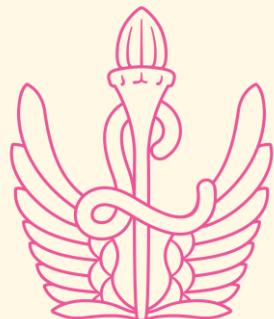


MALINDO Conc: Its development and use in linguistics, language teaching and NLP

Hiroki Nomoto (Tokyo University of Foreign Studies)

ISMIL 27 @ NUS, 5 June 2024



What is MALINDO Conc?

- Open online concordancer for Malay/Indonesian
- Developed and maintained by Hiroki Nomoto, Asako Shiohara (TUFS) and Shiro Akasegawa (LagoNLP).
- Open → Free of charge cf. Sketch Engine
- Online → Web-based cf. AntConc
- Concordancer → Generates concordance lines
 (= examples including a key word)
- For Malay/Indonesian → Tailored to Malay/Indonesian



Past talks on MALINDO Conc

- ISMIL 22 at UCLA (2018)
‘Building an open online concordancer for Malay/Indonesian’
- KOLITA 17 at UNIKA Atma Jaya (2019)
‘Pengembangan sumber bahasa digital dan konsep asas dalam linguistik Melayu/Indonesia’
- SEALTLs at NUS (2019)
‘Using MALINDO Conc for Malay/Indonesian language classes’

Studies using MALINDO Conc

- Shiohara, Sakon & Nomoto (2019)
Discourse functions of the two non-active voices in Indonesian:
Based on the web corpus data in MALINDO Conc.
NUSA 67: 77-101.
- Nomoto (2020)
Passive subtypes in Sarawak Malay.
NUSA 68: 139-158.
- Nomoto (2022)
Crossed control revisited: The structure and interpretations of
'want' and so on + passive verb in Malay/Indonesian.
Wacana 22(2): 338-364.

This talk

1. How to use MALINDO Conc
 - Basic functions
 - Corpora included
 - Updates since 2018/2019
2. How MALINDO Conc was developed
 - Corpus selection
 - Language identification
 - Morphological annotation
3. Usage examples in linguistics and language teaching
4. Future plans

1. How to use MALINDO Conc

<https://malindo.aa-ken.jp/conc/>

Overview and user guides (in Malay, Indonesian, Japanese)

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

<https://malindo.aa-ken.jp/conc/>

The screenshot shows the MALINDO Conc search interface. The top navigation bar includes the title "MALINDO Conc" and a language selection dropdown set to "Bahasa". The search interface features a "Pencarian" (Search) section with a search input field and a "Cari" (Search) button. Below this are several search filters:

- Kata Kunci:** A list of search types:
 - Bentuk Jadian tanpa spesifikasi
 - Akar Kata tanpa spesifikasi
 - Prefiks tanpa spesifikasi
 - Sufiks tanpa spesifikasi
 - Konfiks tanpa spesifikasi
 - Reduplikasi tanpa spesifikasi
- Kolokasi:** A search type: Kolokasi tanpa spesifikasi.
- Korpus:** A search type: Korpus tanpa spesifikasi.

On the right side, there are three buttons: "Lihat" (View), "KWIC" (highlighted in pink), and "Kalimat". At the bottom, there are navigation icons for previous/next pages, a page number "Halaman 0 dari 0", and a "Hasil pencarian tidak ditemukan" (Search result not found) message. There is also a "Unduh" (Download) button. The footer contains the copyright notice "MALINDO Conc Copyright © 2017-2018 Tokyo University of Foreign Studies, Nanyang Technological University and Lago Institute of Language. All rights reserved."

Simple search

- Minimum 3 steps
1. Specify **Keyword** • • • “What”
 2. Specify **Corpora** • • • “From where”
 3. Hit **Go**

Exercise 1

1. Search for sentences containing the word *sifar*.
2. List forms derived from the root *suka*.
3. Is it possible to attach the prefix *meN-* to *suka*?
4. Can *suka* occur with *meN-* but without any suffix?
5. What can you find with

Surface Form unspecified

?

Root unspecified

Prefix all of (meN-, per-)

Suffix equals "none"

Circumfix unspecified

Reduplication unspecified

Exercise 2

1. Find examples in which *yang* occurs within three words to the left of *suka*.
2. Find examples in which *suka* is immediately followed by a *di*-verb.
 - What does sentence 25 mean?

We can share the same results by using MALINDO Conc 😊

(Search results change constantly with search engines such as Google)

(1) Specify keyword

Keyword Clear All 

- Surface Form unspecified** ← For simple search
- Root unspecified
- Prefix unspecified
- Suffix unspecified
- Circumfix unspecified
- Reduplication unspecified

Keyword Clear All 

- Surface Form unspecified
- Root unspecified
- Prefix unspecified
- Suffix unspecified
- Circumfix unspecified
- Reduplication unspecified

↓ Type here

Surface Form equals  **malindo**

OK Cancel

(2) Specify corpora

Keyword

Surface Form equals "malindo"

Root unspecified

Prefix unspecified

Suffix unspecified

Circumfix unspecified

Reduplication unspecified

Collocate unspecified

Corpora unspecified

② Tick

Corpora specified

Leipzig Corpora Collection

IND

IND MXD2012

IND WEB2012

IND WKP2016

ZSM

ZSM MXD2012

ZSM WEB2012

ZSM WKP2016

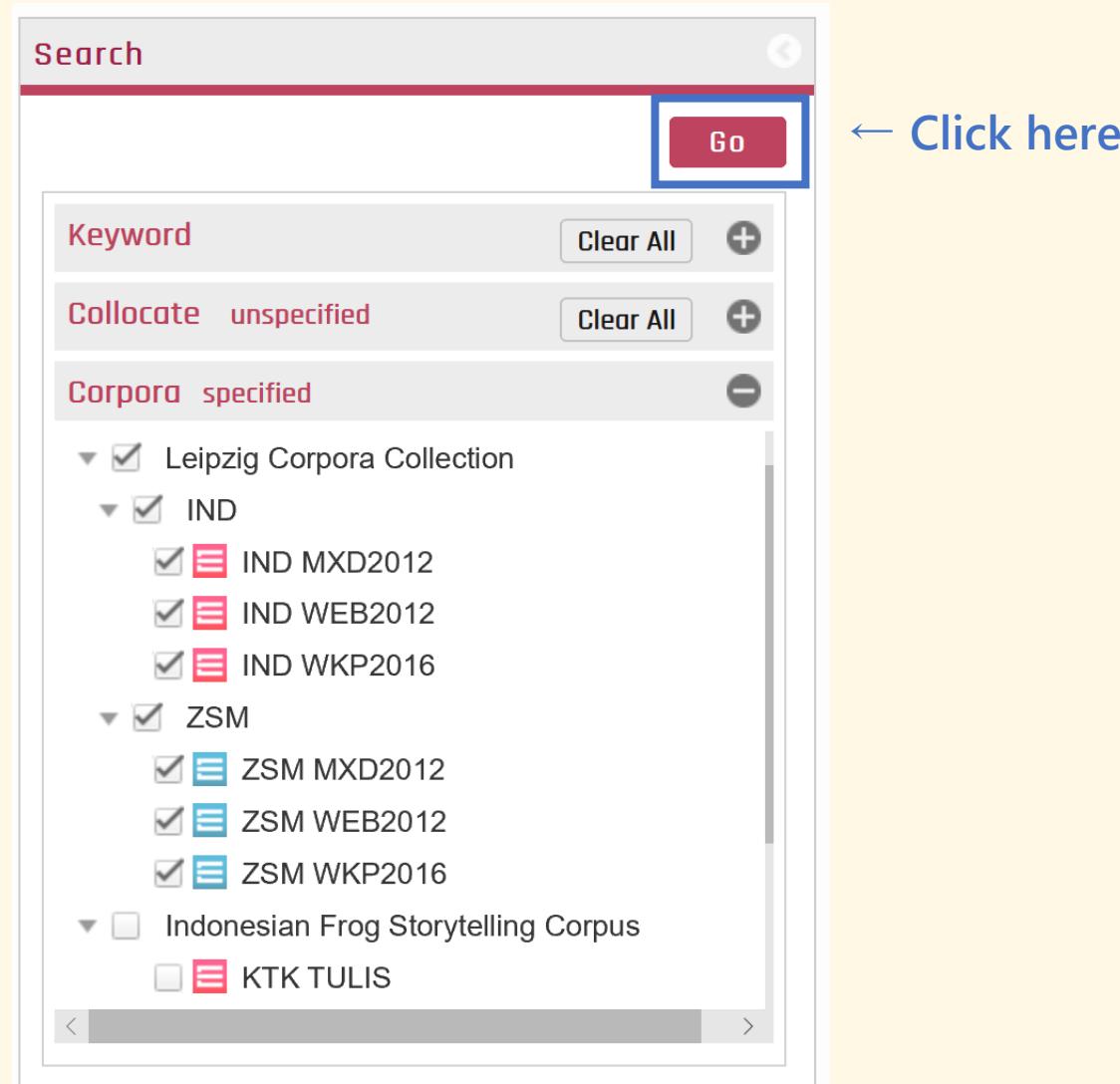
Indonesian Frog Storytelling Corpus

KTK TULIS

< >

① Click here

(3) Hit **Go**



Result

The screenshot shows a concordancer interface with the following elements:

- View**, **KWIC**, and **Sentence** buttons at the top right.
- A numbered list on the left side (1, 2, 3, 4, 5) corresponding to the search results.
- WKP2016** buttons aligned vertically on the right side.
- Four numbered callouts (1, 2, 3, 4) pointing to specific parts of the results.
- Result 1:** Sedangkan seluruh pesawat Boeing akan ditransfer kepada **Malindo** Air yang akan rebranding menjadi Batik Air Malaysia dan menyus...
1,000 lebih ahli yang terdiri oleh Kru Kabin dari MAS, Firefly, Air Asia, **Malindo** Airways dan Eagle Express.
- Result 2:** **Malindo** Air bercadang untuk memulakan operasi pada 1 Mei 2013 dari te...
- Result 3:** A. Ali sedang mencari penyanyi yang boleh menyanyikan lagu pop **Malindo** untuk filem usahasama Malaysia dan Indonesia berjudul " Gelora...
- Result 4:** Tetapi sehingga kini (2014) hanya **Malindo** Air sahaja yang beroperasi.

② ↑

③

④

- This is called a **concordance**
- **Concordancer** is a tool to generate it

Let's try!

What will happen if you clik on ① – ④?

Complex search (1): Morphological search

Keyword Clear All

Surface Form unspecified	Actual form
Root unspecified	Form without affixes and reduplication; dictionary top entry
Prefix unspecified	
Suffix unspecified	
Circumfix unspecified	
Reduplication unspecified	

Prefixes

Prefix

includes any of ✓

meN- N- di- ber- per- ter-

peN- pe- ke- se- none

Used to exclude
words with
prefixes

DK

Cancel

Suffixes

Suffix

includes any of ✓

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

none



Used to exclude words with suffixes

OK

Cancel

Circumfixes

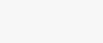
Circumfix

includes any of ▾

ber-...-an ber-...-kan ke-...-an

peN-...-an per-...-an pe-...-an

se-...-nya **none**



Used to exclude words with circumfixes

OK

Cancel

Reduplication

Reduplication

includes any of ▾

Full

Partial

Rhythmic

none

e.g. *buku-buku*

e.g. *seseorang*

e.g. *teka-teki*

↑ Used to exclude reduplicated words

OK

Cancel

Complex search (2): Collocation

The screenshot shows a complex search interface for collocation analysis. At the top, a grid of tokens is displayed in two rows:

L 5	Y	r	s	D	Key word	R 1	R 2	R 3	R 4	R 5
L 4	e	n	'	o	occurred	at	specific	in	the	key

Below the grid, several buttons allow specifying the type of collocate:

- Root unspecified
- Prefix unspecified
- Suffix unspecified
- Circumfix unspecified
- Reduplication unspecified

A callout bubble with the number ① points to the "Collocate unspecified" button.

A modal dialog box is open, containing the following text:

② Click here to specify the position of the other word

Find collocate between L5 and R5

Surface Form unspecified

Root unspecified

Find collocate between L5 and R5

OK Cancel

At the bottom, there are additional buttons for "Collocate unspecified" and "Reduplication unspecified".

Downloading the result

View **KWIC** Sentence

1	Berwatak jahat karena suka memakan korban manusia.	WEB2012
2	Biasanya sundel bolong juga diceritakan suka mengambil bayi-bayi yang baru saja dilahirkan.	WEB2012
3	hluk halus berwujud anak kecil yang kerdil, perawakannya gundul, dan suka mencuri.	WEB2012
4	Entah kenapa, Abi sangat suka menjilati seputar memek Tita, selain berbau harum juga sangat in…	WEB2012
5	tersebut belum bisa menyentuh kembali kesadarannya, penderita lebih suka menghindarinya.	WEB2012
6	Igiyakknya pun tidak mungkin, karena Loro Jonggrang memang tidak suka dengan Bandung Bondowoso.	WEB2012
7	Para pria suka bila Anda mengatakan bagian mana yang ingin disentuh.	WEB2012
8	" jawabku sedikit malu tapi mau dan suka sekali.	WEB2012
9	ar menghamburkan uang, tidak acuh pada masalah administrasi, tidak suka humor dan menghambat kemajuan berpikir.	WEB2012
10	al tentu memikat bila dipanggungkan oleh dramawan-dramawan yang suka mengangkat mitologi dalam bentuk futuris macam Robert Wilson.	WEB2012
11	Mereka mengaku beragama Islam, namun justru suka menyiksa manusia, menyiksa pembantu, berbohong, menghardik o…	WEB2012
12	e rest of this entry Saya sadar sepenuhnya bahwa dari awal saya lebih suka menyampahi blog ini dengan konsep-konsep pemrograman yang t…	WEB2012
13	Jika kau lihat orang yang tidak menyukai dinar dan dirham, tidak suka pujian dan sanjungan, aku bisa pastikan bahwa ia orang yang ikhla…	WEB2012
14	Yah kegedean dikit sih, tapi emang aku suka sepatu yang agak kebesaran supaya bisa terpakai sampai nanti.	WEB2012
15	Mereka lebih suka mencari informasi melalui Internet.	WEB2012
16	5. Setelah bertengkar, jangan biasakan diri suka menyendiri, kemudian menjauhkan diri dari pasangan hidup.	WEB2012
17	Dengan rendah hati, wanita keturunan ini justru lebih suka disebut Sales & Marketing Director, seperti yang tertulis di kartu na…	WEB2012

Page 1 of 183 | 1 - 50 of 9116

[Click here →](#) [Download](#)

(Processing will take some time)

Using the spreadsheet “Sort” function

nodus	knoteks kanan	
iat karena	suka	memakan korban manusia.
iceritakan	suka	mengambil bayi-bayi yang bar
indul, dan	suka	mencuri.
abi sangat	suka	menjilati seputar memek Tita,
erita lebih	suka	menghindarinya.
hang tidak	suka	dengan Bandung Bondowoso.
Para pria	suka	bila Anda mengatakan bagian
i mau dan	suka	sekali.
rasi, tidak	suka	A↓ 昇順(S) Z↓ 降順(O)
wan yang	suka	選択したセルの色を上に表示(C) 選択したフォントの色を上に表示(E)
un justru	suka	選択したセルのアイコンを上に表示(I) ユーザー設定の並べ替え(U)...

切り取り(I)
コピー(C)
貼り付けのオプション:
形式を選択して貼り付け(S)...
挿入(I)...
削除(D)...
数式と値のクリア(N)
クイック分析(Q)
フィルター(E)
並べ替え(O)
コメントの挿入(M)
セルの書式設定(F)...
ドロップダウンリストから選択(K)...
ふりがなの表示(S)
名前の定義(A)...
ハイパーリンク(I)...

Corpora included in MALINDO Conc

- Leipzig Corpora Collection, reclassified version (Goldhahn et al. 2012; Nomoto et al. 2018a) **ind**, **zsm**
- Indonesian Frog Storytelling Corpus (Moeljadi 2014) **ind**
- Corpus of Malay Varieties
 - Colloquial Standard Malay **zsm**
 - Sabah Malay **msa**
 - Sarawak Malay **msa**
 - Brunei Malay **msa**
 - Colloquial Jakarta Indonesian **ind**

Leipzig Corpora Collection (LCC)

- Data collected from web pages in 291 languages (as of May 2024)
- Open (available for anyone, free of charge)
- <https://corpora.uni-leipzig.de/en>



CORPORA COLLECTION
LEIPZIG

Search in 1019 Corpus-Based Monolingual Dictionaries for 291 Languages.

Selected language: German > News 2023

Enter a word



?

Search suggestions: Beginn · heimischen · Initiative · dagegen · 6

More information about: German > News 2023

Change corpus

The corpus **deu_news_2023** is a German news corpus based on material from 2023. It contains 33,142,449 sentences and 520,813,152 tokens. [Details](#)



DOWNLOADS

Download parts of this corpus.



STATISTICS

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

Further services:



WORDS OF
THE DAY

Current terms selected from daily newspapers and news services.

REST API

There are RESTful webservices for this language.

LCC data in MALINDO Conc (Indonesian)

Subcorpus	IND MXD2012	IND WEB2012	IND WKP2016
File name	Ind_mixed_2012	Ind_web_2012	Ind_wikipedia_2016
Content	Miscellaneous (e.g. blogs)	Mostly newspapers	Wikipedia
Size (sentences)	300,000	300,000	300,000
Size (tokens)	5,428,067	5,540,573	5,634,138

Known issue: Overlaps between these two subcorpora

LCC data in MALINDO Conc (Malay)

Subcorpus	ZSM MXD2012	ZSM WEB2012	ZSM WKP2016
File name	zsm_mixed_2012	zsm_web_2012	zsm_wikipedia_2016
Content	Miscellaneous (e.g. blogs)	Mostly newspapers	Wikipedia
Size (sentences)	300,000	300,000	300,000
Size (tokens)	4,184,270	4,180,551	5,787,015

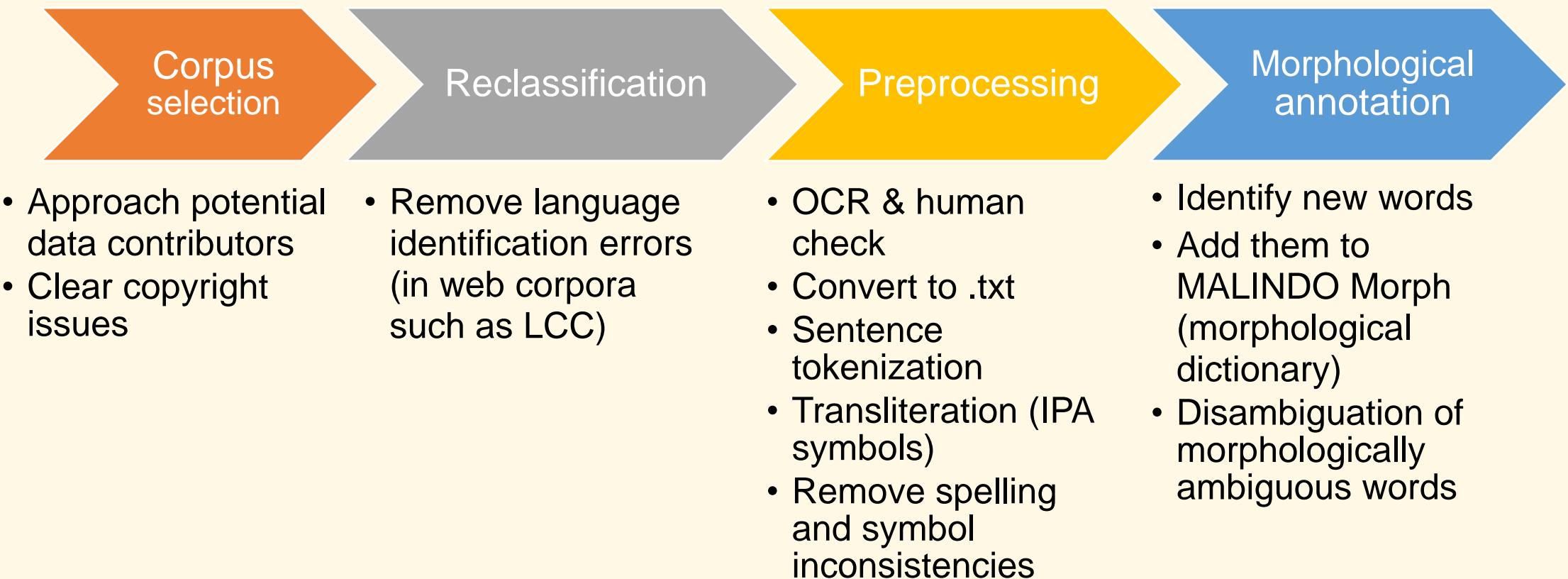
Known issue: Overlaps between these two subcorpora

Corpus of Malay Varieties

- Colloquial Standard Malay **zsm**
 - Standard Lisan (Nomoto 2018)
 - Appendices in Nor Hashimah (2003) and Nor Hashimah et al. (2005)
- Sabah Malay **msa**
 - Sabah (Kartini & Nomoto 2018)
- Sarawak Malay **msa**
 - Sarawak (Nomoto 2019)
 - Example sentences from Mohd. Ali (2015)
- Brunei Malay **msa**
 - Shiohara & Mohd Izzuddin Fitri (2021)
 - Haji Abdul Hamid & Palaniappan (1998)
- Colloquial Jakarta Indonesian **ind**
 - Jakarta Lisan (Shiohara & Yanti 2020)
 - Gil et al. (2015)

2. How MALINDO Conc was developed

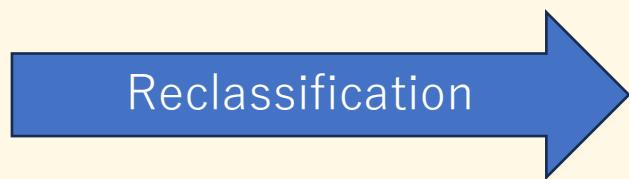
Corpus development processes



Reclassification of LCC (Nomoto et al. 2018a)

- Malay and Indonesian subcorpora in the Leipzig Corpora Collection (LCC) available in 2017 had numerous language identification (LangID) errors.
- We reclassified the corpora using a rule-based language identifier.
- However, some LangID errors still remain.

zsm zsm **ind** zsm zsm zsm
zsm zsm zsm **ind** zsm zsm
zsm zsm zsm zsm **ind** zsm
zsm zsm zsm zsm zsm **ind**
ind zsm zsm zsm zsm zsm zsm
zsm **ind** zsm zsm zsm zsm



zsm zsm zsm zsm zsm
ind zsm zsm zsm zsm zsm
zsm zsm zsm zsm zsm

Malay and Indonesian in LCC now

- Problematic subcorpora seem to have been removed.
- Side effect:
Most of the Malay subcorpora are either from newspapers or Wikipedia.
cf. Indonesian subcorpora are more varied in terms of register.

LangID accuracy (recall) now and then

- Far better than a decade ago, very close to native speakers
- But not good enough for linguists (perhaps good for NLP researchers/practitioners)

System/speaker	Malay	Indonesian
Lui (2014) langid.py	24.2%	77.1%
Google Translate (data used in Nomoto 2023)	92.0%	95.0%
Malay speaker (Farhan)	93.7%	89.3%
Indonesian speaker (David)	86.3 %	96.7%

Conversion to .txt JFS Jakarta Indonesian data (Gil et al. 2015)

Original (csv)	Session name	Speaker	Text
"425845090207170707","1506","BTJ-010607","476","0:17:33","","870","EXPERN","tapiq kan mesti qnya kaloq mauq majuq yaq... ini ajaq kan... berangkat ajaq, gitu.", "tapi kan məstjna kalɔ mau:? maju:? ya: ?ini ?aja kan bəraŋkat ?aja gitu", "tapi kan mesti -nya kalo mau maju ya ini aja kan ber- angkat aja gitu", "but KAN should -NYA TOP want move.forward yes this just KAN BER- lift just like.that", "but in fact, if he wants to be success... er... he should leave.", "", "referring to TIS' son."	BTJ-010607	BTJTIS	
"984293090207170707","1506","BTJ-010607","477","0:17:38","","481","BTJTIS","xxx, iyaq sebetulnya.", "xxx ?iya səbtulna", "xxx iya se- betul -nya", "xxx yes SE.NYA- right -CIRC", "xxx, yes, that's right.", "", ""	BTJ-010607	BTJTIS	
"729924090207170707","1506","BTJ-010607","478","0:17:41","","481","BTJTIS","tapiq karna diaq yaq mungkin xxx, karna ngga diijinkan a(gaq), agaq takut jugaq, gituq.", "təpi karna diyā? ya muŋkin yew karna ŋga diŋjinkan ?a ?aga? takut juga? gitu?", "tapi karna dia ya mungkin xxx karna ngga di- ijin -kan agaq agaq takut juga gitu", "but because 3 yes possible xxx because NEG DI- permission -KAN fairly fairly fear also like.that", "but maybe because he didn't get permission, it seemed that he was in distress.", "", "still talking about her son."	BTJ-010607	BTJTIS	

Conversion to .txt

JFS Jakarta Indonesian data (Gil et al. 2015)

- Extract Jakarta Indonesian data based on **session names**.
- Extract **speaker** and **text**.
- Format as a tab-separated text.

Formatted (txt)

EXPERN: tapiq kan mesti qnya kaloq mauq majuq yaq... ini ajaq kan... berangkat ajaq, gitu.

BTJTIS: xxx, iyaq sebetulnya.

BTJTIS: tapiq karna diaq yaq mungkin xxx, karna ngga diijinkan a(gaq), agaq takut jugaq, gituq.

Morphological annotation: Identify new words

- JFS Jakarta Indonesian data was tricky because of

➤ “q”

Representing glottal stops (but not always)

➤ “()”

Not uttered, but added by the transcribers

Formatted (txt)

EXPERN: tapiq kan mesti qnya kaloq mauq majuq yaq... ini ajaq kan... berangkat ajaq, gitu.

BTJTIS: xxx, iyaq sebetulnya.

BTJTIS: tapiq karna diaq yaq mungkin xxx, karna ngga diijinkan a(gaq), agaq takut jugaq, gituq.

Morphological annotation: Add new words to MALINDO Morph

- MALINDO Morph: Morphological dictionary for Malay/Indonesian (Nomoto et al. 2018b)

ec-62080	mau	mauq	0	0	0	0	Indo-Jakarta-Lisan	mau	mau
ec-62081	gandaria	gandariaq	0	0	0	0	Indo-Jakarta-Lisan	gandaria	gandaria
ec-62082	variasi	variasiqnya	0	-nya	0	0	Indo-Jakarta-Lisan	variasi+dia	variasi+dia
ec-62083	kena	kenaq	0	0	0	0	Indo-Jakarta-Lisan	kena	kena
ec-62084	merek	mérek	0	0	0	0	Indo-Jakarta-Lisan	merek	merek
ec-62085	orang	seseòrang	se-	0	0	R-separa	Indo-Jakarta-Lisan	seseorang	seseorang

ID	root	surface form	pre fix	suf fix	cir cum fix	redupli cation type	source	stem	lemma

Can morphological analysis be automated?

- MorphInd (Larasati et al. 2011)
 - Widely used (e.g. Universal Dependencies)
 - Cannot identify circumfixes
pengiriman (= *kirim* + *peN-...-an*) ‘delivery’
→ ^peN+kirim<v>+an_NSD\$
- SANTI-morf (Prihantono 2021)
 - Can handle circumfixes
 - Analyses differ from MALINDO Morph’s

Form	SANTI-morf	MALINDO Morph
<i>perbaiki</i> ‘to fix’	<i>baik</i> + <i>per-...-i</i>	<i>per-</i> + <i>baik</i> + <i>-i</i>
<i>pertahankan</i> ‘to maintain’	<i>tahan</i> + <i>per-...-kan</i>	<i>per-</i> + <i>tahan</i> + <i>-kan</i>
<i>seharian</i> ‘all day long’	<i>hari</i> + <i>se-...-an</i>	<i>se-</i> + <i>hari</i> + <i>-an</i>

ChatGPT (GPT-4o)

User:

Please conduct a morphological analysis for the following Malay sentence.

Saya perlu memperbaharui pasport saya secepat mungkin.

ChatGPT:

[...]

3. Memperbaharui

- **Prefix:** mem-
- **Infix:** -per-
- **Root:** baharu
- **Suffix:** -i

[...]

User: Can you make a table?

Word	Root	Prefix	Infix	Suffix	Part of Speech	Meaning
Saya	saya	-	-	-	Pronoun	I
Perlu	perlu	-	-	-	Verb	need
Memperbaharui	baharu	mem-	-per-	-i	Verb	to renew
Pasport	pasport	-	-	-	Noun	passport
Saya	saya	-	-	-	Possessive Pronoun	my
Secepat	cepat	se-	-	-	Adverb	as quickly
Mungkin	mungkin	-	-	-	Adverb	possible

User: Analyze the word "ketidakberperikemanusiaan".

Component	Type	Meaning
ke-	Prefix	state or condition
tidak-	Prefix	not, lack of
ber-	Prefix	having or possessing
peri-	Prefix	about or concerning
ke-	Prefix	state or condition (within "kemanusiaan")
manusia	Root	human
-an	Suffix	noun (state, condition, or quality)

No circumfixes
(cf. MorphInd)

User: Are there any circumfixes in the word?

Yes, the word "ketidakberperikemanusiaan" contains circumfixes. A circumfix is a combination of a prefix and a suffix that together modify a root word. [...]

Component	Type	Meaning
Ke-...-an (outer circumfix)	Circumfix	state or condition (entire word)
ke-...-an (inner circumfix)	Circumfix	state or condition (within "kemanusiaan")
Tidak	Prefix	not, lack of
Ber-	Prefix	having or possessing
Peri-	Prefix	about or concerning
Manusia	Root	human
-an	Suffix	noun (state, condition, or quality)



User: No, circumfixes are not a simple combination of a prefix and a suffix.

You're right. A circumfix is **a single morphological unit** that consists of **two parts, a prefix and a suffix**, which together surround a root word to create a new meaning. [...]

Component	Type	Meaning
Ke-....-an	Circumfix	State or condition
Tidak	Prefix	Not, lack of
Ber-	Prefix	Having or possessing
Peri-	Prefix	About or concerning
Ke-....-an (within kemanusiaan)	Circumfix	State of being human (humanity)
Manusia	Root	Human

Morphological analysis is not an interesting topic in NLP any more?

- Generative AIs do not necessarily require linguistically correct analyses. <https://platform.openai.com/tokenizer>

Many words map to one token, but some don't: `indivisible`.

Unicode characters like emojis may be split into many tokens containing the underlying bytes: `???????????`

Sequences of characters commonly found next to each other may be grouped together: `1234567890`

- Given the **usage- and frequency-based** nature of generative AIs, has our conception of “linguistically correct” been completely wrong? cf. language as a **symbolic system**

Disambiguation of morphologically ambiguous words

1. Xiong adalah **beruang**, mao berarti kucing. ‘bear (animal)’
Root *beruang*
2. Tak seperti lazimnya reptil, buaya memiliki jantung **beruang** empat, sekat rongga badan (diafragma) dan cerebral cortex. ‘to have room’
Root *ruang* + prefix *ber-*
3. Mau ke pesta tidak **beruang**, akhirnya nongkrong di pinggir jalan. ‘to have money’
Root *uang* + prefix *ber-*

Annotation file (xml)

```
<doc url="https://github.com/matbahasa/Jakarta_Lisan/tree/main/JFS/JFS-Jakarta1506.txt" subcorpus="var_jkt-l-JFS">  
<s id="var_jkt-l-JFS.0000468" subcorpus="var_jkt-l-JFS" url="https://github.com/matbahasa/Jakarta_Lisan/tree/main/JFS/JFS-Jakarta1506.txt">  
    <w rt="tapi" stem="tapi" lemma="tapi">tapiq</w>  
    <w rt="kan" stem="kan" lemma="kan">kan</w>  
    <w rt="mesti" s1="-nya" stem="mestinya@mesti+dia" lemma="mestinya@mesti+dia">mestiqnya</w>  
    <w rt="kalo" stem="kalo" lemma="kalo">kaloq</w>  
    <w rt="mau" stem="mau" lemma="mau">mauq</w>  
    <w rt="maju" stem="maju" lemma="maju">majuq</w>  
    <w rt="ya" stem="ya" lemma="ya">yaq</w>  
    <g/>  
    <pu>...</pu>  
    <w rt="ini" stem="ini" lemma="ini">ini</w>  
    <w rt="aja" stem="aja" lemma="aja">ajaq</w>  
    <w rt="kan." stem="kan." lemma="kan.">kan</w>  
    <g/>  
    <pu>.</pu>  
    <pu>.</pu>  
    <w rt="angkat" p1="ber-" stem="berangkat" lemma="berangkat">berangkat</w>  
    <w rt="aja" stem="aja" lemma="aja">ajaq</w>  
    <g/>  
    <pu>.</pu>  
    <w rt="gitu" stem="gitu" lemma="gitu">gitu</w>  
    <g/>  
    <pu>.</pu>  
    </s>  
</doc>
```

3. Usage examples in linguistics and language teaching

Use in Nomoto (2022)

Nomoto, Hiroki. 2022. [Crossed control revisited: The structure and interpretations of 'want' and so on + passive verb in Malay/Indonesian](#). *Wacana* 22(2): 338-364.

Kucing=nya ma(h)u/suka di-cium oleh Siti.²
cat=3 want/like PASS-kiss by Siti

(i) 'Her cat wants/likes to be kissed by Siti.'

(non-crossed reading; attitude holder = her cat)

(ii) 'Siti wants/likes to kiss her cat.'

(crossed reading; attitude holder = Siti)

- Critical assessment of (i) Jeoung (2020), (ii) Berger (2019), Kroeger and Frazier (2020)

(i) Jeoung's (2020) doubt on the empirical robustness of crossed reading

Jeoung, Helen. 2020. [Categorial ambiguity in *mau*, *suka*, and other Indonesian predicates](#). *Language* 96 :e157–e172.

- The crossed reading is not as common as previously thought.
- What has been analysed as such in fact involves a different meaning, more specifically an aspectual meaning of the same predicate:
 - ‘about to, will’ for *mau*
 - ‘often’ for *suka*
- My conclusion: Not supported by corpus data!

Methodology

Corpora

- LCC [14,470,873 tokens]
 - IND MXD2012
 - IND WEB2012
 - IND WKP2016
- Indonesian Frog Storytelling Corpus (Moeljadi 2014) [22,446 tokens]

Query (for *mau/suka + di-* passive)

Keyword

Surface Form = **mau|suka**

Collocate

- Find Collocate between **R1** and **R1**
- Prefix = **di-**

Results: *mau* ‘*want*; *about to*, *will*’

Mau + di- passive

Interpretation	Frequency
non-crossed	34
crossed	6
temporal	2
crossed or temporal	7
unclear	1
total	50

Mau + bare passive

Interpretation	Frequency
non-crossed	0
crossed	4
crossed > temporal	22
temporal	0
temporal > crossed	5
crossed or temporal	3
total	34

Results: *suka* ‘like; often’

Suka + di- passive

Interpretation	Frequency
non-crossed	41
crossed	0
temporal	3
crossed or temporal	3
total	47

Suka + bare passive

Interpretation	Frequency
crossed	2
crossed > temporal	1
total	3

(ii) Restructuring analysis of crossed reading

Berger (2019), Kroeger and Frazier (2020)

- Matrix and embedded clauses share a single Voice feature (Voice agreement).
- That Voice feature is realized in the embedded clause.



Kucing=nya|ma(h)u/suka|di-cium oleh Siti.²
cat=3 want/like PASS-kiss by Siti

(i) 'Her cat wants/likes to be kissed by Siti.'

∅

di-

(non-crossed reading; attitude holder = her cat)

(ii) 'Siti wants/likes to kiss her cat.'

(crossed reading; attitude holder = Siti)

Primary evidence

[Voice: pass] can be realized in

- a. both matrix and embedded clauses
- b. matrix clause, but not embedded clause

[Voice: ____] [Voice: pass]

- a. Buah itu kemudian **di-coba** di-makan, ternyata rasa=nya asam.
fruit that subsequently PASS-try PASS-eat perceived flavor=3 sour
'The fruit was then tried to be eaten, and its taste was found to be sour.'²⁵
- b. Tapi kita tak tahu apa yang **di-suka** makan...
but 1PL not know what REL PASS-like eat
'But we don't know what they like to eat...' ²⁶

(Kroeger and Frazier 2020)

(a) Double realization in MALINDO Conc

Perampok di-coba di-tangkap polisi.
thief PASS-try PASS-catch police

- (i) *'The thief tried to be caught by the police.' (non-crossed reading)
(ii) 'The police tried to catch the thief.' (crossed reading)

- Common, so I did not check MALINDO Conc in the paper.
cf. *dicoba* + *di-* V (17), *dimau* + *di-* V (0), *disuka* + *di-* (0)
 - My conclusion: Not a crossed control construction at all, but a standard biclausal structure with a CP complement, paralleling (cf. Saddy 1991)

Perampok itu di-katakan [CP di-tangkap polisi].
thief that PASS-say PASS-catch police
'The thief is said to have been caught by the police.'

cf. Patrianto and Chen (2023):
This is the sole analysis for
the Javanese crossed control
equivalent.

(b) Matrix realization of [Voice: pass]

- Only one instance in MALINDO Conc (900K sentences)
- Elicitation: 4/8 judged as “absolutely unnatural”

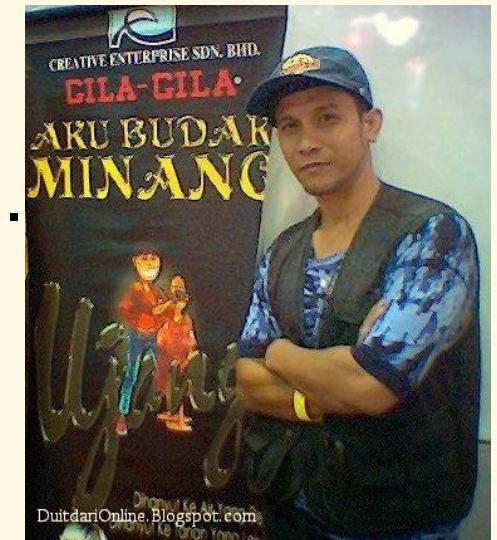
- a. ?*Ayah di-coba/di-mau/di-suka obati oleh ibu. [4/8]
father PASS-try/PASS-want/PASS-like treat(medically) by mother
'Mother tries/wants/likes to treat Father.'
- b. *?Ayah tentu ku=coba/ku=mau/ku=suka obati. [4/8]
father certainly 1SG=try/1SG=want/1SG=like treat(medically)
'I certainly try/want/like to treat Father.'

- My conclusion: compound verbs (for some speakers)

Use in language teaching

Autonomous vocabulary learning

1. Choose five words from *Aku Budak Minang* that are not found on the reading assistance system [Malay Reading Tutor](#) (Nomoto & Kawamura 2019).
2. For each word, describe its meaning, quote the sentence in *Aku Budak Minang* where it is used and translate that sentence.
3. Find three other sentences containing the word using MALINDO Conc and translate them.



Words chosen by the students

- **Simply unavailable words:**
baligh, belatuk, bergenang, berkhatan, berlaga, berlenggang, bertahun-tahun, bertimpa-timpa, hulur , jelingan, kampit, kang, karun, menjinjit, menghulur, mengorat, meraung, merempat, parut, raga, reban, tafsir, tempelak, tenung, tergagap-gagap,
tergamak, terkedu, terpa, tersadai, terserempak, tersisih, tertanya-tanya
[= siapa], sorang-sorang [= seorang-seorang], taruk [= taruh], udah [= sudah]
- **Colloquial forms:**
atuk [= datuk], bagitau [= beritahu], berpekdah [= berfaedah], blah, bukak [= buka], caya [= percaya], heran [= hairan], jugak [= juga], kejap [= sekejap], mintak [= minta], pedulik [= peduli], sapa
[= benar], camno [= macam mana], meh [= mari]
- **Dialectal forms:**
beno [= benar], camno [= macam mana], meh [= mari]
- **Non-standard spellings:**
bersopan-santun [= bersopan santun], saguhati [= sagu hati]

4. Future plans

Data update

- More recent data is available on the LCC download page.
- Currently available corpora do not contain words such as *Covid*.

Now	Future
MXD2012	MXD2012
WEB2012	Newscrawl 2016
WKP2016	Wikipedia 2021

Move to a new server

- MALINDO Conc will be down for about 10 days sometime between July 2024 – March 2025.

Terima kasih!

References

- Berger, Mike. 2019. [Indonesian crossed control: Expanding the typology of restructuring](#). In Richard Stockwell, Maura O'Leary, Zhongshi Xu and Z.L. Zhou (eds.) *Proceedings of the 36th West Coast Conference on Formal Linguistics*, 61–70. Somerville, MA: Cascadilla Proceedings Project.
- Gil, David, Uri Tadmor, John Bowden & Bradley Taylor. 2015. [Data from the Jakarta Field Station, Department of Linguistics, Max Planck Institute for Evolutionary Anthropology, 1999-2015](#).
- 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)*.
- Haji Abdul Hamid bin Jaludin & Periannan Palaniappan. 1998. [Folklor Kampong Aying: Himpunan Hasil Projek Folklor Kampong Ayer](#). Gadong: Jabatan Kesusastraan Melayu, Fakulti Sastera dan Sains Kemasyarakatan, Universiti Brunei Darussalam.
- Jeoung, Helen. 2020. [Categorial ambiguity in *mau*, *suka*, and other Indonesian predicates](#). *Language* 96 :e157–e172.
- Kartini Abd. Wahab & Hiroki Nomoto. 2018. [Korpus Variasi Bahasa Melayu: Sabah](#).
- Kroeger, Paul and Kristen Frazier. 2020. [Crossed-control in Malay/Indonesian as long-distance passivization](#). In Illeana Paul (ed.) *Proceedings of the twenty-sixth meeting of the Austronesian Formal Linguistics Association (AFLA)*, 159–174. University of Western Ontario.

- Larasati, Septina Dian, Vladislav Kuboň and Daniel Zeman. 2011. Indonesian morphology tool (MorphInd): Towards an Indonesian corpus. In Cerstin Mahlow et al. (eds.) *Systems and Frameworks for Computational Morphology*, 119–129. Springer, Verlag.
- Lui, Marco Hoiyin. 2014. [Generalized Language Identification](#). University of Melbourne dissertation.
- Moeljadi, David. 2014. Usage of Indonesian possessive verbal predicates: A statistical analysis based on storytelling survey. *Tokyo University Linguistic Papers* 35: 155–176.
- Mohd. Ali Salim. 2015. *Dialek Melayu Sarawak: Perkataan, Ayat dan Perbualan*. Kuala Lumpur: Dewan Bahasa dan Pustaka.
- Nomoto, Hiroki. 2018. [Korpus Variasi Bahasa Melayu: Standard Lisan](#).
- Nomoto, Hiroki. 2019. [Korpus Variasi Bahasa Melayu: Sarawak](#).
- Nomoto, Hiroki. 2020. [Passive subtypes in Sarawak Malay](#). *NUSA* 68: 139-158.
- Nomoto, Hiroki. 2022. [Crossed control revisited: The structure and interpretations of 'want' and so on + passive verb in Malay/Indonesian](#). *Wacana* 22(2): 338-364.

- Nomoto, Hiroki. 2023. [Issues surrounding the use of ChatGPT in similar languages: The case of Malay and Indonesian](#). In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, 76-82. Association for Computational Linguistics.
- Nomoto, Hiroki, Shiro Akasegawa & Asako Shiohara. 2018a. [Reclassification of the Leipzig Corpora Collection for Malay and Indonesian](#). *NUSA: Linguistic Studies of Languages in and around Indonesia* 65: 47–66.
- Nomoto, Hiroki, Hannah Choi, David Moeljadi and Francis Bond. 2018b. [MALINDO Morph: Morphological dictionary and analyser for Malay/Indonesian](#). In Kiyoaki Shirai (ed.) *Proceedings of the LREC 2018 Workshop "The 13th Workshop on Asian Language Resources"*, 36-43.
- Nomoto, Hiroki & Yoshiko Kawamura. 2019. Mareego gakushuusha o taishou nishita dokkai shien shisutemu no kaihatsu [On the development of a reading assistance system for learners of Malay]. *Gaikokugo Kyouiku Kenkyuu* 22: 215-229.
- Nor Hashimah Jalaluddin. 2003. *Bahasa dalam Perniagaan: Satu Analisis Semantik dan Pragmatik*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

- Nor Hashimah Jalaluddin, Harishon Radzi, Maslida Yusof, Raja Masittah Raja Ariffin & Sa'adiah Ma'alip. 2005. *Sistem Panggilan dalam Keluarga Melayu: Satu Dokumentasi*. Kuala Lumpur: Dewan Bahasa dan Pustaka.
- Patrianto, Hero and Victoria Chen. 2023. [Crossed control as an illusion: Insights from Javanese](#). Paper presented AFLA 30.
- Prihantono. 2021. *An Automatic Morphological Analysis System for Indonesian*. Lancaster University dissertation.
- Saddy, Douglas. 1991. WH scope mechanism in Bahasa Indonesia. In Lisa L.S. Cheng and Hamida Demirdash (eds.) *MIT working papers in linguistics 15: More papers on wh-movement*, 183–218.
- Shiohara, Asako & Yanti. 2020. Korpus Variasi Bahasa Melayu: Bahasa Indonesia Jakarta Lisan.
- Shiohara, Asako, Yuta Sakon and Hiroki Nomoto. 2019. [Discourse functions of the two non-active voices in Indonesian: Based on the web corpus data in MALINDO Conc](#). *NUSA* 67: 77-101.

Acknowledgements

The development of MALINDO Conc was supported by the following grants:

- JSPS Program for Advancing Strategic International Networks to Accelerate the Circulation of Talented Researchers “A Collaborative Network for Usage-Based Research on Lesser-Studied Languages”
- JSPS Program for Fostering Globally Talented Researchers “Developing Human Resources Taking the Lead in Research on Endangered and/or Minority Languages in an International Network”
- JSPS Grants #26770135, #JP18K00568, #20K00599 and #23H00639.
- Tokyo University of Foreign Studies IRC (Information Resources Center Project (AY2019-2020)