Pengembangan sumber bahasa digital dan konsep asas dalam linguistik Melayu/Indonesia

Hiroki Nomoto
Universitas Kajian Asing Tokyo
KOLITA 17, 10-12/04/2019

東京外国語大学
Tokyo University of Foreign Studies

Linguistik pada era digital

List Chart Collocates Compare KWIC

[POS]

Find matching strings Reset

Sections Texts/Virtual Sort/Limit Options

(HIDE HELP) NOT LOGGED IN

Same corpus - new location: English-Corpora.org

The Corpus of Contemporary American English (COCA) is the largest freely-available corpus of English, and the only large and balanced corpus of American English. COCA is probably the most widely-used corpus of English, and it is related to many other corpora of English that we have created, which offer unparalleled insight into variation in English.

The corpus contains more than 560 million words of text (20 million words each year 1990-2017) and it is equally divided among spoken, fiction, popular magazines, newspapers, and academic texts.

Click on any of the links in the search form to the left for context-sensitive help, and to see the range of queries that the corpus offers. You might pay special attention to the comparisons between genres and years and the (new) virtual corpora, which allow you to create personalized collections of texts related to a particular area of interest.

More help files

BRITISH NATIONAL CORPUS

About

What is the BNC?
Creating the BNC
BNC Products
Copyright
Contact Us
Contents A-Z

Using the BNC

What can I do with the BNC?
FAQ

Obtaining

How to download

Search the British National Corpus online

Various online services offer the possibility to search and explore the BNC via different interfaces. Some of the most notable are listed below:

- BNCWeb at Lancaster University (registration required - sign up here)
- BYU-BNC (Brigham Young University)
- Intellitext (University of Leeds)
- Phrases in English
- Audio BNC - access the digital audio files from the spoken corpus

Please note that we cannot answer queries about using any of these services, which are provided by other institutions. If you have a service for querying the BNC online, get in touch and we'll consider adding it to the list.

About the BNC

The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of British English, both spoken and written, from the late twentieth century. [more]

Here are some of the most popular links to information about the BNC:

Center for Corpus Development, NINJAL

At the NINJAL Center for Corpus Development, we create language corpora that aim to capture the breadth of the Japanese language.

日本語 | NINJAL

Corpora | Tools | Subscription | KOTONOHA Project | Reports

Please click the name of the corpus of interest.

- Balanced Corpus of Contemporary Written Japanese
- Corpus of Historical Japanese
- Corpus of Spontaneous Japanese
- NINJAL Web Japanese Corpus
- International Corpus of Japanese As a Second language -In Japanese Only-
- Nagoya University Conversation Corpus
- Gen-Nichi-Ken Corpus of Workplace Conversation
- Corpus of Everyday Japanese Conversation
- Corpus of Modern Japanese -In Japanese Only-
- Corpus Annotation -In Japanese Only-

How to Apply

Try our online corpus tools!

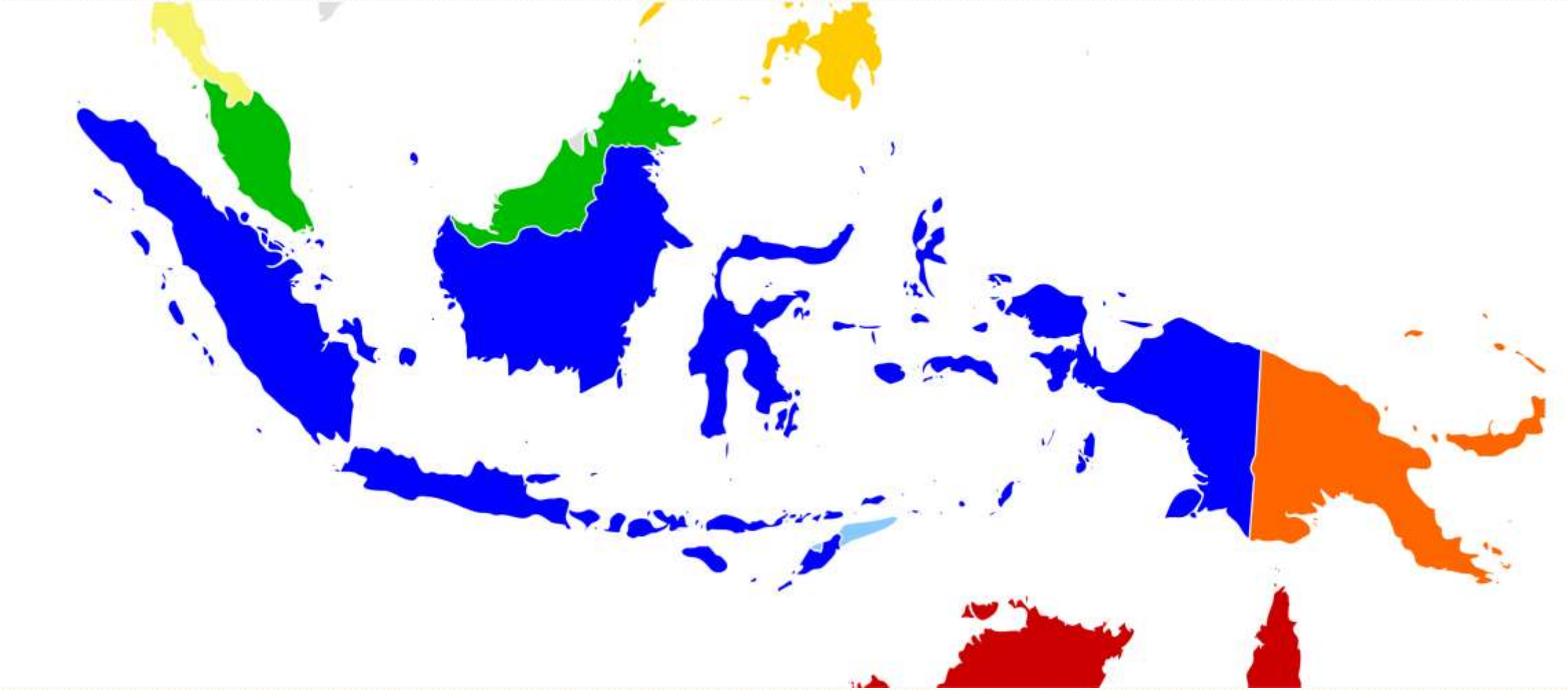
No Registration — Shonagon

Registration Required — Chunagon

UniDic
Morphological Analysis Dictionary

An electronic dictionary which segments Japanese texts into words, and allows for morphological analysis.

Korpus bahasa Melayu/Indonesia



Selamat datang!

Ini adalah laman resmi pencarian kata dalam Korpus Indonesia (Koin)

Pilihan menu pencarian:

Frekuensi	Kalimat	Kolokat	Kelas Kata

Korpus DBP - Penyelidik

Dewan Bahasa dan Pustaka

Carian Konkordans | Carian Konkordans (Lama) | Analisis Kata | Statistik Bahan | Senarai Bahan | Muatnaik Bahan | Panduan | Log Keluar

Carian Konkordans

Kata [ ] Cari

Sumber Bahan | Tarikh/Tahun Terbit Bahan | Kriteria Penulis

Jenis Bahan

- ☐ Akhbar (115,530)
- ☐ Buku (675)
- ☐ Majalah (1,981)
- ☐ Efemeral (36)
- ☐ Teks Sastera (708)
- ☐ Kertas Kerja (129)
- ☐ Perbahanan

Tajuk [ ]

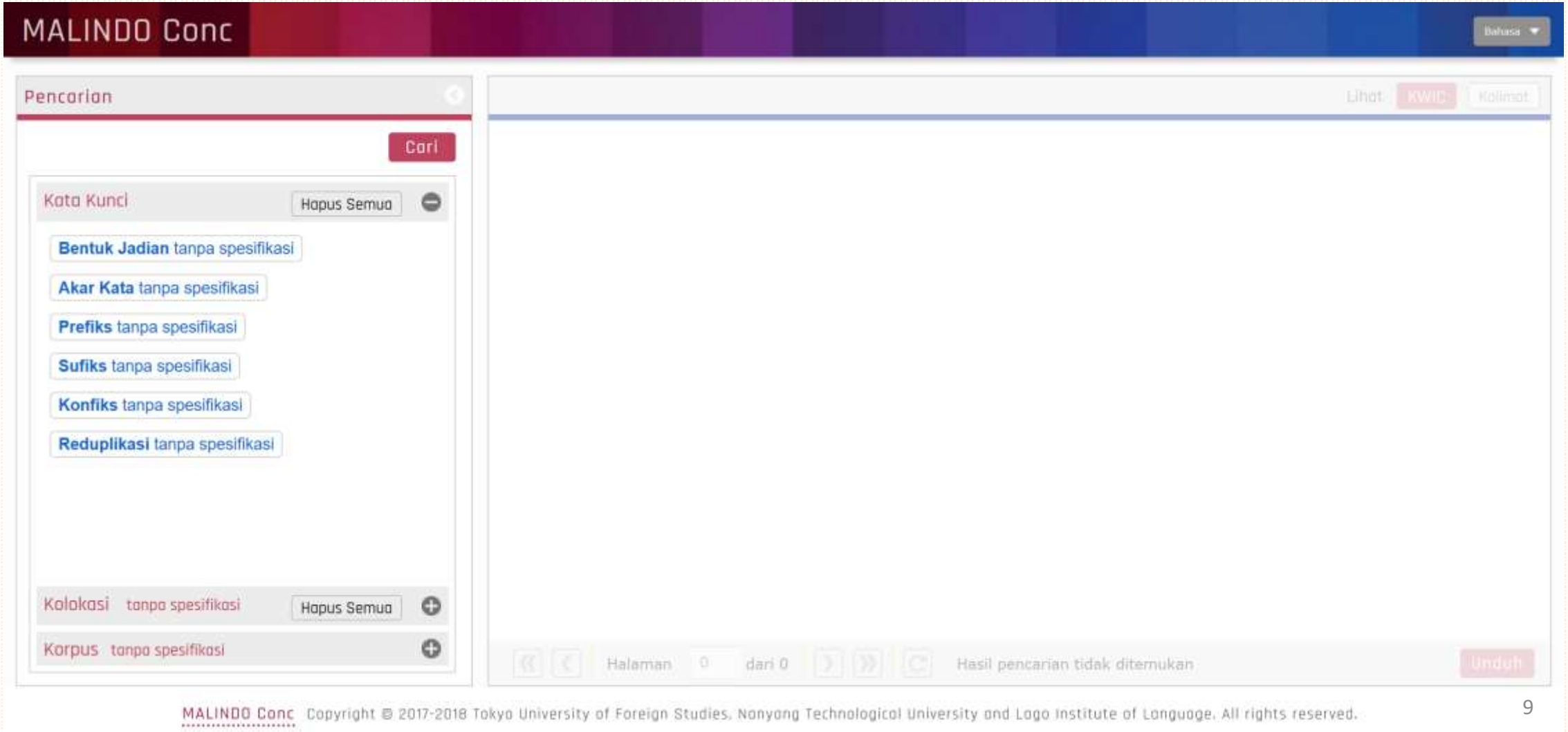
Tajuk penuh atau sebahagian. Pisahkan nama-nama dengan koma (,) untuk memasukkan lebih dari satu nama.

Bidang [-- Semua Bidang --]

Konteks ◉ Hadkan lingkungan konteks [8] kata ke kiri dan ke kanan.

Bilangan Output ◉ [100] kata ○ [0.1] % ○ Semua

MALINDO Conc (Nomoto dkk. 2018c)
https://malindo.aa-ken.jp/conc/



Organisasi presentasi ini

- Memperkenalkan MALINDO Conc
- Dua macam persediaan untuk mengembangkan MALINDO Conc dan tantangan yang kami hadapi
  1. Korpus yang dijadikan data MALINDO Conc

     → Masalah identifikasi bahasa
  2. Anotasi morfologi

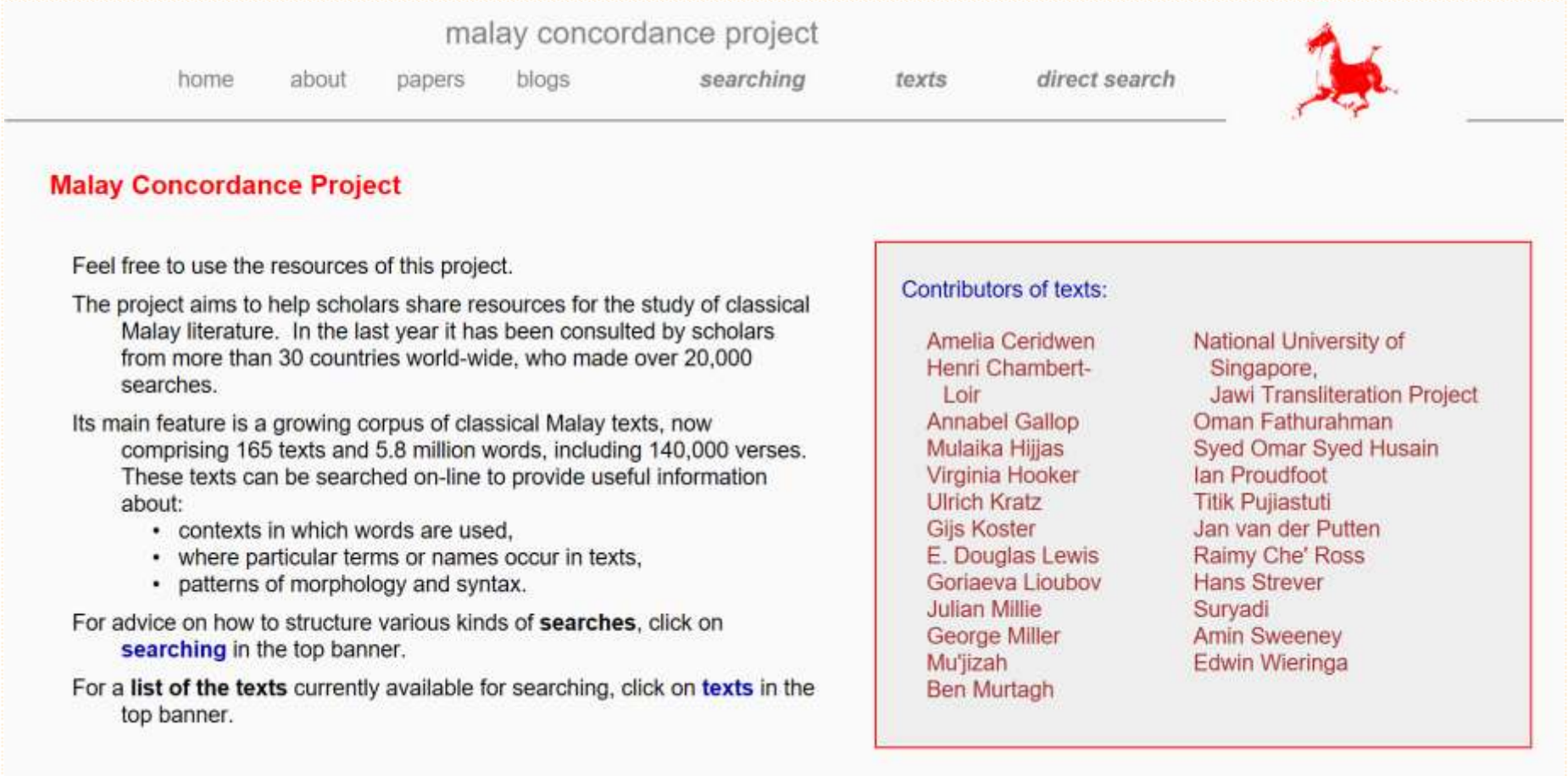
     → Alat penganalisis yang ada kurang memuaskan untuk tujuan linguistik

MALINDO Conc

Ciri-ciri MALINDO Conc

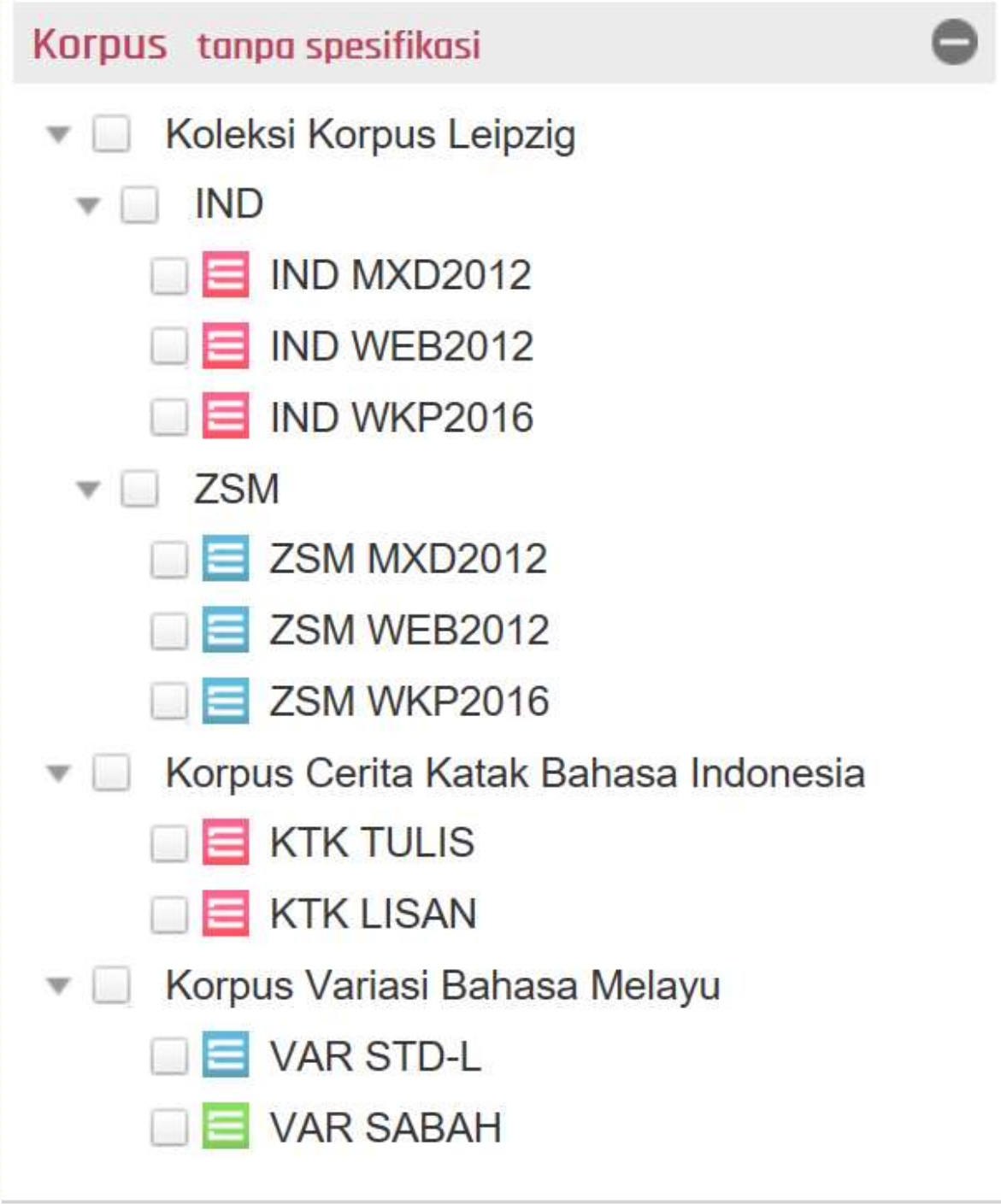
- Sistem pencarian, khususnya pengkonkordans (*concordancer*)
- Dikembangkan sebagai alat untuk semua peneliti linguistik Melayu/Indonesia
- Lintas variasi: bukan hanya Indonesia, bukan hanya Malaysia
- Gratis
- Mudah untuk diguna
- Tetapi bisa melakukan pencarian yang cukup baik
  - Pencarian morfologis
  - Kolokasi

Teladan MALINDO Conc



[1] Lintas variasi

- MALINDO Conc coba menargetkan berbagai variasi bahasa Melayu di Nusantara.
- Sistem lain
  - KOIN: Indonesia
  - Korpus DBP: Malaysia
  - SEALang Library Corpus (Malay): Malaysia, Singapura, Brunei
  - SEALang Library Corpus (Indonesian): Indonesia



Indonesia

Malaysia, Singapura, Brunei

Sabah, Malaysia

ı Radliyallaahu anhu bahwa seorang	**budak**	perempuan hitam mempunyai tenda di dalam masjid, ia sering d···	MXD2012
ın yang masih sendirian, begitu pula	**budak**	laki-laki dan budak perempuan kalian yang shaleh, jika mereka fa···	MXD2012
irian, begitu pula budak laki-laki dan	**budak**	perempuan kalian yang shaleh, jika mereka faqir maka allah akan···	MXD2012
yang lebih dari sekedar majikan dan	**budak**	.	MXD2012
ala kondisi kecuali terhadap istri dan	**budak**	perempuannya.	MXD2012
ı oleh Abu Lukluk (Fairuz) , seorang	**budak**	pada saat ia akan memimpin salat Subuh.	MXD2012
5. Riqab, adalah	**budak**	yang ingin memerdekakan diri dengan membayar uang tebusan.	MXD2012
kepada setiap perintah Allah alias Si	**Budak**	Allah.	MXD2012
	Budak	menangih, bapak makan yang tak sudah, mak sibuk on handpho···	MXD2012
Baik biarkan jek	**budak**	tu sebab kalau tolong and jadi apa2 yang lebih teruk kat budak t···	MXD2012
g and jadi apa2 yang lebih teruk kat	**budak**	tu, diorang yang kena tanggung.	MXD2012
Masa tu	**budak**	tu berada di tgh2 antara tayar depan dan belakang.	MXD2012
	Budak	nak berlari, mencuri hatta nak membunuh sekalipun, hanya ward···	MXD2012
Setibanya di rumah	**budak**	itu, kawan kita ni menekan loceng pintu.	MXD2012
Apa lah malang nasib	**budak**	perempuan tu.	MXD2012
O. klah, aku maleh nak gaduh ngan	**budak**	hingusan cam ko ni (jalak tua, menunjukkan eksen nyer) so apa···	MXD2012

[2] Pencarian morfologis

Korpus dapat dicari dengan informasi jenis afiksasi dan reduplikasi seperti

- Verba *di-* diikuti verba *meN-*
- Bentuk-bentuk infleksi untuk *pikir* dan *pikirkan*
- Verba *ber-…-kan*
- Verba *meN-X-X* & *X-meN-X*
- *ingin* + verba *di-* & *ingin* + kata (cth. *untuk*) + verba *di-*

MALINDO Conc

Kata Kunci

Hapus Semua

Bentuk Jadian tanpa spesifikasi

Akar Kata tanpa spesifikasi

Prefiks tanpa spesifikasi

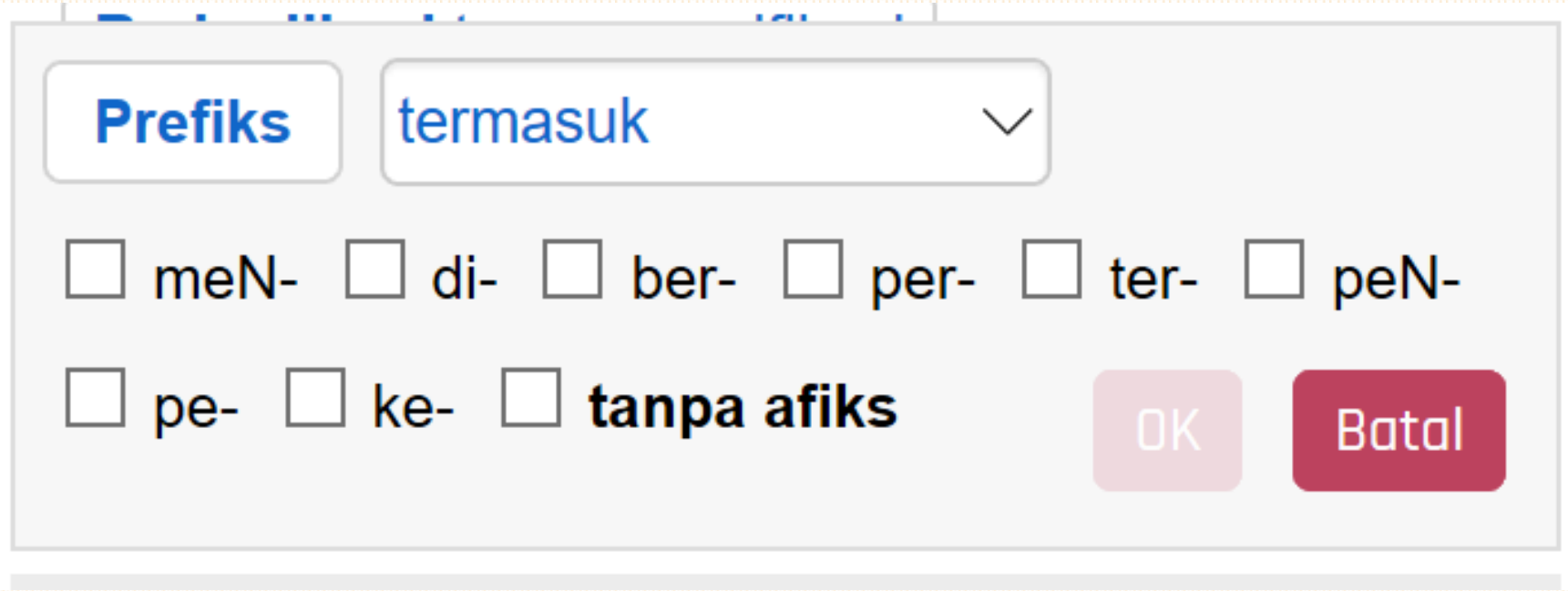
Sufiks tanpa spesifikasi

Konfiks tanpa spesifikasi

Reduplikasi tanpa spesifikasi

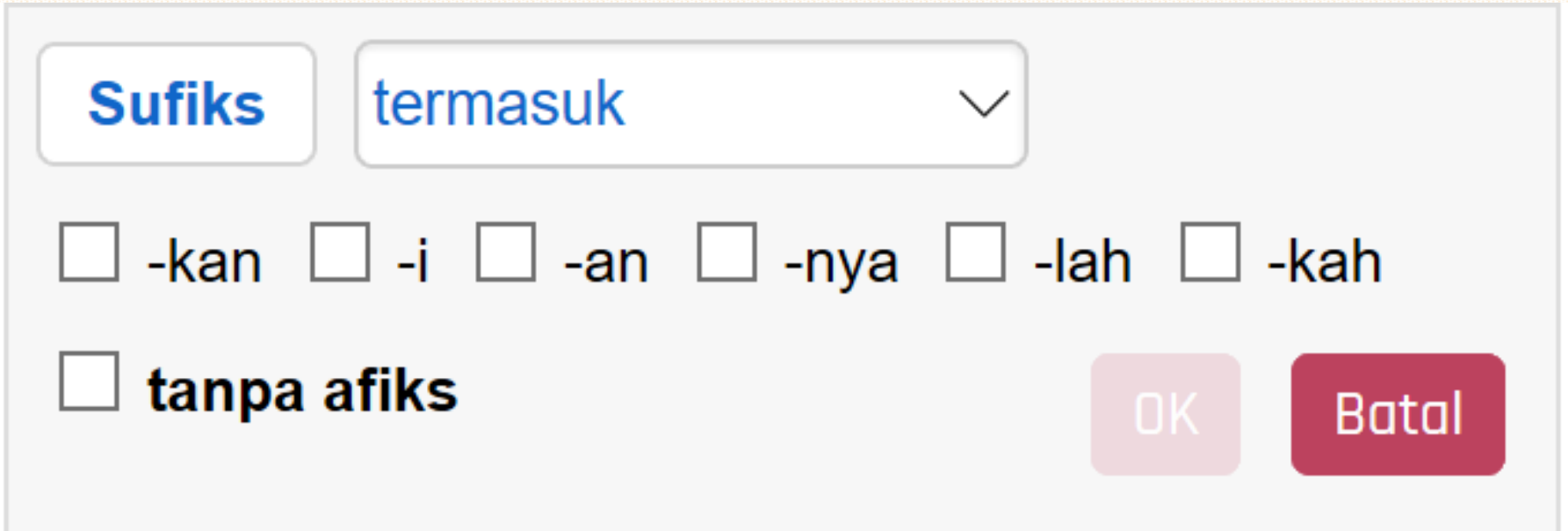
MALINDO Conc

Kata Kunci > Prefiks



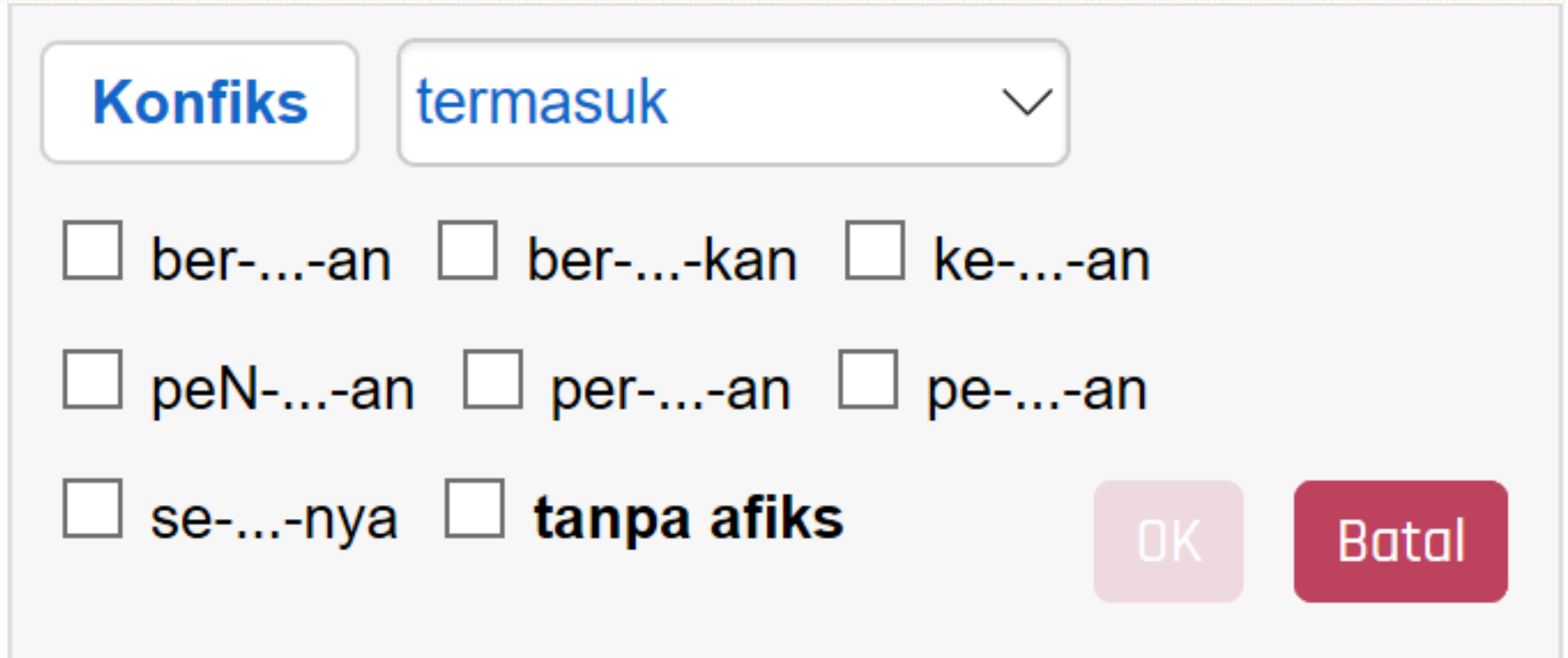
MALINDO Conc

Kata Kunci > Sufiks



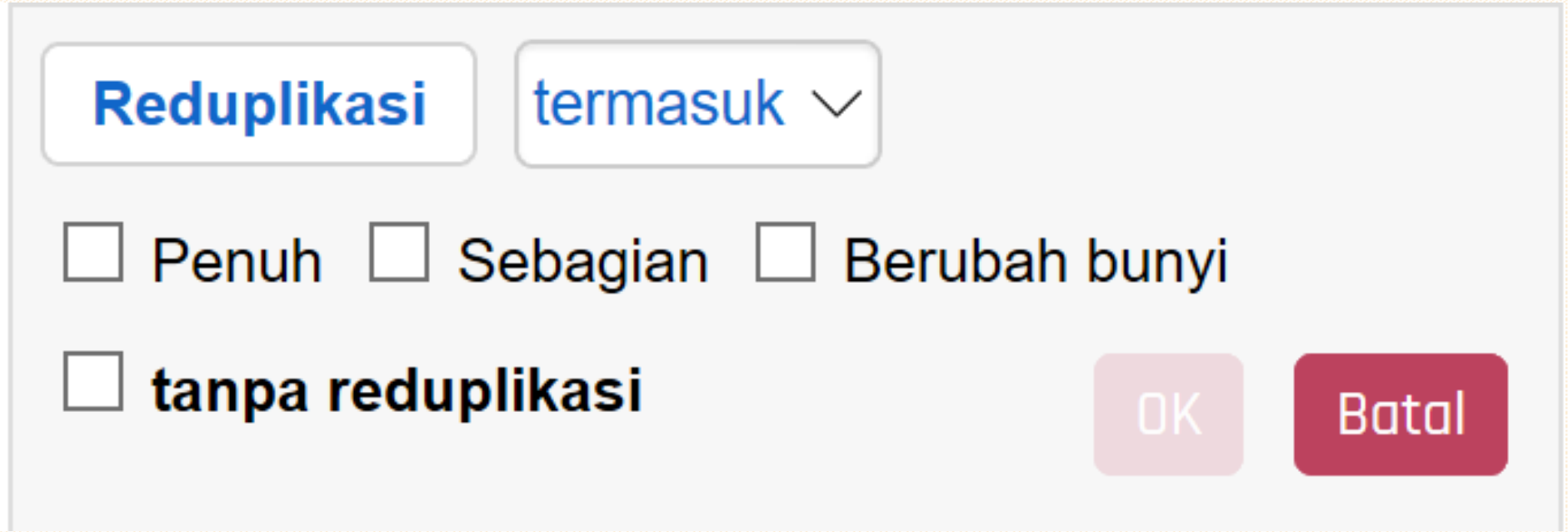
MALINDO Conc

Kata Kunci > konfiks



MALINDO Conc

Kata Kunci > Reduplikasi



Contoh 1:
Bentuk-bentuk infleksi untuk *pikir/pikirkan*

Pikir, memikirkan, dipikir, pikirkan, dipikirkan…

14	oleh seorang yang penuh keyakinan akan sangat mempengaruhi pola **pikir** pembacanya.	WEB2012
15	Mbak Tina yang lebih banyak **memikirkan** masalah produksi.	WEB2012
16	Padahal jika **dipikir** dengan bahasa komunikasi, foto tersebut jelas-jelas menggamba…	WEB2012
17	Lawan kerakusan APINDO yg hny **memikirkan** keuntungan perusahaan tanpa memikirkan kesehatan gene…	WEB2012
18	kerakusan APINDO yg hny memikirkan keuntungan perusahaan tanpa **memikirkan** kesehatan generasi mdatang! mungkin mereka tdk dpt ASI…	WEB2012
19	ıt ' sulit' diterima akal sehat atau malah terkesan mengada-ada, maka **pikirkan** lagi alasan tersebut.	WEB2012
20	Harus **dipikirkan** juga, jika sudah mengunduh, berapa biaya untuk memperba…	WEB2012
21	'ang telah banyak mencuri masa yang sepatutnya saya gunakan untuk **memikirkan** apa yang seharusnya saya buat demi manfaat ummat.	WEB2012
22	Cuma **pikirkan** satu hal, saya pun memikirkannya, bahwa kita akan mati.	WEB2012
23	Cuma pikirkan satu hal, saya pun **memikirkannya** , bahwa kita akan mati.	WEB2012
24	Aku **pikir** mungkin di arena pacuan kuda sama saja soalnya dari kamarku jel…	WEB2012
25	Setelah aku **pikir-pikir** lagi ternyata sepertinya aku tak mampu untuk pindah ke kam…	WEB2012
26	Padahal, jika **dipikirkan** dengan serius, manakah yang lebih hebat kekuatannya, apak…	WEB2012
27	Tidak **memikirkan** kandungan makna Al-Qur` an dan hadits, karena akan me…	WEB2012
28	Abul Fadl kembali bertanya, Coba **pikirkan** dan ingat-ingatlah sebuah amal yang kamu kerjakan ikhlas kar…	WEB2012
29	Pola **pikir** semacam ini adalah pemikiran dua kelompok sempalan Islam yaitu…	WEB2012

Contoh 2:
Reduplikasi dengan *meN-*

Mengibas-ngibaskan, mengaku-ngaku, menyapu-nyapu, mengada-ada…

363	Dengan gerakan demonstratif sambil terbatuk-batuk tangannya	**mengibas-ngibaskan**	asap rokok yang dikepulkan sepasang kekasih per··· WEB2012
364	Guncangan terasa kuat dan	**terus-menerus**	. WEB2012
365	Sayangnya, dia bertemu dengan orang-orang yang	**mengaku-ngaku**	saja. WEB2012
366	Dan yg lebih memprihatinkan, Megawati dalam pidatonya selalu	**mengagung-agungkan**	Soekarno, padahal dulu ibunya dibuat sakit hati··· WEB2012
367	Langsung kubimbing tangannya untuk	**mengelus-elus**	dan mengurut seluruh bagian penis dan kedua bijinya. WEB2012
368	;an batang kemaluannya menyumpal mulutku sambil sesekali lidahnya	**menyapu-nyapu**	dinding vulvaku. WEB2012
369	, tanyanya bingung sambil tetap	**mengelus-elus**	batang kejantananku. WEB2012
370	di suatu kekuasaan yang penuh setelah kerajaan barbar memecah dan	**membagi-bagi**	daerah kekuasaan Romawi. WEB2012
371	Dengan penuh ambisi, ia	**mencari-cari**	orang yang mengikuti ajaran Tuhan Yesus untuk ditangka··· WEB2012
372	Pergumulan politik saat itu telah	**mencabik-cabik**	Eropa. WEB2012
373	Sambil terisak-isak bahagia, Ningsih memeluk tubuhku dan	**mengelus-elus**	punggungku. WEB2012
374	Sementara itu, kegiatan	**masak-memasak**	dilakukan di dapur kotor atau yang biasa disebut area··· WEB2012
375	Mak Asiah	**memuji-muji**	saya sebagai seorang anak yang berbudi, Cuma ketika be··· WEB2012
376	ng yang terjun dalam penelitian ilmiah) Saya rasa anda terlalu banyak	**mengulang-ulang**	kata absolut. WEB2012
377	Dia mengatakan jika pengakuan Kemad adalah palsu dan	**mengada-ada**	. WEB2012
378	Perlu digarisbawahi, upaya yang kami lakukan bukanlah untuk	**mencari-cari**	kesalahan pihak lain atau aparat penegak hukum. WEB2012

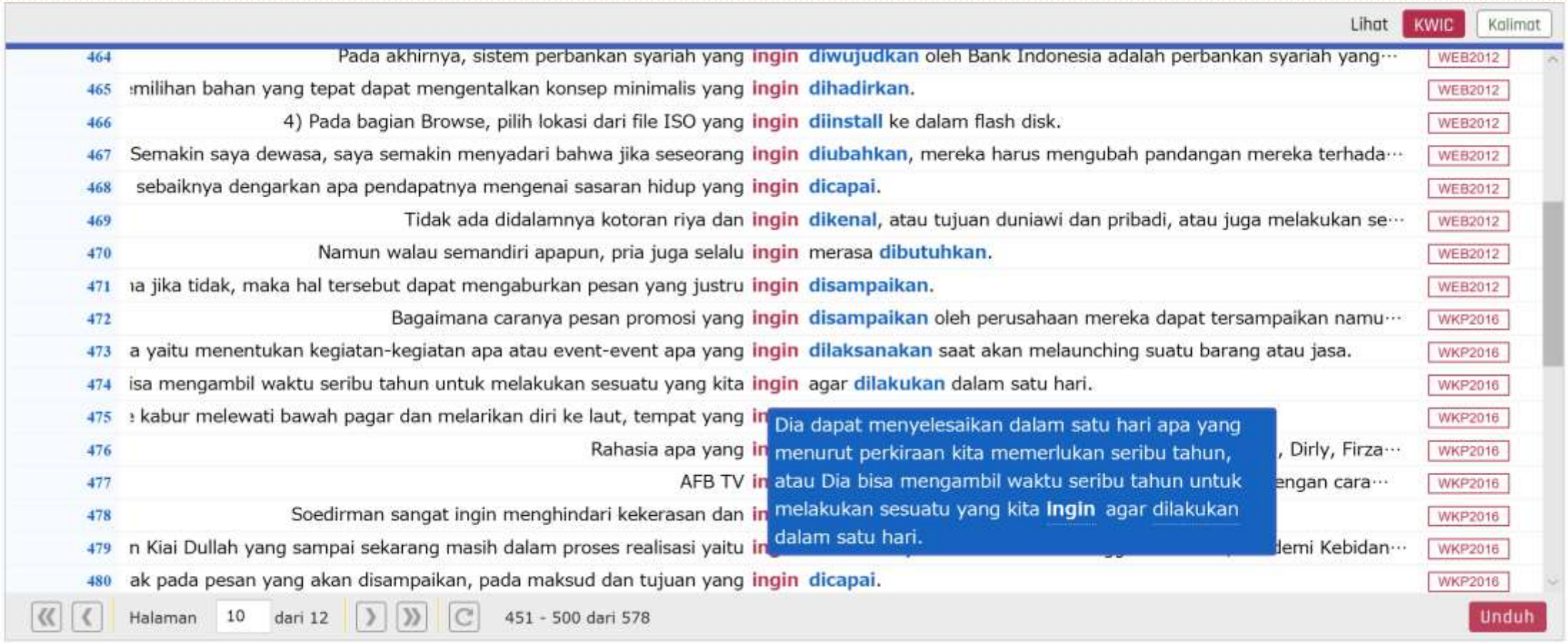
MALINDO Conc

Kiri 5	Kiri 4	Kiri 3	Kiri 2	Kiri 1	Kata Kunci	Kanan 1	Kanan 2	Kanan 3	Kanan 4	Kanan 5

Contoh 3: *ingin* (+ kata) + verba *di-*

Kiri 5	Kiri 4	Kiri 3	Kiri 2	Kiri 1	ingin +	di-	di-	Kanan 3	Kanan 4	Kanan 5

pesan promosi yang ingin disampaikan oleh perusahaan mereka



Komp-tikas (Nomoto & Choi 2018)

… sesuatu yang kita ingin agar dilakukan dalam satu hari.

<u>Kesan komplementiser-tikas (*Complementiser-trace effect*)</u>

*[$_{CP}$ Komp *t* …

(1) a. Who do you believe [$_{CP}$ *t* married Naomi]?

b. *Who do you believe [$_{CP}$ that *t* married Naomi]?

(2) sesuatu yang kita ingin [$_{CP}$ agar *t* dilakukan dalam satu hari].

Dengan pencarian morfologis, kita bisa…

- Mengacu pada kelas-kelas abstrak
  cth. “kata-kata terbitan untuk *pikir*”
- Studi morfosintaksis
  Kategori sintaksis biasanya bisa diprediksi berdasarkan afiks yang di paling luar

cf. Sistem lain

- Hanya pencarian kata kunci sederhana saja
- Tidak bisa menggunakan RegEx (kecuali * dan ? dalam Korpus DBP)
- Pencarian mesti berdasarkan item leksikal tertentu.
  → Penelitian korpus terbatas ke penelitian leksikal.

Bagaimanakah MALINDO Conc dibuat?

1. Datanya dari mana?
2. Mengapa bisa melakukan pencarian morfologis?

Dua macam persediaan & tantangannya

1. Korpus yang dijadikan data MALINDO Conc

   → Masalah identifikasi bahasa

2. Anotasi morfologi

   → Alat penganalisis yang ada kurang memuaskan untuk tujuan linguistik

[1] Data MALINDO Conc

- Data mesti besar, sekurang-kurangnya 1 milyar token.
- Bebas isu hak cipta.
- Dalam tempoh waktu dan uang anggaran proyek

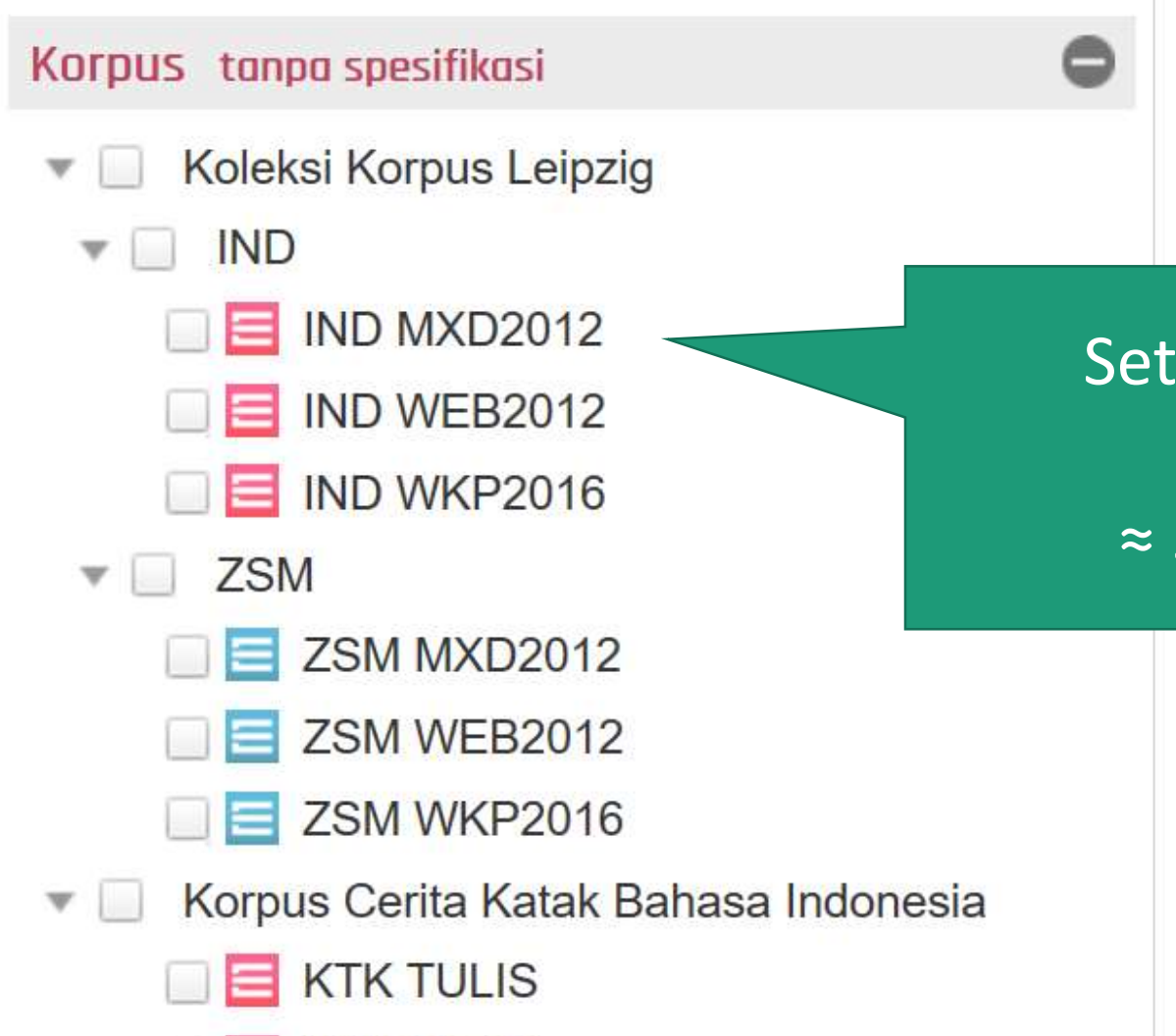
→ Korpus web

(= korpus yang menggunakan data yang dikumpulkan dari situs web)

Ukuran korpus utk bahasa Melayu/Indonesia

Alat	Ukuran (token)	Korpus
Malay Concordance Project	5,7 juta	Karya sastera klasik Melayu
KOIN	5,5 juta	Artikel ilmiah
Korpus DBP	135 juta	Data sendiri
SEAlang Malay	2,5 juta	An Crúbadán (korpus web)
SEAlang Indonesian	5 juta	

Data dari Koleksi Korpus Leipzig



Setiap subkorpus mempunyai
300 ribu kalimat
≈ 5.847 ribu token (> KOIN)

Koleksi Korpus Leipzig (Goldhahn dkk. 2012)

- http://corpora.uni-leipzig.de/
- Sekumpulan korpus web ekabahasa dengan sebanyak 236 bahasa
- Dikembangkan oleh Jurusan Pemrosesan Bahasa Alami, Fakultas Matematika dan Sains Komputer, Universitas Leipzig
- Boleh diundur secara gratis dan tanpa pendaftaran, dengan ukuran maksimnya 3 juta patah kata

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...

2012		10K	30K	100K	300K	1M
					Newscrawl	
Year	**Country**	**Downloads**				
2011		10K	30K	100K	300K	1M
2012		10K	30K	100K	300K	1M
2015		10K	30K	100K	300K	1M
2016		10K	30K	100K	300K	1M
					Web	
Year	**Country**	**Downloads**				
2011		10K	30K	100K	300K	1M
2012		10K	30K	100K	300K	1M
2013	Indonesia	10K	30K	100K	300K	1M
2015	Brunei	10K	30K	100K	300K	1M
2015	India	10K	30K	100K	300K	1M
2015	Indonesia	10K	30K	100K	300K	1M
2018	com	10K	30K	100K	300K	1M

Serius??

Masalah identifikasi bahasa

- Terdapat sekian banyak kesalahan identifikasi bahasa dalam subkorpus bahasa Melayu dan bahasa Indonesia
- Nomoto dkk. (2018a): menyusun kembali Koleksi Korpus Leipzig dengan membetulkan kesilapan identifikasi bahasanya

http://ms.wikipedia.org/wiki/Adi_dan_Ayah

1. Adi yang cerdas ini begitu mengidolakan sang ayah yang seringkali berlaku konyol dan kikuk, tetapi ia selalu menganggap Ayahnya adalah Ayah terbaik dan terhebat di dunia ini.
2. Karena menurut Adi sang Ayah pasti selalu lebih dari ayah-ayah lainnya, maka mau tidak mau sang Ayah harus bisa melakukan kehebatan-kehebatan yang ingin dipamerkan Adi kepada tetangganya, Dana dan Dini, dan Bertha, Ibu mereka.
3. Keinginan-keinginan Adi kebanyakan dikarenakan ulah dari Dana dan Dini, tetangga Adi yang juga merupakan anak dari Bos ayah Adi, dimana mereka selalu pamer dan membandingkan antara Ayah Adi dengan Ayah mereka.
4. Kelucuan-kelucuan di setiap episodenya muncul saat bagaimana usaha si Ayah dengan sekuat tenaga untuk dapat memenuhi semua keinginan Adi, yang terkadang sepertinya tidak masuk di akal.
5. Lewat serial ini penonton akan melihat hubungan menarik antara seorang anak yang begitu dekat dengan Ayahnya.
6. Serial Adi dan Ayah adalah sebuah drama komedi keluarga yang mengangkat kisah mengenai hubungan antara seorang anak laki-laki bernama Adi dengan sang Ayah.

Hasil reklasifikasi (satuan: token)

Layak sebagai data MALINDO Conc

Bahasa	Kode asal	Melayu (zsm)	Indonesia (ind)	Tidak pasti (msa)
Melayu	msa	17.719.080	687.212	1.272.241
	ind-bn	222.670	1.619	0
	Jumlah	17.941.750	688.831	1.272.241
Indonesia	ind	28.443.247	1.110.083.452	3.653.346
	ind-id	347.935	330.870.557	0
	Jumlah	28.791.182	1.440.954.009	3.653.346

[2] Anotasi morfologi (fail XML)

```
<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>
```

Masalah penganalisis morfologi yg sudah ada

- Alat-alat yang dikembangkan oleh para peneliti pemrosesan bahasa alami berdasarkan pengertian konsep dasar linguistik yang kurang tepat.
- Pembedaan konfiks dari gabungan prefiks+sufiks
  - MorphInd (Larasati dkk. 2011)

    pengiriman → ^peN+kirim<v>+an_NSD$ --- konfiks atau prefiks+sufiks?
  - *meN-…-kan* salah dianggap sebagai konfiks

MALINDO Morph (Nomoto dkk. 2018b)

- Kamus morfologi yang merupakan daftar
  - Akar kata (*root*)
  - Bentuk jadian (*surface form*)
  - Prefiks
  - Sufkis
  - Konfiks
  - Jenis reduplikasi
- Skrip penganalisis sendiri + pemeriksaan hasil analisis otomatis secara manual
- https://github.com/matbahasa/MALINDO_Morph

Pemeriksaan manual sangat mahal tapi perlu

- Kasus ambiguitas morfologis
  1. *penanya*

     (i) *peN-* + *tanya* (ii) *pena* + *-nya*
  2. *pelatih* (bahasa Melayu)

     (i) *peN-* + *latih* (ii) *pe-* + *latih*
- *Mereka* (*mereka* vs. *meN-* + *reka*) tidak diperiksa secara manual karena jumlahnya yang terlalu banayk.

Center for Corpus Development, NINJAL

At the NINJAL Center for Corpus Development, we create language corpora that aim to capture the breadth of the Japanese language.

日本語 | NINJAL

Corpora | Tools | Subscription | KOTONOHA Project | Reports

Please click the name of the corpus of interest.

- Balanced Corpus of Contemporary Written Japanese
- Corpus of Historical Japanese
- Corpus of Spontaneous Japanese
- NINJAL Web Japanese Corpus
- International Corpus of Japanese As a Second language -In Japanese Only-
- Nagoya Univer... Conversation C...
- Gen-Nichi-Ken Corpus of Workplace Conversation
- Corpus of Everyday Japanese Conversation
- Corpus of Modern Japanese -In Japanese Only-
- Corpus Annotation -In Japanese Only-

How to Apply

Try our online corpus tools!

Kamus morfologi oleh Badan Bahasa Jepang

UniDic
Morphological Analysis Dictionary

An electronic dictionary which segments Japanese texts into words, and allows for morphological analysis.

Penambahan bentuk dasar (*stem*) dan lema

- Sudah banyak "stemmer" dan "lemmatizer" untuk bahasa Melayu/Indonesia yang dikembangkan oleh para peneliti di bidang teknik.
- Meski demikian, hasil analisisnya ternyata tidak selalu bentuk dasar atau lema.

Sastrawi stemmer

- https://github.com/sastrawi/sastrawi
- Tidak menghasilkan bentuk dasar (*stem*) tetapi akar kata (*root*).

  Cth.

  menyuarakan → suara

  bersuara → suara

- Sebenarnya, bukan STEMmer tetapi ROOTer.

MorphInd (Larasati dkk. 2011)

- Lema untuk *kirim* → kirim

  Lema untuk *mengirim* → mengirim

- Tetapi *kirim* dan *mengirim* bukan dua kata berlainan seperti *kucing* dan *ayam*.

- Lema bagi kedua kata *kirim* dan *mengirim* mesti sama.

Konsep ‘bentuk dasar (*stem*)’

- **<u>Bentuk dasar</u>**: bentuk yang menjadi dasar untuk proses morfologi
- Bahasa Inggris (bahasa isolatif)
  eats = eat + -s *eat* = bentuk dasar untuk sufiksasi *-s*
- Bahasa Jepang (bahasa aglutinatif)
  tabe-rare-ta ‘telah dimakan’
  1. tabe-rare *tabe* = bentuk dasar untuk sufiksasi *-rare*
  2. tabe-rare-ta *tabe-rare* = bentuk dasar untuk sufiksasi *-ta*
- Bahasa Indonesia (bahasa aglutinatif)
  1. suara-kan *suara* = bentuk dasar untuk sufiksasi *-kan*
  2. meny-[s]uara-kan *suara-kan* = bentuk dasar untuk prefiksasi *meN-*

➢ Bentuk dasar (*stem*) ≠ “kata dasar”/akar kata (*root*)

Pilih bentuk dasar yang mana untuk MALINDO Morph?

1. *suara* = bentuk dasar untuk sufiksasi *-kan* (derivasi)
2. *suara-kan* = bentuk dasar untuk prefiksasi *meN-* (infleksi)

- Bentuk dasar untuk infleksi lebih berguna.
- *mengakui* → *aku* atau *akui*?

  Hasil pencarian Google untuk *mengakui* perlu mengandung contoh-contoh kata *aku*?

Infleksi untuk *suarakan*

suarakan	menyuarakan	disuarakan
• aktif kosong • kalimat perintah (aktif) • pasif kosong (“semu”)	• aktif morfologis	• pasif morfologis • kalimat perintah (pasif)

Konsep ‘lema’

- **Lema**: bentuk wakil untuk sekelompok bentuk kata yang berkaitan
- *kirim* & *mengirim*: wakilnya yang mana?
- Pilih yang biasa untuk penutur asli

  → bentuk *meN-*

  (Untuk penutur asing, bentuk kosong mungkin lebih baik.)

Akar kata (*root*), bentuk dasar (*stem*) dan lema: Contoh (1)

Bentuk jadian	Akar kata (*root*)	Bentuk dasar (*stem*)	Lema
menyuarakan	*suara*	*suarakan*	*menyuarakan*
disuarakan	*suara*	*suarakan*	*menyuarakan*
suarakan	*suara*	*suarakan*	*menyuarakan*
suara	*suara*	*suara*	*suara*

Akar kata (*root*), bentuk dasar (*stem*) dan lema: Contoh (2)

Bentuk jadian	Akar kata (*root*)	Bentuk dasar (*stem*)	Lema
membukukan	*buku*	*bukukan*	*membukukan*
dibukukan	*buku*	*bukukan*	*membukukan*
bukukan	*buku*	*bukukan*	*membukukan*
buku	*buku*	*buku*	*buku*
buku-buku	*buku*	*buku*	*buku*

“Kata dasar”, “kata akar”

- Tidak semua akar kata (*root*) adalah kata.
- Bahasa Inggris
  receive (prefiksasi *re-*; *ceive* bukan morfem bebas, yaitu bukan “kata”)
- Bahasa Indonesia
  anai-anai (reduplikasi penuh; *anai* bukan morfem bebas atau “kata”)
- Istilah “kata akar” (dan “root word”) bermasalah.
- Apakah *anai-anai* kata dasar? Kata akar?
  → Jika ya, “kata dasar” kadang-kadang sama dengan akar kata (*root*) dan kadang-kadang tidak… Apa itu sebenarnya?

Peneliti teknik mempercayai ahli linguistik

Sastrawi

Sastrawi is a simple PHP library which allows you to reduce inflected words in Indonesian Language (Bahasa Indonesia) to their base form (stem). Despite its simplicity, this library is designed to be high quality and well documented. For more information in english, see README.

Development	Master	Releases	Statistics
build passing · coverage 96 % · Scrutinizer 9.59	build passing	stable v1.2.0	downloads 36.62 k

Stemming

Stemming adalah proses mengubah kata berimbuhan menjadi kata dasar. Contohnya:

- menahan => tahan
- berbalas-balasan => balas

Stemming adalah proses mengubah kata berimbuhan menjadi **kata dasar**.

Kita yang bertanggungjawab memastikan….

- Teknologi bahasa untuk bahasa Melayu/Indonesia berkembang berlandaskan pengertian konsep linguistik dasar yang saksama.
- Data yang diperlukan untuk pengembangan tenknologi bahasa untuk bahasa Melayu/Indonesia disediakan untuk para peneliti bidang teknik secara terbuka dan dalam bentuk yang mudah diguna.
- Pastikan penutur bahasa Melayu/Indonesia dapat menikmati teknologi yang dinikmati oleh penutur bahasa Inggris.

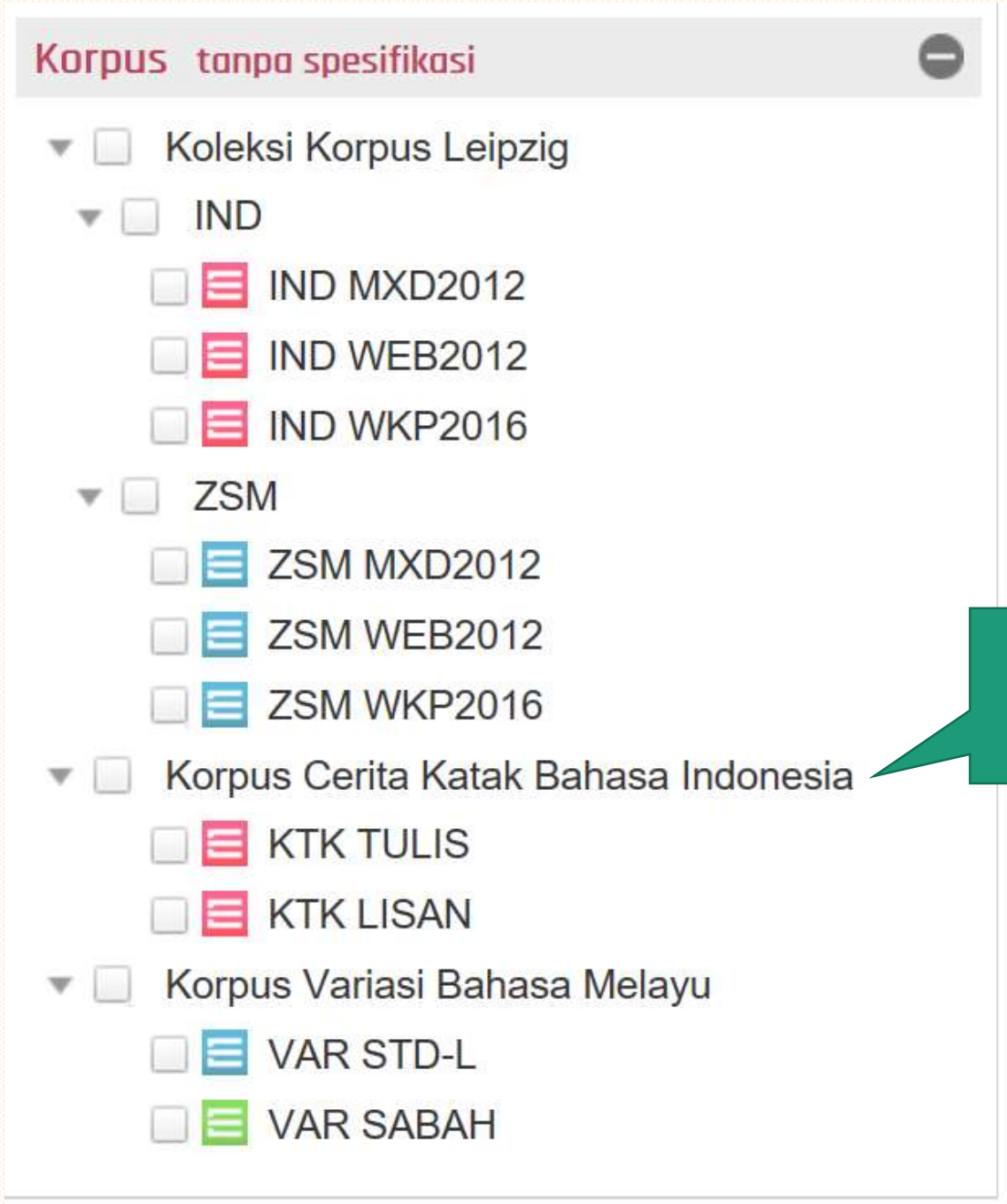
Simpulan

- Sistem pencarian korpus MALINDO Conc 

- Kamus morfologi MALINDO Morph

https://malindo.aa-ken.jp/

- Versi reklasifikasi Koleksi Korpus Leipzig: sudah dihantar ke tim Leipzig, tetapi belum diunggah (hubungi saya kalau benar-benar maukan)
- Anotasi morfologi (data boleh diguna melalui kolaborasi dengan kami)
- Sumbangan dari kolega
  - Ruang server (data bahasa Indonesia lebih banyak tetapi uang untuk meminjam ruang server tidak cukup)
  - Data korpus



Korpus Cerita Kata Bahasa Indonesia
Disumbangkan oleh David Moeljadi

Format: Teks (.txt) tanpa simbol IPA
Tidak boleh: Microsoft, ELAN, FLEX

Daftar acuan

- 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)*.
- Larasati, Septina Dian, Vladislav Kuboň & Daniel Zeman. 2011. Indonesian morphology tool (MorphInd): Towards an Indonesian corpus. Dlm Cerstin Mahlow dan Michael Piotrowski (peny.) *Systems and Frameworks for Computational Morphology*, 119-129. Verlag: Springer.
- Nomoto, Hiroki, Shiro Akasegawa & Asako Shiohara. 2018a. Reclassification of the Leipzig Corpora Collection for Malay and Indonesian. *NUSA* 65: 47-66.
- Nomoto, Hiroki, Hannah Choi, David Moeljadi & Francis Bondb. 2018b. MALINDO Morph: Morphological dictionary and analyser for Malay/Indonesian. Kiyoaki Shirai (ed.) *Proceedings of the LREC 2018 Workshop "The 13th Workshop on Asian Language Resources"*, 36-43.
- Nomoto, Hiroki, Shiro Akasegawa & Asako Shiohara. 2018c. Building an open online concordancer for Malay/Indonesian. Presentasi di ISMIL 22.
- Nomoto, Hiroki & Hannah Choi. 2018. The Apparent lack of a complementizer-trace effect in Indonesian. Presentasi di ISMIL 22.

Penghargaan

Pengembangan MALINDO Conc dilakukan dengan dana JSPS “Program for Advancing Strategic International Networks to Accelerate the Circulation of Talented Researchers” yang ditawarkan ke Univeritas Kajian Asing Tokyo untuk proyek berjudul “A Collaborative Network for Usage-Based Research on Less-Studied Languages” dan dana JSPS #26770135 serta #18K00568. Pemakalah juga merakam penghargaan ke Universiti Teknologi Nanyang karena menerima pemakalah sebagai peneliti pelawat selama setahun.