagur | UV | | |




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

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

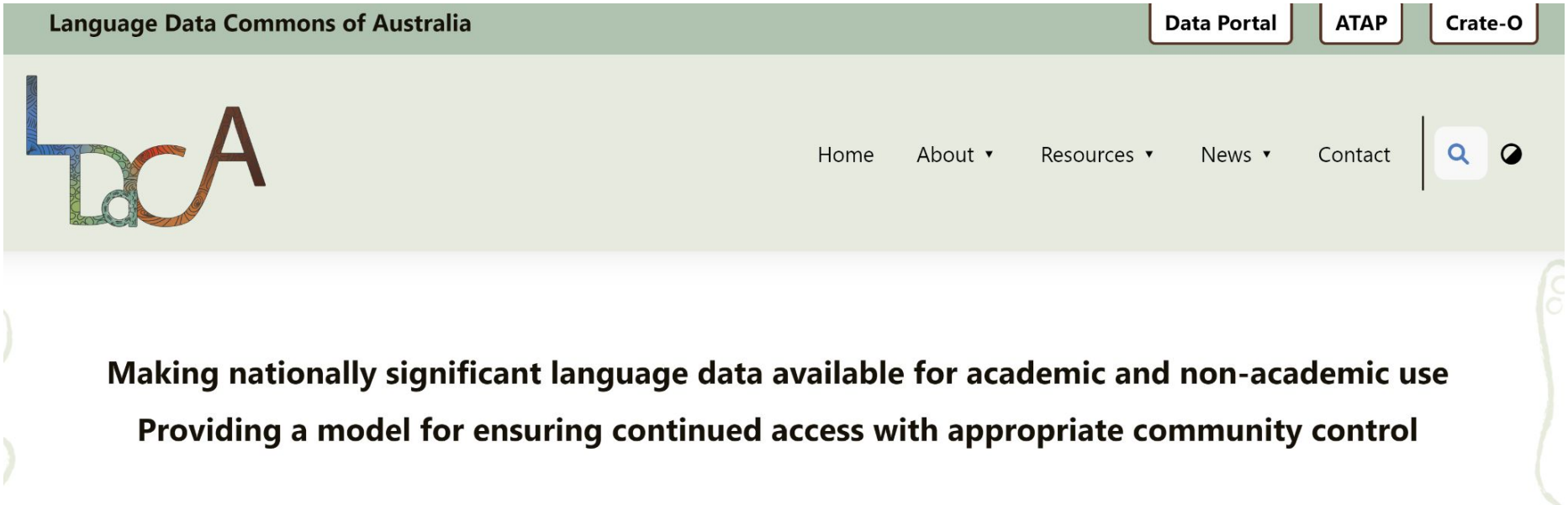
# Butir pembahasan

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) 📌



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

# LDaCA (Language Data Commons of Australia)




https://www.ldaca.edu.au/
https://ardc.edu.au/project/language-data-commons-of-australia/

# LDaCA (Language Data Commons of Australia)


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

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 yester day:
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 govern ment 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 (LDaCA) (sumber data + daya + infrastruktur nasional)~~ 📌



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

# 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



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

# 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