

Building an open concordancer for Malay/Indonesian


Hiroki Nomoto[†] Shiro Akasegawa[‡] Asako Shiohara[†]

[†]Tokyo University of Foreign Studies

[‡]Lago Institute of Language

ISMIL 22@ UCLA, 12/05/2018

Organization

- MALINDO Conc 
 - A new open online concordancer for Malay/Indonesian
 - Designed as a common tool among researchers of Malay/Indonesian
 - Free of charge
 - Easy to use
 - Yet allows moderately sophisticated search queries
- Compare it with the existing open concordancers.

Corpus search tools for Malay/Indonesian

Tool	Size (million)	Corpus
Malay Concordance Project	5.7 tokens	Classical Malay literature
Korpus DBP	135 tokens	Own data
SEAlang Malay	2.5 tokens	An Crúbadán (web corpora)
SEAlang Indonesian	5 tokens	
MALINDO CONC	1.8 sents (will upgrade to 4.8 sents)	Leipzig Corpora Collection (web corpora)

<https://malindoconc.lagoinst.info> (temporary URL)

MALINDO Conc

Bahasa

Pencarian

Lihat KEM KEM

Cari

Kata Kunci Hapus Semua

Bentuk Jadian tanpa spesifikasi

Akar Kata tanpa spesifikasi

Prefiks tanpa spesifikasi

Sufiks tanpa spesifikasi

Konfiks tanpa spesifikasi

Reduplikasi tanpa spesifikasi

Kolokasi tanpa spesifikasi Hapus Semua

Korpus tanpa spesifikasi

Halaman 0 dari 0 Hasil pencarian tidak ditemukan Unduh



CORPORA COLLECTION

LEIPZIG UNIVERSITY

Search in 263 Corpus-Based Monolingual Dictionaries for
236 Languages.



Corpus: Indonesian (ind_mixed_2013)

Indonesian mixed corpus based on material from 2013

Sentences: 74,329,815 · Types: 7,964,109 · Tokens: 1,206,281,985

German

English

French

Arabic

Russian

 all...

To download a corpus select a language and corpus size and download the corresponding data file.

Mixed

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

Newscrawl

Year	Country	Downloads					
2011		10K	30K	100K	300K	1M	3M
2012		10K	30K	100K	300K	1M	3M

Web

Year	Country	Downloads					
2011		10K	30K	100K	300K	1M	3M
2012		10K	30K	100K	300K	1M	3M
2013	Indonesia	10K	30K	100K	300K	1M	3M
2015	Brunei	10K	30K	100K	300K	1M	3M
2015	Indonesia	10K	30K	100K	300K	1M	3M

Wikipedia

Year	Country	Downloads					
2016		10K	30K	100K	300K	1M	3M

MALINDO Conc and the Malay Concordance Project

- MALINDO Conc was modelled after the Malay Concordance Project, an open online concordancer for Classical Malay.
(<http://mcp.anu.edu.au/>).
- Good features MALINDO Conc inherits:
 1. Any variety of Malay
 2. Morphological search
 3. Contributions from users

[1] Any variety of Malay

- MALINDO Conc intends to include **any variety of Malay** across the archipelago.
- The existing open concordancers deal with a particular geopolitical variety of Malay.
 - [Korpus DBP](#): Malaysian Malay
 - [SEALang Library Corpus \(Malay\)](#): Malaysian, Singaporean, Bruneian Malay
 - [SEALang Library Corpus \(Indonesian\)](#): Indonesian

Korpus tanpa spesifikasi

▼ Koleksi Korpus Leipzig

▼ IND

 IND MXD2012

 IND WEB2012

 IND WKP2016

▼ ZSM

 ZSM MXD2012

 ZSM WEB2012

 ZSM WKP2016

- 300K sents each
- 10 more IND
subcorpora coming soon

[2] Morphological search

One can search the corpus for forms with a particular morphological profile.

- Inflected forms of *fikir* and *fikirkan*
- *ber-...-kan* verbs
- *meN-X-X* & *X-meN-X* verbs
- *ingin* + *di-* verb & *ingin* + word (e.g. *untuk*) + *di-* verb

Kata Kunci

Hapus Semua



Bentuk Jadian tanpa spesifikasi

Akar Kata tanpa spesifikasi

Prefiks tanpa spesifikasi

Sufiks tanpa spesifikasi

Konfiks tanpa spesifikasi

Reduplikasi tanpa spesifikasi

Keyword > Prefixes

Prefiks termasuk

meN- di- ber- per- ter- peN-

pe- ke- **tanpa afiks**

OK Batal

Keyword > Suffixes

Sufiks

termasuk

-kan -i -an -nya -lah -kah

tanpa afiks

OK

Batal

Keyword > Circumfixes

Konfiks

termasuk

- ber-...-an ber-...-kan ke-...-an
 peN-...-an per-...-an pe-...-an
 se-...-nya **tanpa afiks**

OK

Batal

Keyword > Reduplication types

Reduplikasi

termasuk ▾

Penuh Sebagian Berubah bunyi

tanpa reduplikasi

OK

Batal

Example 1:

Inflected forms of *fikir/fikirkan*

Kata Kunci

Hapus Semua



Bentuk Jadian tanpa spesifikasi

Akar Kata sama dengan "fikir"

Prefiks any of (meN-, di-, tanpa afiks)

Sufiks any of (-kan, tanpa afiks)

Konfiks tanpa spesifikasi

Reduplikasi tanpa spesifikasi

fikir, memikir, fikirkan, memikirkan, difikirkan

Lihat **KWIC** Kalimat

1035	Bila saya fikir balik.	MXD2012
1036	uk kesimpulannya, aku rasa sudah sampai masanya bagi Melodi untuk memikirkan tentang KUALITI (info yang bagus) , bukan KUANTITI (ra...	MXD2012
1037	Dan saya tidak fikir dia (Umno) gembira kalau semua masalah selesai.	MXD2012
1038	Tension memikirkan hal ini.	MXD2012
1039	Saya juga pernah untung dan saya juga pernah rugi jadi saya fikir saya juga layak untuk berkongsi pengalaman.	MXD2012
1040	rsama-kekuatan pengendalian yang tidak ada orang lain bahkan boleh memikirkan pencocokan.	MXD2012
1041	Saya asyik memikirkan , macam mana la hidup mereka hari2 di asrama begini, sed...	MXD2012
1042	Tak terjangkau akal bila memikir .	MXD2012
1043	Anda tidak perlu memikirkan motif dan latar permaidani yang sesuai bergandingan deng...	MXD2012
1044	nenghadapi hari mendatang dan berkejaran mencari cinta yang seolah difikirkan sudah hilang.	MXD2012
1045	Maknanya: Fikirkan masak-masak sebelum mengeluarkan kata-kata yang pedas.	MXD2012
1046	Fikir dengan 4 modal utama ini, apa yang boleh anda lakukan.	MXD2012
1047	Team RS pun kena fikir balik.	MXD2012
1048	Gusar hati aku memikirkan kalian yang masuk ke hutan selama 2 hari, 2 malam.	MXD2012
1049	Saya hanya fikir ia betul-betul bagus untuk melihat bumi, dan melihatnya bergantun...	MXD2012
1050	" Tidakkah mereka mahu memikirkan dan meyakini bahawa Allah yang menciptakan langit dan b...	MXD2012

Example 2: Reduplication with *meN-*

Kata Kunci Padam Semua 

- Bentuk Lahir** tidak ditentukan
- Akar** tidak ditentukan
- Awalan** sama dengan "**meN-**"
- Akhiran** tidak ditentukan
- Apitan** tidak ditentukan
- Penggandaan** sama dengan "**Penuh**"

menutup-nutupi, meninjau-ninjau, kena-mengena, mengada-ngada, mengolok-olok

		Lihat	KWIC	Ayat
151	Itulah yang disebut melatah dan mengada-ngada , berpura-pura dan kehilangan harga diri.			MXD2012
152	Ustaz Syamil mengalihkan pandangan dan mencari-cari pemilik suara orang yang memanggil namanya.			MXD2012
153	9) Berhati-hatilah dari suka mengolok-olok terhadap cara berbicara orang lain, seperti orang yang t...			MXD2012
154	Satu hari masa bulan puasa sebab tak ada kerja punya pasal, aku merayau-rayau di halaman rumah sendiri.			MXD2012
155	24 Murka TUHAN yang menyala-nyala itu tidak akan surut sampai Ia telah melaksanakan dan...			MXD2012
156	Mereka telah membuat-buat cerita bahawa jika Kristus adalah seorang Tuhan yang I...			MXD2012
157	mang tidak dapat dinafikan, kadangkala ada mertua yang suka sangat mencari-cari kesalahan anak menantu.			MXD2012
158	an sebagainya maka apakah SPRM perlu berlari enjerit2 mencari Teoh merata-rata ?			MXD2012
159	Tapi, kenapa ada negara yang menutup-nutupi ?			MXD2012
160	idlan kita mencebik dan mempersenda orrang-orsng miskin yang tidak meminta-minta dengan gelar jumud, kolot, culas dan kotor.			MXD2012
161	Di ketika meninjau-ninjau itulah saya ternampak suami allahyarham, Pahamin s...			MXD2012
162	Jadi ingin saya kongis sedikit luahan rasa yang membuak-buak penasaran setiap kali Islam itu dihina.			MXD2012
163	Ini sedikit sebanyak ada kena-mengena dengan ajaran agama dan didikan keluarga.			MXD2012
164	Pada saya, ini bukan kerana media mengada-ngada , tetapi berasakan mereka sendiri perlu ikut dan diizinkan...			MXD2012
165	Dia membelek-belek wajahnya, macam tak pernah tengok muka sendiri pu...			MXD2012
166	Tentu fikrie sedang mencari-cari alasan untuk mengherdik-herdik dirinya sekali lagi.			MXD2012
167	Matanya melirik badan Ahmad dari atas kehawah sambil mengaleng-aleng kenalanya ala-ala Tubi Chawla bile nyanyi ngan Sh...			MXD2012

Kolokasi tanpa spesifikasi

Hapus Semua



Muncul ✓

Cari kolokasi di antara **Kiri 5** dan **Kanan 5**

Bentuk Jadian tanpa spesifikasi

Akar Kata tanpa spesifikasi

Prefiks tanpa spesifikasi

Sufiks tanpa spesifikasi

Konfiks tanpa spesifikasi

Reduplikasi tanpa spesifikasi

Example 3:

ingin 'to want' (+ word) + *di-* verb

Kata Kunci

Hapus Semua



Bentuk Jadian sama dengan "ingin"

Akar Kata tanpa spesifikasi

Prefiks tanpa spesifikasi

Sufiks tanpa spesifikasi

Konfiks tanpa spesifikasi

Reduplikasi tanpa spesifikasi

Example 3:

ingin 'to want' (+ word) + *di-* verb

Kolokasi dengan spesifikasi Hapus Semua ⊖

Muncul ✓

Cari kolokasi di antara **R1** dan **R2**

Bentuk Jadian tanpa spesifikasi

Akar Kata tanpa spesifikasi

Prefiks sama dengan "**di-**"

Sufiks tanpa spesifikasi

Konfiks sama dengan "**tanpa afiks**"

Reduplikasi tanpa spesifikasi

pesan promosi yang ingin disampaikan oleh perusahaan mereka

Lihat KWID Kalimat

464	Pada akhirnya, sistem perbankan syariah yang ingin diwujudkan oleh Bank Indonesia adalah perbankan syariah yang...	WEB2012
465	milihan bahan yang tepat dapat mengentalkan konsep minimalis yang ingin dihadirkan.	WEB2012
466	4) Pada bagian Browse, pilih lokasi dari file ISO yang ingin diinstall ke dalam flash disk.	WEB2012
467	Semakin saya dewasa, saya semakin menyadari bahwa jika seseorang ingin diubah, mereka harus mengubah pandangan mereka terhada...	WEB2012
468	sebaiknya dengarkan apa pendapatnya mengenai sasaran hidup yang ingin dicapai.	WEB2012
469	Tidak ada didalamnya kotoran riya dan ingin dikenal, atau tujuan duniawi dan pribadi, atau juga melakukan se...	WEB2012
470	Namun walau semandiri apapun, pria juga selalu ingin merasa dibutuhkan.	WEB2012
471	ia jika tidak, maka hal tersebut dapat mengaburkan pesan yang justru ingin disampaikan.	WEB2012
472	Bagaimana caranya pesan promosi yang ingin disampaikan oleh perusahaan mereka dapat tersampaikan namu...	WKP2016
473	a yaitu menentukan kegiatan-kegiatan apa atau event-event apa yang ingin dilaksanakan saat akan melaunching suatu barang atau jasa.	WKP2016
474	isa mengambil waktu seribu tahun untuk melakukan sesuatu yang kita ingin agar dilakukan dalam satu hari.	WKP2016
475	akabur melewati bawah pagar dan melarikan diri ke laut, tempat yang in	WKP2016
476	Rahasia apa yang in	WKP2016
477	AFB TV in	WKP2016
478	Soedirman sangat ingin menghindari kekerasan dan in	WKP2016
479	n Kiai Dullah yang sampai sekarang masih dalam proses realisasi yaitu in	WKP2016
480	ak pada pesan yang akan disampaikan, pada maksud dan tujuan yang ingin dicapai.	WKP2016

Halaman 10 dari 12 451 - 500 dari 578 Unduh

COMP-trace (Nomoto & Choi 2018)

... sesuatu yang kita **ingin** agar **dilakukan**
dalam satu hari.

sesuatu yang kita ingin [**agar** *t* di-lakukan
something REL we want so.that PASS-do
'something that we want to be done in a day'
(lit. *'something that we want that ___ is done
in a day')

Morphological search enables...

- Reference to abstract classes
e.g. “derivatives of of *fikir*”
- Morphosyntactic studies
The syntactic category of an affixed word is often predictable from the outermost affix in it.

cf. Korpus DBP and SEALang Library Corpora

- Only simple keyword search
- No support for RegEx (but * and ? in Korpus DBP)
- Search must be based on a particular lexical item, limiting possible corpus-based studies mostly to lexical ones.

[3] Contributions from users

- Currently, MALINDO Conc's corpora consists only of the reclassified version of the Leipzig Corpora Collection (Goldhahn et al. 2012; Nomoto, to appear).
- In the future, we will also include in the corpora, data collected by others as well as ourselves.
 1. Multilingual Spoken Corpus (Malay) (Shoho et al. 2005)
 2. David Moeljadi's Indonesian Frog Storytelling Corpus (Moeljadi 2014)
 3. Michael Ewing, František Kratochvíl, ...

To contribute your corpus

1. Publish (to become citable)
2. Get permission from the speakers/authors
OR take responsibility for their rights
3. Anonymize (strongly recommended)
4. Format (so computers can handle, ordinary people can type easily)
 - Text file (No Microsoft, ELAN, FLEX)
 - Avoid special characters (e.g. IPA)
 - No multiple punctuation marks (e.g. *iya:::*)

Morphological annotation

Morphological annotation using

- **MALINDO Morph** morphological dictionary (Nomoto et al. 2018)
https://github.com/matbahasa/MALINDO_Morph
- Ranking information for morphologically ambiguous tokens
- Manual disambiguation
 - *penanya* = (i) *peN-* + *tanya*, (ii) *pena* + *-nya*
 - *pelatih* (Malay) = (i) *peN-* + *latih*, (ii) *pe-* + *latih*

Annotated sentence part (XML file)

```
<w rt="ada" s1="-lah">
    Adalah</w>
<w rt="mudah">mudah</w>
<w rt="bagi">bagi</w>
<w rt="anak" r="R-penuh">
    anak-anak</w>
<w rt="yang">yang</w>
<w rt="sudah">sudah</w>
<w rt="biasa">biasa</w>
<w rt="didik" p1="ter-">
    terdidik</w>
```

```
<w rt="atas">atas</w>
<w rt="sikap">sikap</w>
<w rt="bakti" p1="ber-">
    berbakti</w>
<w rt="dan">dan</w>
<w rt="hormat" p1="meN-"
s1="-i">menghormati</w>
<w rt="dua" p1="ke-">
    kedua</w>
<w rt="ibu bapa" s1="-nya">
    ibubapanya</w>
```

Features not found in the Malay Concordance Project

1. Not only for English-speaking people.
 - User interface: Malay, Indonesian, English
 - Manual (in preparation): Malay, Indonesian, Japanese
2. Search results are downloadable (currently not working).

Both features are found with Korpus DBP, but not with SEALang Library Corpora.

References

- Goldhahn, Dirk, Thomas Eckart & Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*.
- Moeljadi, David. 2014. Usage of Indonesian possessive verbal predicates: A statistical analysis based on storytelling survey. *Tokyo University Linguistic Papers* 35: 155-176.
- Nomoto, Hiroki, Shiro Akasegawa, and Asako Shiohara. to appear. Reclassification of the Leipzig Corpora Collection for Malay and Indonesian. *NUSA*.
- Nomoto, Hiroki and Hannah Choi. 2018. The Apparent Lack of a Complementizer-trace Effect in Indonesian. ISMIL presentation.
- Nomoto, Hiroki, Hannah Choi, David Moeljadi and Francis Bond. 2018. MALINDO Morph: Morphological dictionary and analyser for Malay/Indonesian. Kiyooki Shirai (ed.) *Proceedings of the LREC 2018 Workshop "The 13th Workshop on Asian Language Resources"*, 36-43.
- Shoho, Isamu, Zaharani Ahmad, Hiroshi Uzawa, Hiroki Nomoto and Anida Saruddin. 2005. *Multilingual Spoken Corpora (Malay)*.

<https://malindoconc.lagoinst.info>
(temporary URL)

The development of MALINDO Conc was conducted under the JSPS grant “Program for Advancing Strategic International Networks to Accelerate the Circulation of Talented Researchers” offered to Tokyo University of Foreign Studies for a project entitled “A Collaborative Network for Usage-Based Research on Less-Studied Languages” as well as the JSPS Grant-in-Aid for Young Scientists (B) (#26770135). We are grateful to JSPS and Nanyang Technological University (NTU) for supporting the first author’s stay at NTU.