An annotated news corpus of Malaysian Malay

Siaw-Fong CHUNG and Meng-Hsien SHIH

National Chengchi University, Taiwan and National Taiwan University
sfchung@nccu.edu.tw, simon.xian@gmail.com

Malaysian Malay corpora and Malay annotation tools have been developed separately, and as such a complete tagged corpus is not yet available. The aim of this study was to create an annotated corpus of Malaysian Malay newspaper articles that supports accurate searches of parts-of-speech, affixes, and lemmas, producing a tagged corpus with morphological information from accessible resources such as morphology analyzers and parts-of-speech taggers. With this annotated corpus, precise searches for lemmas and affixes will be allowed. In this paper, the results of the annotation tool and its limitations will be presented, as well as a demonstration of a precise search for the prefix *ber-* and how its linguistic behaviors were analyzed. This study will make an important contribution to Malay linguistics, as a corpus with complete tagging and morphology information is not yet openly available.

1. Introduction[1]

A corpus provides linguistic examples that are beyond imagination. An annotated corpus provides even more fine-grained information needed for specific linguistic analysis. There are several corpora for Malaysian Malay (hereafter "Malay") and the most cited one is the online DBP Corpus (Dewan Bahasa and Pustaka).[2] This corpus has served as core material for the development of many other sub-corpora in Malaysia. Some studies (e.g., the Malay Practical Grammar Corpus [MPGC], Imran Ho Abdullah et al., 2004; MALay LEXicon [MALEX], Zuraidah Mohd Don, 2010) have taken a small portion of this corpus to create other grammar references and lexical databases. In addition to the use of the DBP Corpus, many scholars have also built and used their own corpora. Among these are Chung's (2010, 2011, 2013, 2014, 2019) self-collected Malay newspaper articles, Lee and Low's (2011) Malay textbook corpus, Mohd Hanafi Ahmad Hijazi et al.'s (2016) Facebook and Tweet posts, and many other translation databases.[3]

Most of the existing Malay corpora, including the DBP Corpus and the self-created corpora mentioned above, either have limited access or are not lemmatized or part-of-speech (POS) tagged.[4] For the DBP Corpus, for example, an ordinary user can obtain sufficient sample sentences from the corpus but without tag information, while a trained user who has a self-collected Malay corpus that can be processed using any concordancer might use running texts with keyword in-context (KWIC) without tag information. Such shortcomings have prevented the research of Malay corpus linguistics to advance to another semi-automatic level. A tagged Malay corpus would enable more accurate searches of POS, affixes, and lemmas, none of which are easily obtained today. Scholars who have worked on POS taggers in Malaysia usually worked separately from corpus

[1] The authors would like to thank the Taiwan Ministry of Science and Technology Project 106-2410-H-004-109-MY2 and 108-2410-H-004-095- for supporting the research herein.

[2] http://sbmb.dbp.gov.my/korpusdbp/SelectUserCat.aspx

[3] Malay names are cited as full names.

[4] This evaluation was made based on a public user's perspective, regardless of the existence of any complete, in-house tagged corpora that are not open to the public.

CHUNG, Siaw-Fong and Meng-Hsien SHIH, 2019. 'An annotated news corpus of Malaysian Malay'. In Hiroki Nomoto and David Moeljadi, eds. *Linguistic studies using large annotated corpora*. NUSA 67: 7–34. [Permanent URL: http://repository.tufs.ac.jp/handle/10108/94451] [doi: 10.15026/94451]

builders. Some resources may provide a lemmatizer but not a POS tagger, and vice versa. This makes the research of Malay corpus linguistics difficult to some extent. This situation must change if further development is expected for Malay corpus linguistics.

In this study, we evaluated several available resources before selecting a tagger and a morphology analyzer to perform the lemmatizing and POS tagging of a self-collected corpus. The selection was made after testing different tools. Based on the annotation results, we analyzed the distributions of different lemmas, affixes, and morphological combinations in Malay. We then ran a search for *ber-* to see how deeper semantic annotation could be carried out, which is needed for further Malay corpus linguistics research. The following are our research questions:

(a) What are the advantages and disadvantages of using existing Malay corpus annotation tools?

(b) What were the distributions of lemmas, affixes, and morphological combinations found in the annotated Malay news corpus?

(c) If further semantic annotation is needed in a tagged Malay corpus, how can it be carried out?

All three research questions were answered by first (a) running an experiment to test the annotation of a raw Malay news corpus; (b) analyzing the tagged results by calculating the proportions of different lemmas, affixes, and morphological combinations in the corpus; and (c) retrieving all the instances of *ber-* and categorizing them based on their different meanings.

Hereafter, Malaysian Malay will be termed "Malay." Unless otherwise stated, a "tagged" corpus will mean a corpus with both POS and morphological tag information. The "lemmatizer" will be called a "morphology analyzer" in a later part of this paper.

The next section will present the evaluation of the current existing corpus tools for Malay; it will also introduce some tools mainly used for Indonesian data.

2. Evaluation of existing corpora and tools

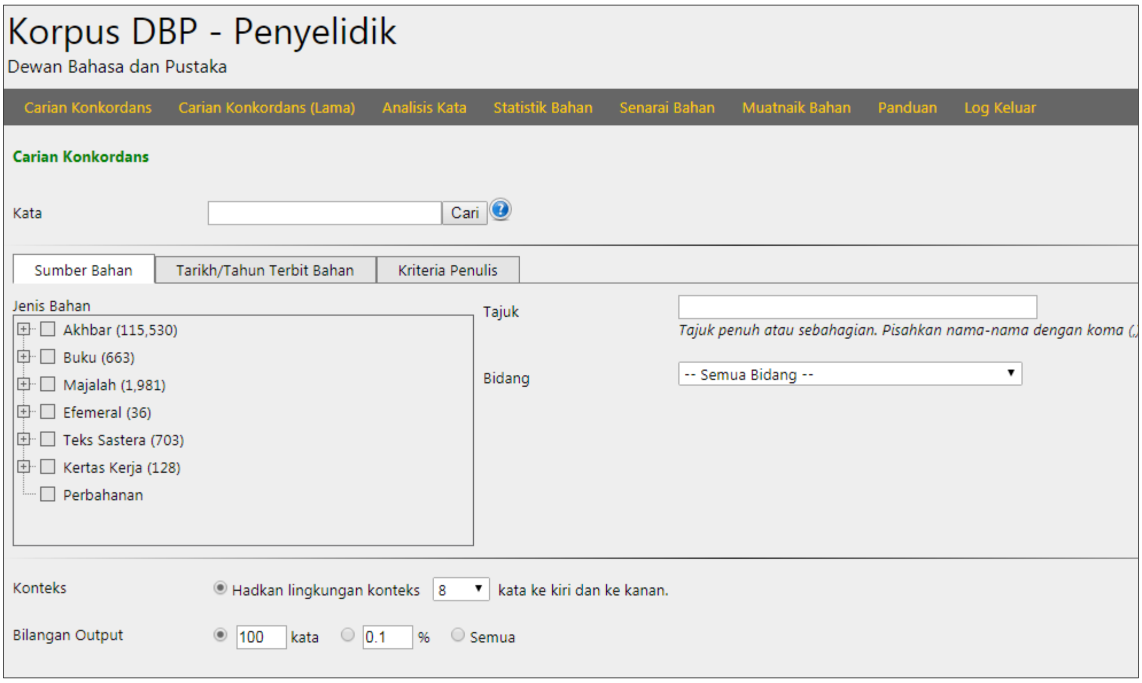
The online DBP Corpus contains *DBP*-published works such as books, literary texts, and other materials, including newspapers, magazines, working papers, etc. According to the statistics of the corpus materials provided by the DBP Corpus website, the corpus consists of 115,530 news articles, 1,981 magazine articles, 703 literary texts, 663 books, 128 working papers, and 36 "ephemeral" materials.[5] However, the number of words in the books and articles is unknown. In total, there are 118,913 texts (one book was considered one long text). Among the newspaper texts, a majority of them (86,885; 75%) came from *Berita Harian*, and 17,539 (15%) came from *Utusan Malaysia*, the same resource used for our own corpus. The remaining texts were from smaller newspapers such as *Harian Metro*, *Berita Minggu*, *Harakah*, and *Metro Ahad* (accessed February 23, 2019). However, the corpus details (i.e., the dates, the size, and the number of words) were not found on the DBP Corpus website. The following is an evaluation of using their platform.

The DBP Corpus interface has made several improvements over the years. The new interface now allows users to log in as a public user or as a researcher.[6] After logging in

[5] No further explanation could be obtained for what was meant by "ephemeral" materials.

[6] Several trials were carried out at different times. In February 2019, even when the setting was set to *semua* 'all', only 100 instances were shown (accessed February 23). Another search in the corpus in August 2019

as a researcher, a search for *ber** was carried out (an asterisk was used as a wildcard as there is no other better way to search for a prefix in a non-annotated corpus). This command returned noise, such as *beri* 'give' and *berita* 'news', which are words that start with *ber* but not as a prefix. The interface is shown in Figure 1 below. The search returned as many as 191,746 instances, with each page containing a maximum of 10 instances as a default.[7] When we clicked on the KWIC, we were provided with a single sentence containing the keyword. Two options were given—either to obtain the whole sentence (*Keseluruhan ayat*) or the whole paragraph (*Keseluruhan perenggan*)—but often the two were of little difference as a paragraph in Malay newspaper articles often consists of one whole long sentence (see the bottom of Figure 1). With several trials, we found only a few hits that returned more than one sentence. When we tried to download the whole results, it took a long time and we did not proceed to the next step. Such an interface may be suitable for learners and teachers looking for examples, but since researchers cannot download the data for further analysis, searches for specific aspects, such as prefixes, are difficult.



Jenis Bahan
- ☐ Akhbar (115,530)
- ☐ Buku (663)
- ☐ Majalah (1,981)
- ☐ Efemeral (36)
- ☐ Teks Sastera (703)
- ☐ Kertas Kerja (128)
- ☐ Perbahanan

Tajuk

Tajuk penuh atau sebahagian. Pisahkan nama-nama dengan koma (,

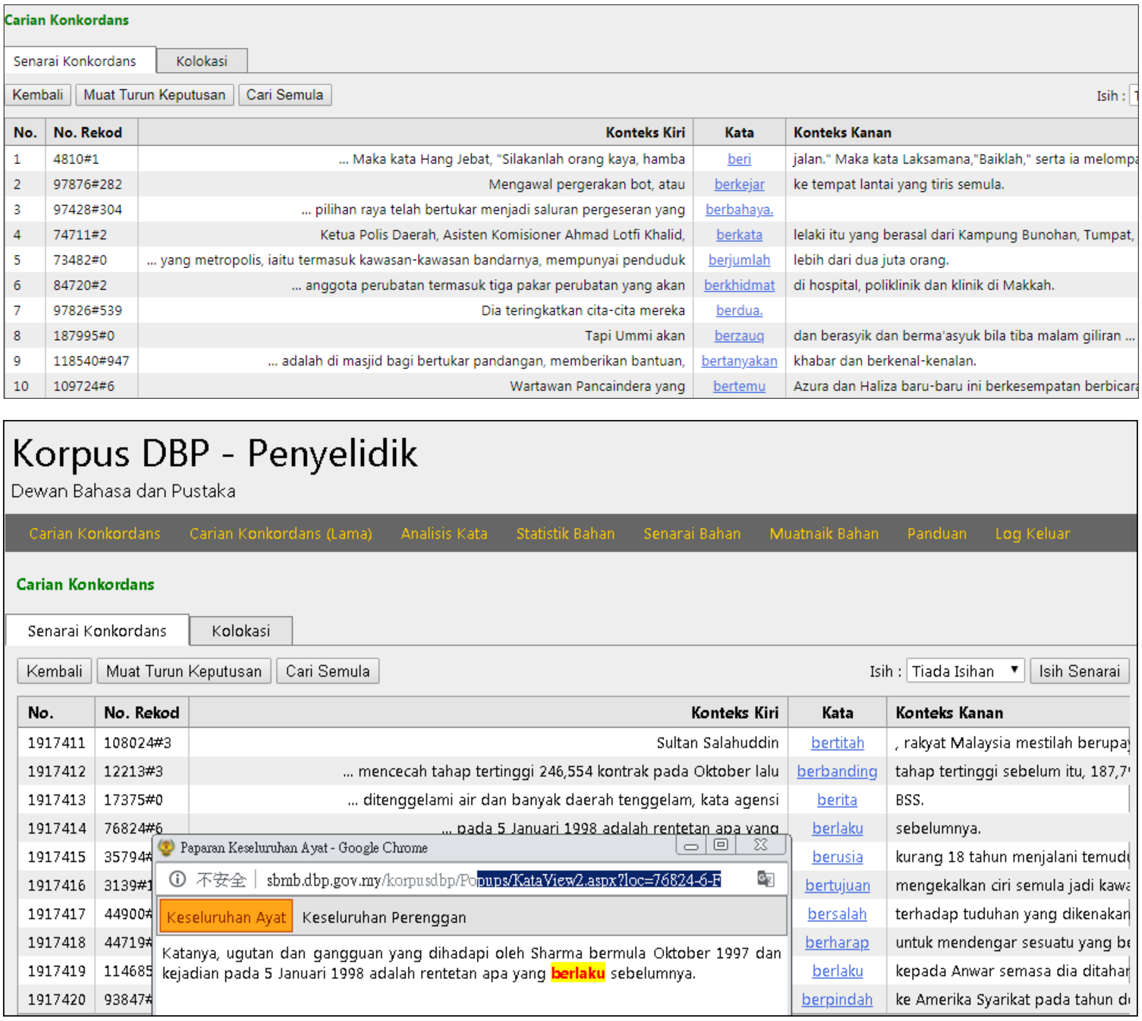
Bidang -- Semua Bidang --

Konteks ● Hadkan lingkungan konteks 8 kata ke kiri dan ke kanan.

Bilangan Output ● 100 kata ○ 0.1 % ○ Semua

returned only a maximum of nine instances (accessed August 23), no matter how many instances we set as our search criterion. On September 23, 2019, the full results were finally obtained when the setting was *semua*. These different trials show that the website was not stable across time, and if this situation continues, this will make the linguistic analysis of Malay difficult for researchers.

[7] We tried to change this to 500 hits per page but no changes were found after a long wait. We then let the program run for a while, and the requested 500 hits per page were finally loaded, but this was too long of a wait for only one search (accessed September 23).



No.	No. Rekod	Konteks Kiri	Kata	Konteks Kanan
1	4810#1	... Maka kata Hang Jebat, "Silakanlah orang kaya, hamba	beri	jalan." Maka kata Laksamana,"Baiklah," serta ia melompa
2	97876#282	Mengawal pergerakan bot, atau	berkejar	ke tempat lantai yang tiris semula.
3	97428#304	... pilihan raya telah bertukar menjadi saluran pergeseran yang	berbahaya.	
4	74711#2	Ketua Polis Daerah, Asisten Komisioner Ahmad Lotfi Khalid,	berkata	lelaki itu yang berasal dari Kampung Bunohan, Tumpat,
5	73482#0	... yang metropolis, iaitu termasuk kawasan-kawasan bandarnya, mempunyai penduduk	berjumlah	lebih dari dua juta orang.
6	84720#2	... anggota perubatan termasuk tiga pakar perubatan yang akan	berkhidmat	di hospital, poliklinik dan klinik di Makkah.
7	97826#539	Dia teringkatkan cita-cita mereka	berdua.	
8	187995#0	Tapi Ummi akan	berzauq	dan berasyik dan berma'asyuk bila tiba malam giliran ...
9	118540#947	... adalah di masjid bagi bertukar pandangan, memberikan bantuan,	bertanyakan	khabar dan berkenal-kenalan.
10	109724#6	Wartawan Pancaindera yang	bertemu	Azura dan Haliza baru-baru ini berkesempatan berbicara

No.	No. Rekod	Konteks Kiri	Kata	Konteks Kanan
1917411	108024#3	Sultan Salahuddin	bertitah	, rakyat Malaysia mestilah berupa
1917412	12213#3	... mencecah tahap tertinggi 246,554 kontrak pada Oktober lalu	berbanding	tahap tertinggi sebelum itu, 187,7
1917413	17375#0	... ditenggelami air dan banyak daerah tenggelam, kata agensi	berita	BSS.
1917414	76824#6	... pada 5 Januari 1998 adalah rentetan apa yang	berlaku	sebelumnya.
1917415	35794#		berusia	kurang 18 tahun menjalani temudu
1917416	3139#1		bertujuan	mengekalkan ciri semula jadi kawa
1917417	44900#		bersalah	terhadap tuduhan yang dikenakan
1917418	44719#		berharap	untuk mendengar sesuatu yang be
1917419	114685		berlaku	kepada Anwar semasa dia ditahan
1917420	93847#		berpindah	ke Amerika Syarikat pada tahun d

Paparan Keseluruhan Ayat - Google Chrome
不安全
sbmb.dbp.gov.my/korpusdbp/Popups/KataView2.aspx?loc=76824-6-F
Keseluruhan Ayat
Keseluruhan Perenggan

Katanya, ugutan dan gangguan yang dihadapi oleh Sharma bermula Oktober 1997 dan kejadian pada 5 Januari 1998 adalah rentetan apa yang **berlaku** sebelumnya.

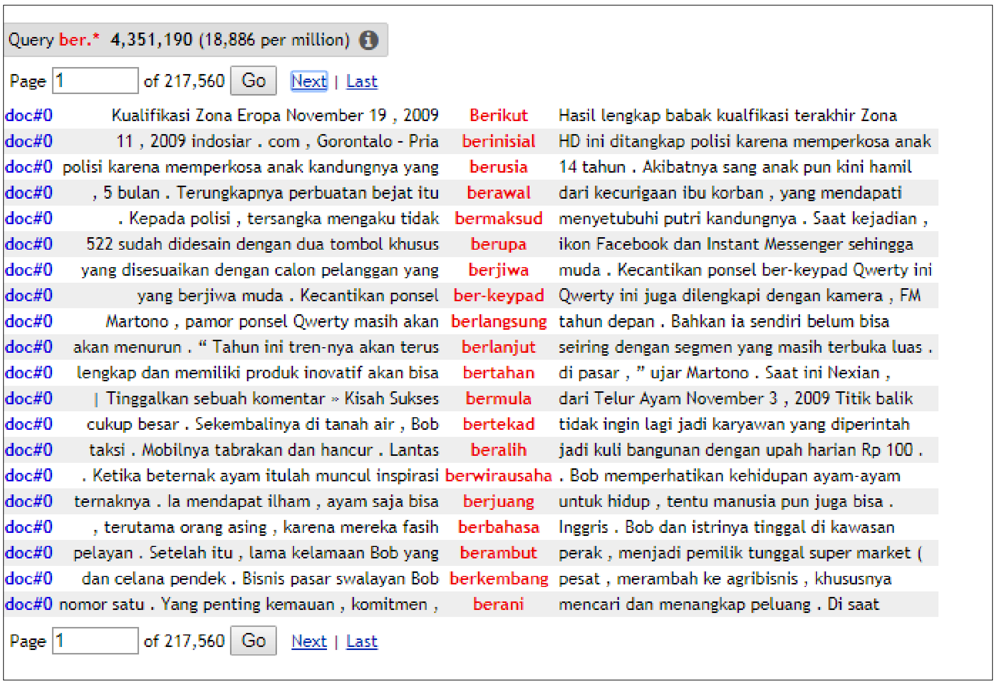
Figure 1. The DBP Corpus

The non-availability of morphology searches could have been due to the limitation of the user's access, or it could have also been due to the fact that the corpus is not annotated in terms of morphology and POS information.

In addition to the DBP Corpus, other corpora that have been collected by different researchers are listed in Table 1. Among these, the largest Malay corpus (230 million words) is the MalaysianWaC Corpus provided in the Sketch Engine interface (Kilgarriff & Tugwell, 2002). Although this corpus may be the largest corpus of Malay, it is not a corpus suitable for the analysis of Malay prefixes or lemmas for several reasons. First, the corpus is only labelled with shallow tagsets for the Wordsketch function (nouns, verbs, adjectives, adverbs, etc., according to the "Apertium Indonesian and Malaysian tagset") (taken from the description of the corpus). For concordance, when we searched for *ber** in the corpus, the results in Figure 2 were returned. In the final line in Figure 2, we found *berani* 'brave', which is not a prefix use. For the use of this corpus, the elimination of noise is needed. Second, the web results were not entirely Malaysian Malay. This is common in web data as the writer's native language is unknown. The most powerful functions of Sketch Engine, the Word Sketch and Sketch-Diff functions, which provide a quick sketch of the linguistic behaviors of one or more searched words, do not work for Malay. Therefore, the zsmWaC Corpus only offers a collection of web information with searches of running texts—the same kind of data that we could obtain by using a self-collected corpus running on any concordancer, though on a smaller scale.

Table 1. Malay corpora

Corpus	Feature	Website
MalaysianWac (or zsmWac) Corpus, Sketch Engine (Kilgarriff et al. 2010)	Non-tagged raw texts	https://www.sketchengine.eu/zsmwac-malaysian-corpus/
SEAlang Library Malay Text Corpus	Non-tagged raw texts	http://sealang.net/malay/corpus.htm
Malay Practical Grammar Corpus (MPGC) (Imran Ho Abdullah et al. 2004)	A section of the DBP Corpus	Not available
MALEX (MALay LEXicon) (Yap et al. 2010; Zuraidah Mohd Don, 2010)	A list of Malay lexicons (with morphology information), with English translations	Not available
Malay textbooks for primary schools (Lee & Low 2011)	Malaysian language textbooks for primary schools	http://www.mybaca.org/

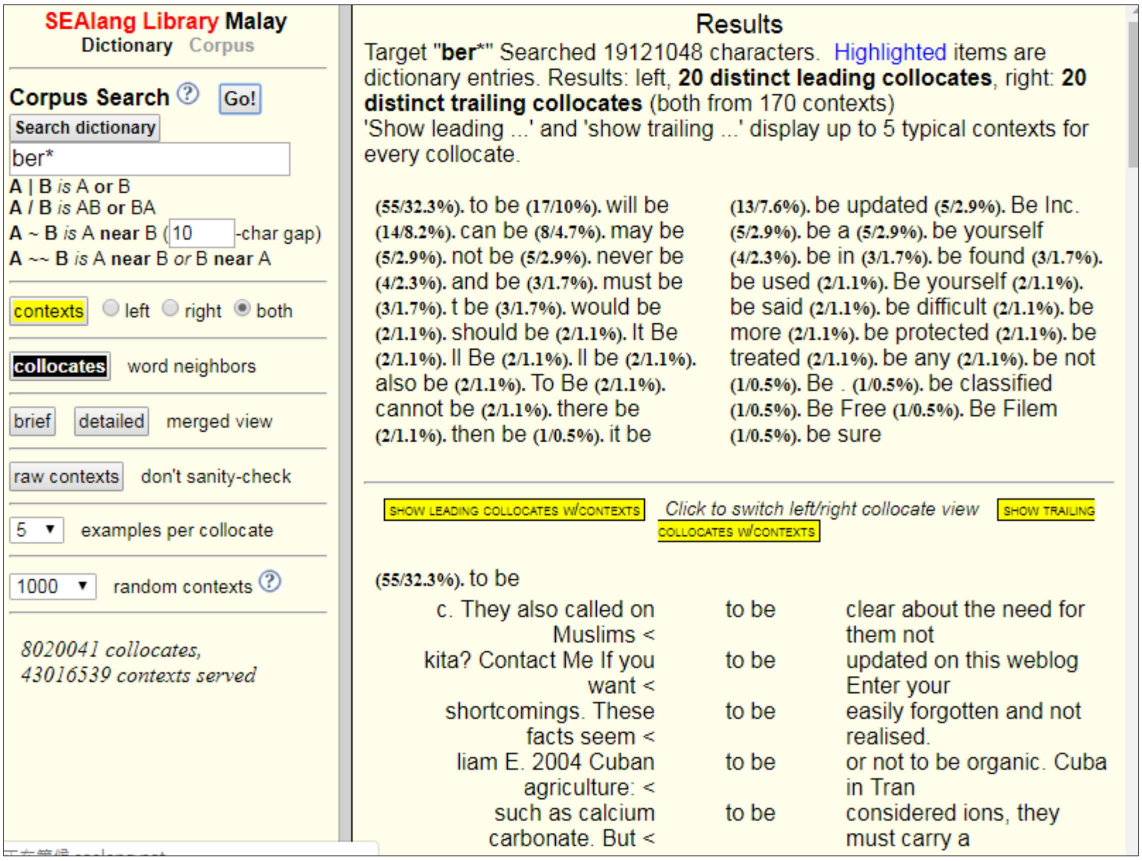


Query ber.* 4,351,190 (18,886 per million)

doc#0	Kualifikasi Zona Eropa November 19 , 2009	Berikut	Hasil lengkap babak kualfikasi terakhir Zona
doc#0	11 , 2009 indosiar . com , Gorontalo - Pria	berinisial	HD ini ditangkap polisi karena memperkosa anak
doc#0	polisi karena memperkosa anak kandungnya yang	berusia	14 tahun . Akibatnya sang anak pun kini hamil
doc#0	, 5 bulan . Terungkapnya perbuatan bejat itu	berawal	dari kecurigaan ibu korban , yang mendapati
doc#0	. Kepada polisi , tersangka mengaku tidak	bermaksud	menyetubuhi putri kandungnya . Saat kejadian ,
doc#0	522 sudah didesain dengan dua tombol khusus	berupa	ikon Facebook dan Instant Messenger sehingga
doc#0	yang disesuaikan dengan calon pelanggan yang	berjiwa	muda . Kecantikan ponsel ber-keypad Qwerty ini
doc#0	yang berjiwa muda . Kecantikan ponsel	ber-keypad	Qwerty ini juga dilengkapi dengan kamera , FM
doc#0	Martono , pamor ponsel Qwerty masih akan	berlangsung	tahun depan . Bahkan ia sendiri belum bisa
doc#0	akan menurun . “ Tahun ini tren-nya akan terus	berlanjut	seiring dengan segmen yang masih terbuka luas .
doc#0	lengkap dan memiliki produk inovatif akan bisa	bertahan	di pasar , ” ujar Martono . Saat ini Nexian ,
doc#0	| Tinggalkan sebuah komentar » Kisah Sukses	bermula	dari Telur Ayam November 3 , 2009 Titik balik
doc#0	cukup besar . Sekembalinya di tanah air , Bob	bertekad	tidak ingin lagi jadi karyawan yang diperintah
doc#0	taksi . Mobilnya tabrakan dan hancur . Lantas	beralih	jadi kuli bangunan dengan upah harian Rp 100 .
doc#0	. Ketika beternak ayam itulah muncul inspirasi	berwirausaha	. Bob memperhatikan kehidupan ayam-ayam
doc#0	ternaknya . Ia mendapat ilham , ayam saja bisa	berjuang	untuk hidup , tentu manusia pun juga bisa .
doc#0	, terutama orang asing , karena mereka fasih	berbahasa	Inggris . Bob dan istrinya tinggal di kawasan
doc#0	pelayan . Setelah itu , lama kelamaan Bob yang	berambut	perak , menjadi pemilik tunggal super market (
doc#0	dan celana pendek . Bisnis pasar swalayan Bob	berkembang	pesat , merambah ke agribisnis , khususnya
doc#0	nomor satu . Yang penting kemauan , komitmen ,	berani	mencari dan menangkap peluang . Di saat

Figure 2. Search results for *ber in the zsmWaC Corpus in the Malaysian Sketch Engine**

The SEAlang Library Malay Text Corpus is a “monolingual corpus that consists of Malay texts retrieved from a variety of Internet sources” (cited from http://sealang.net/malay/corpus.htm). It is said to contain Scannell’s (2007) corpus (about 2.5 million words) collected from the web using a crawler. The SEAlang Corpus displays co-appearing patterns of the searched words (which are called ngrams), but it cannot deal with prefixes. When we searched for *ber*, we found only three instances (accessed August 23). When we searched for *ber**, it returned many English examples, such as ‘to be’, ‘can be’, and ‘may be’ (see Figure 3), indicating that this corpus contains many English sentences, which is a serious problem if it is used as a monolingual Malay corpus.[8] However, when we searched for a specific word such as *sayur* ‘vegetable’, we were given the correct ngram patterns (e.g., *dan sayur* ‘and vegetable’). However, a corpus linguist needs more than these co-occurring patterns.



SEAlang Library Malay
Dictionary Corpus

Corpus Search ? Go!
Search dictionary
ber*
A | B *is* A or B
A / B *is* AB or BA
A ~ B *is* A **near** B (10 -char gap)
A ~~ B *is* A **near** B *or* B **near** A

contexts ○ left ○ right ● both

collocates word neighbors

brief detailed merged view

raw contexts don't sanity-check

5 ▼ examples per collocate

1000 ▼ random contexts ?

8020041 collocates,
43016539 contexts served

Results

Target "**ber***" Searched 19121048 characters. Highlighted items are dictionary entries. Results: left, **20 distinct leading collocates**, right: **20 distinct trailing collocates** (both from 170 contexts)
'Show leading ...' and 'show trailing ...' display up to 5 typical contexts for every collocate.

(55/32.3%). to be (17/10%). will be (14/8.2%). can be (8/4.7%). may be (5/2.9%). not be (5/2.9%). never be (4/2.3%). and be (3/1.7%). must be (3/1.7%). t be (3/1.7%). would be (2/1.1%). should be (2/1.1%). It Be (2/1.1%). ll Be (2/1.1%). ll be (2/1.1%). also be (2/1.1%). To Be (2/1.1%). cannot be (2/1.1%). there be (2/1.1%). then be (1/0.5%). it be

(13/7.6%). be updated (5/2.9%). Be Inc. (5/2.9%). be a (5/2.9%). be yourself (4/2.3%). be in (3/1.7%). be found (3/1.7%). be used (2/1.1%). Be yourself (2/1.1%). be said (2/1.1%). be difficult (2/1.1%). be more (2/1.1%). be protected (2/1.1%). be treated (2/1.1%). be any (2/1.1%). be not (1/0.5%). Be . (1/0.5%). be classified (1/0.5%). Be Free (1/0.5%). Be Filem (1/0.5%). be sure

SHOW LEADING COLLOCATES W/CONTEXTS *Click to switch left/right collocate view* SHOW TRAILING COLLOCATES W/CONTEXTS

(55/32.3%). to be

c. They also called on Muslims <	to be	clear about the need for them not
kita? Contact Me If you want <	to be	updated on this weblog Enter your
shortcomings. These facts seem <	to be	easily forgotten and not realised.
liam E. 2004 Cuban agriculture: <	to be	or not to be organic. Cuba in Tran
such as calcium carbonate. But <	to be	considered ions, they must carry a

Figure 3. SEAlang Library Malay Text Corpus

For POS taggers, there have been attempts to create Malay taggers (Knowles & Zuraidah Mohd Don 2003; Norshuhani Zamin et al. 2012; Rayner, Mujat & Obit 2013), and different versions of tagsets have been found. We reviewed those in accordance with the needs of the current work. Knowles and Zuraidah Mohd Don (2003: 424; 2006) proposed a tagset for Malaysian Malay to cope with the “syntactic drift” in Malay, such as that found in the following example: “*masuk* is the normal word for ‘enter’, which makes it a kind of verb; but it is used in such a way on buildings and in carparks that it could also be taken to be a noun ‘entrance’.” To overcome this phenomenon, Knowles and Zuraidah Mohd Don (2003: 424) proposed tagging the DBP Corpus by analyzing sentences:

> For example, in *bulan samar* ‘dim moon’, the ‘adjective’ *samar* behaves as expected and follows the noun as a modifier. In *Seman terlalu gembira* ‘Seman was extremely happy’, the ‘adjective’ *gembira* follows the intensifier *terlalu*. The English translation makes it still look like an adjective, but the structure is one of a large set relating to the verbal group, and our parser treats

[8] We thank the reviewer for this added explanation. Another reviewer also pointed out that the use of * meant “zero or more repetition of the preceding character.” In this case, the minimum match was *be* in this corpus. We thank the reviewer for this added explanation.

gembira as a kind of ‘verb’. In *ibu bapanya membangkang keras* ‘his mother and father disagree strongly’ the parser treats the ‘adjective’ *keras* as an ‘adverb’ after the ‘verb’ *membangkang*.

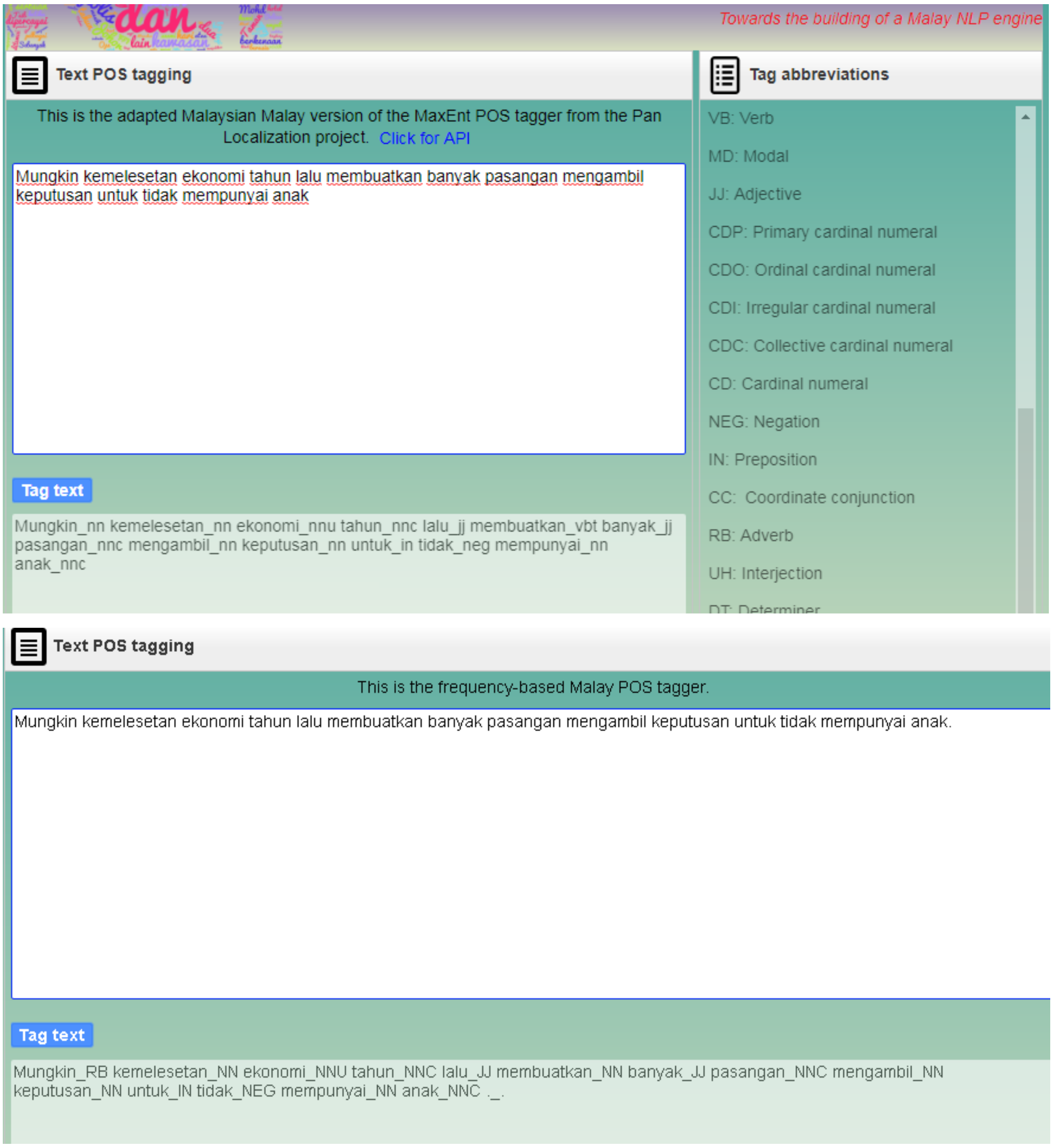
This method solved some ambiguity problems, but it did not provide a morphological analysis of the words. For example, more information is needed to distinguish [*masuk*_VERB] from its other forms, such as [*ke*+*masuk*+*an*_NOUN]. If a tag is only provided for *ke-masuk-an* without marking the root *masuk* as a verb, it will lose the morphological information within it.

On the other hand, Chu et al. (2016: 115) created Mi-POS, a Malay part-of-speech tagger using a “probabilistic approach with information from the context.” The tagger revised Norshuhani Zamin et al.’s (2012) “Lazy Man’s Way” tagset by simplifying several verb tags (i.e., MD, VB, VBD, VBG, VBN, VBP, and VBZ) under the “VB” tagset. The Mi-POS tagset has some similarities with the tagset in the Malay NLP tagger, the first tagger that was available online, developed by Rohana Binti Mahmud and her colleague Mohamed Lubani in the Department of Artificial Intelligence, University of Malaya. The online version was used in the first approach of the current study (http://malaynlp.appspot.com/), which will be introduced and elaborated in the next section. One result is displayed in Figure 4.

As can be seen, although an online tagger was eagerly hoped for, the tagger has some inaccuracies that cannot be avoided. The Malay NLP project has two versions of Malay taggers—one is based on the “MaxEnt POS tagger from the Pan Localization project” (cited from http://malaynlp.appspot.com//, accessed September 1), and the other is frequency-based. For the former, when we inserted a short sentence into the system as shown in Figure 4 (see the top screenshot), we found that some words were incorrectly tagged, namely ‘mungkin_nn’, ‘mengambil_nn’, and ‘mempunyai_nn’. Although this tagger has everything that we were looking for (e.g., online tagger, API upon request, etc.), it would require a high cost of human correction if all words needed to be checked.[9] The frequency-based version provided a better result (see bottom screenshot of Figure 4), but it did not include an API that allowed us to tag our own corpus.[10]

[9] Assistance from Dr. Rohana Binti Mahmud through Mohamed Lubani (http://malaynlp.appspot.com/) was highly appreciated. We are sincerely grateful for the availability and extension of the API given by the Malay NLP project team. Comments on the suitability of the tool relied completely on the suitability of the tool for our immediate needs. We were told (Mohamed Lubani, personal communication) that the MaxEnt tagger “is trained to optimise the assignment of a set of POS tags to a given sequence of words.” Manual examination was strongly encouraged.

[10] When we accessed the tagger again on September 25, the website showed that the domain name had expired on September 4, 2019 “and is pending renewal or deletion.” This, again, shows how inconsistent the availability of resources for Malay are, and further proves the need to have an annotated corpus of our own.



This is the adapted Malaysian Malay version of the MaxEnt POS tagger from the Pan Localization project.

Mungkin kemelesetan ekonomi tahun lalu membuatkan banyak pasangan mengambil keputusan untuk tidak mempunyai anak

Mungkin_nn kemelesetan_nn ekonomi_nnu tahun_nnc lalu_jj membuatkan_vbt banyak_jj pasangan_nnc mengambil_nn keputusan_nn untuk_in tidak_neg mempunyai_nn anak_nnc

VB: Verb
MD: Modal
JJ: Adjective
CDP: Primary cardinal numeral
CDO: Ordinal cardinal numeral
CDI: Irregular cardinal numeral
CDC: Collective cardinal numeral
CD: Cardinal numeral
NEG: Negation
IN: Preposition
CC: Coordinate conjunction
RB: Adverb
UH: Interjection
DT: Determiner

This is the frequency-based Malay POS tagger.

Mungkin kemelesetan ekonomi tahun lalu membuatkan banyak pasangan mengambil keputusan untuk tidak mempunyai anak.

Mungkin_RB kemelesetan_NN ekonomi_NNU tahun_NNC lalu_JJ membuatkan_NN banyak_JJ pasangan_NNC mengambil_NN keputusan_NN untuk_IN tidak_NEG mempunyai_NN anak_NNC ._.

Figure 4. Part-of-speech tagger from the Malay NLP project

As shown in Figure 4, we could potentially obtain the POS information of each word from the POS tagger, such as *tidak* as a "negator" (neg) and *banyak* as an "adjective" (jj). However, we also needed the root and morphology information for *mengambil*, which is a transitive verb, as in the combination of [*meng*+*ambil* (v)]. For this purpose, we found another tool that might help us solve the problem, the morphology analyzer created by Tan and colleagues (2017). The output of this analyzer was made available to us through personal communication.[11] The morphology analyzer separated the affixes from the root,

[11] We obtained some output results from Dr. Tien Ping Tan and his colleagues in the School of Computer Sciences, Universiti Sains Malaysia. We are grateful for their help. We also thank Dr. Michael Tanangkingsing from the National Taipei University of Technology and Dr. Jasmina Khaw Yen Min from the Department of Computer Science, Universiti Tunku Abdul Rahman, Malaysia, who provided the connections for their cooperation.

which is something we needed in order to analyze prefixes and the root. This tool and the Malay NLP tagger were used in this study as the first approach with which to annotate our Malay corpus. The results will be reported in the next section.

In addition to the above resources, we also searched for the suitability of resources in Indonesian. We found several open resources that we could access and that were easy to use, and some platforms (e.g., MALINDO Conc) provided a mixture of Malaysian Malay and Indonesian data. These resources are as follows:

(1) a. Leipzig Corpora Collection (Goldhahn, Eckart & Quasthoff 2012)

b. MALINDO Morph (Nomoto et al. 2018)

c. MALINDO Conc (Nomoto, Akasegawa & Shiohara 2018a, b), a reclassification of the Leipzig Corpora Collection

d. MorphInd POS tagger (Larasati, Kuboň & Zeman 2011)

Figure 5 shows the Indonesian component of the Leipzig Corpora Collection (Goldhahn, Eckart & Quasthoff 2012).[12] The information on the website stated that up to 2013, there were 74,329,815 sentences, 7,964,109 types, and 1,206,281,985 tokens in the Indonesian component (accessed September 2019). The platform does not allow searches for affixes or lemmas, but when we typed in *ber-kata* 'ber-say' (without the hyphen in actual written form), we were given the results shown in Figure 5 below. The examples came from online resources.

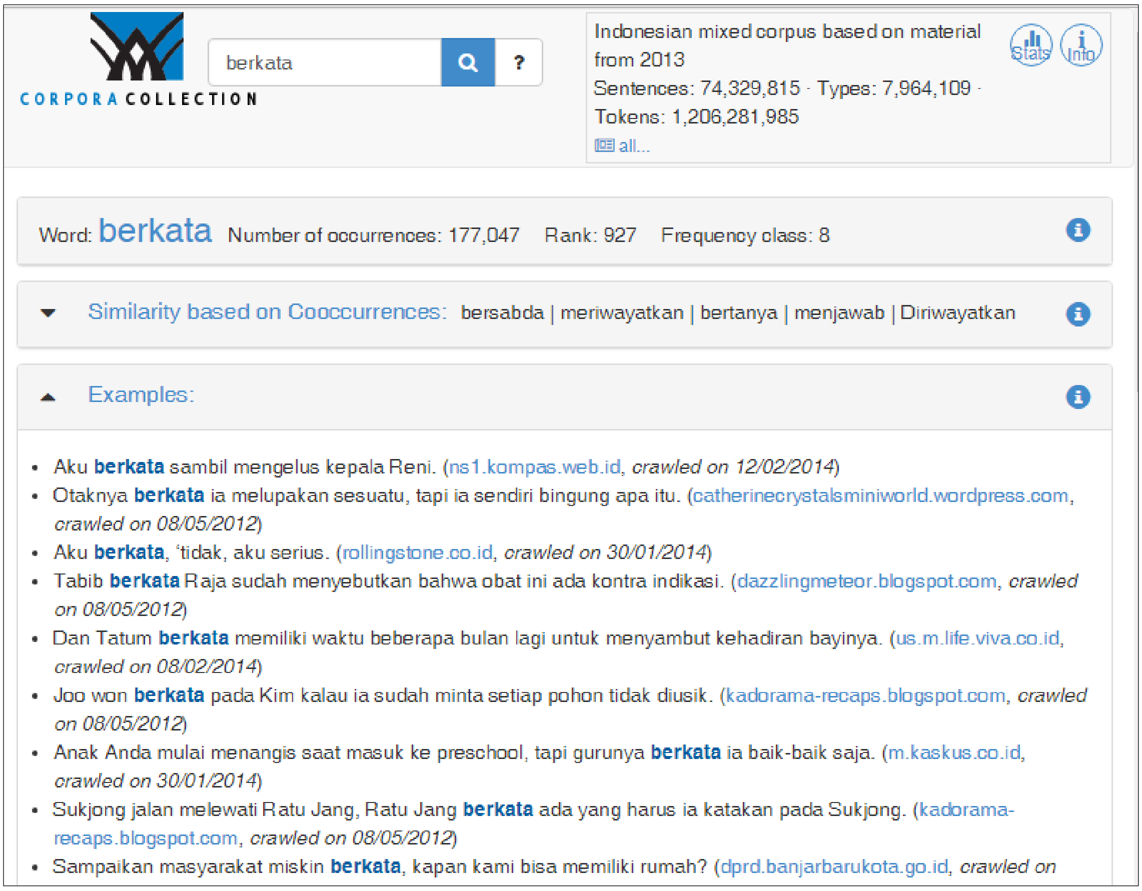
The Leipzig Corpora Collection was later reclassified by Nomoto, Akasegawa & Shiohara (2018a, b) for the creation of MALINDO Conc.[13] The morphological annotation of the searchable corpora using MALINDO Conc is based on MALINDO Morph (Nomoto et al. 2018), a "morphology dictionary," and it is equipped with prefix searches (see Figure 6). Furthermore, this platform separated the Leipzig Corpora Collection according to the Indonesian (IND) and Malaysian (ZSM) websites, which facilitated the research of both languages.[14] Figure 6 below (see the left side) shows the selection of affixes and the output presented (on the right side). This corpus has the affix search option that we had hoped for, but it is not POS-tagged at the current stage (accessed September 23).

As for Indonesian POS taggers, MorphInd (Larasati Kuboň & Zeman 2011) is one of the most complete tools we have encountered so far, as it is equipped with both a morphology analyzer and a POS tagging system. The results of this tool will be evaluated in the following section.

[12] https://corpora.uni-leipzig.de/en?corpusId=ind_mixed_2013

[13] https://malindo.aa-ken.jp/conc/

[14] However, as mentioned, all website information presents some difficulties in distinguishing the writer's identity, and even websites that have a regional domain such as ".id" or ".my" may not guarantee their source of language. Therefore, it was sometimes difficult to tell whether the language was Malaysian or Indonesian.



Indonesian mixed corpus based on material from 2013
Sentences: 74,329,815 · Types: 7,964,109 · Tokens: 1,206,281,985

Word: berkata Number of occurrences: 177,047 Rank: 927 Frequency class: 8

Similarity based on Cooccurrences: bersabda | meriwayatkan | bertanya | menjawab | Diriwayatkan

Examples:

- Aku **berkata** sambil mengelus kepala Reni. (ns1.kompas.web.id, *crawled on 12/02/2014*)
- Otaknya **berkata** ia melupakan sesuatu, tapi ia sendiri bingung apa itu. (catherinecrystalsminiworld.wordpress.com, *crawled on 08/05/2012*)
- Aku **berkata**, 'tidak, aku serius. (rollingstone.co.id, *crawled on 30/01/2014*)
- Tabib **berkata** Raja sudah menyebutkan bahwa obat ini ada kontra indikasi. (dazzlingmeteor.blogspot.com, *crawled on 08/05/2012*)
- Dan Tatum **berkata** memiliki waktu beberapa bulan lagi untuk menyambut kehadiran bayinya. (us.m.life.viva.co.id, *crawled on 08/02/2014*)
- Joo won **berkata** pada Kim kalau ia sudah minta setiap pohon tidak diusik. (kadorama-recaps.blogspot.com, *crawled on 08/05/2012*)
- Anak Anda mulai menangis saat masuk ke preschool, tapi gurunya **berkata** ia baik-baik saja. (m.kaskus.co.id, *crawled on 30/01/2014*)
- Sukjong jalan melewati Ratu Jang, Ratu Jang **berkata** ada yang harus ia katakan pada Sukjong. (kadorama-recaps.blogspot.com, *crawled on 08/05/2012*)
- Sampaikan masyarakat miskin **berkata**, kapan kami bisa memiliki rumah? (dprd.banjarbarukota.go.id, *crawled on*

Figure 5. The Indonesian component the of Leipzig Corpora Collection (Goldhahn, Eckart & Quasthoff 2012)

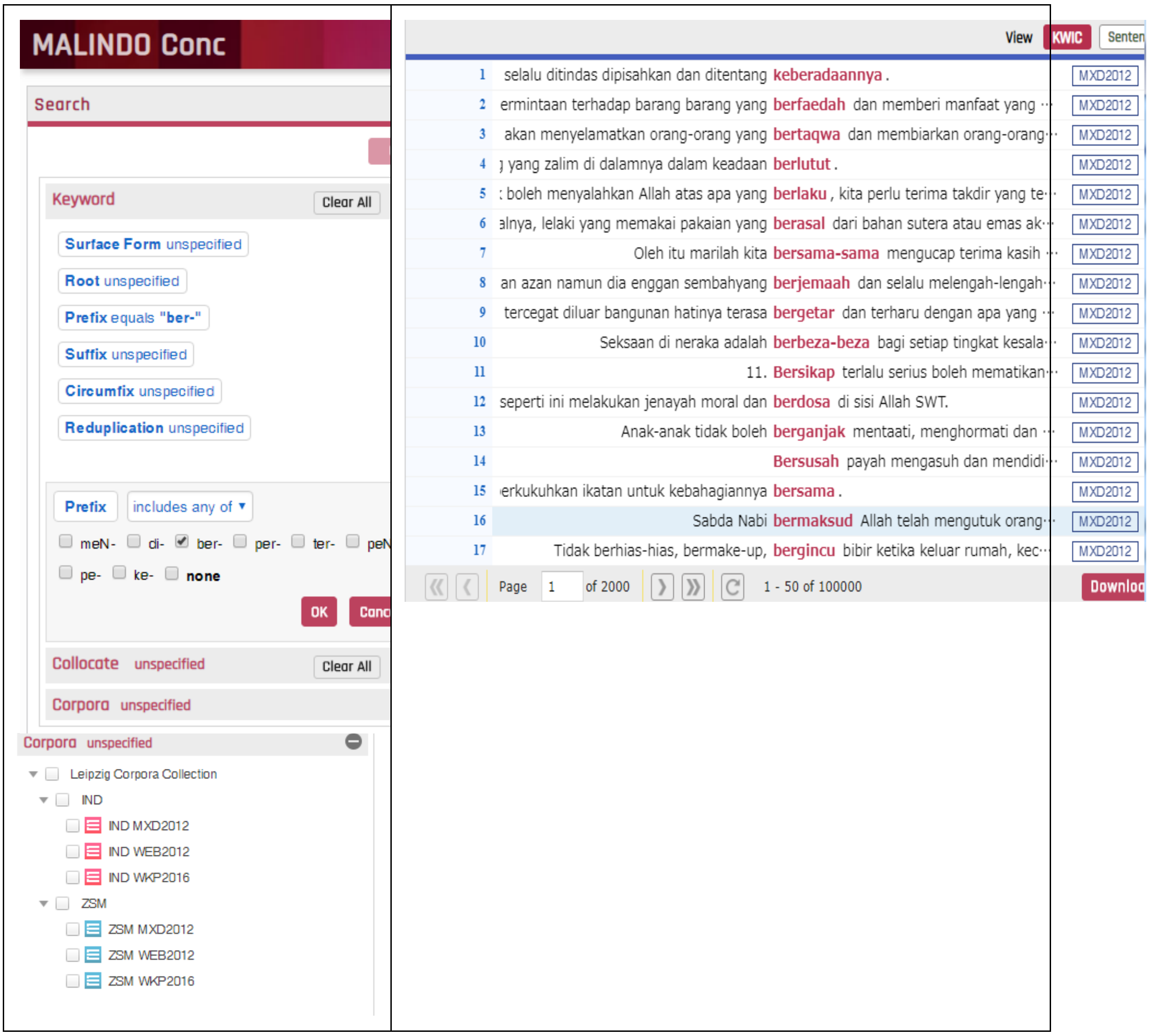
Based on the survey above, it is safe to say that the development of Malaysian corpora has largely been dependent on the DBP Corpus, and even if the corpora included POS taggers and morphology analyzers, they were rarely open resources. In this study, we processed the Malay corpus that has been used in Chung (2010, 2011, 2013, 2014, 2019), which were raw texts from Malay newspaper articles. We annotated the corpus first, and later we will show how research on the Malay prefix *ber-* was carried out.

3. Annotation of a Malaysian Malay news corpus

The corpus used in this work for data analysis was a Malay corpus containing 35,767 newspaper articles collected from *Utusan Malaysia* (a national Malaysian newspaper) between December 16, 2010 and June 14, 2011.[15] All of these newspaper articles were published online and were searchable in the newspaper's archives.

Since news and articles in newspapers may not be entirely news-related, there were features, fictions, discussions, forums, and all other non-news columns in the corpus. We use the term Standard Malaysian Malay here to refer to the language of the corpus, but there might be colloquial uses that are considered not-so-standard by some.

[15] Here, we would like to acknowledge a typo made in Chung (2014) on the year the corpus containing 35,767 newspaper articles was used. It should have been December 16, 2010 to June 14, 2011 (not January 2011).



MALINDO Conc

Search

Keyword Clear All

Surface Form unspecified

Root unspecified

Prefix equals "ber-"

Suffix unspecified

Circumfix unspecified

Reduplication unspecified

Prefix includes any of ▾

□ meN- □ di- ☑ ber- □ per- □ ter- □ peN
□ pe- □ ke- □ none

OK Canc

Collocate unspecified Clear All

Corpora unspecified

Corpora unspecified

- □ Leipzig Corpora Collection
  - □ IND
    - □ IND MXD2012
    - □ IND WEB2012
    - □ IND WKP2016
  - □ ZSM
    - □ ZSM MXD2012
    - □ ZSM WEB2012
    - □ ZSM WKP2016

View KWIC Senten

1	selalu ditindas dipisahkan dan ditentang	keberadaannya	.	MXD2012
2	ermintaan terhadap barang barang yang	berfaedah	dan memberi manfaat yang ···	MXD2012
3	akan menyelamatkan orang-orang yang	bertaqwa	dan membiarkan orang-orang···	MXD2012
4	) yang zalim di dalamnya dalam keadaan	berlutut	.	MXD2012
5	: boleh menyalahkan Allah atas apa yang	berlaku	, kita perlu terima takdir yang te···	MXD2012
6	alnya, lelaki yang memakai pakaian yang	berasal	dari bahan sutera atau emas ak···	MXD2012
7	Oleh itu marilah kita	bersama-sama	mengucap terima kasih ···	MXD2012
8	an azan namun dia enggan sembahyang	berjemaah	dan selalu melengah-lengah···	MXD2012
9	tercegat diluar bangunan hatinya terasa	bergetar	dan terharu dengan apa yang ···	MXD2012
10	Seksaan di neraka adalah	berbeza-beza	bagi setiap tingkat kesala···	MXD2012
11	11.	Bersikap	terlalu serius boleh mematikan···	MXD2012
12	seperti ini melakukan jenayah moral dan	berdosa	di sisi Allah SWT.	MXD2012
13	Anak-anak tidak boleh	berganjak	mentaati, menghormati dan ···	MXD2012
14		Bersusah	payah mengasuh dan mendidi···	MXD2012
15	erkukuhkan ikatan untuk kebahagiannya	bersama	.	MXD2012
16	Sabda Nabi	bermaksud	Allah telah mengutuk orang···	MXD2012
17	Tidak berhias-hias, bermake-up,	bergincu	bibir ketika keluar rumah, kec···	MXD2012

Page 1 of 2000 1 - 50 of 100000 Downloa

Figure 6. Search interface of MALINDO Conc (Nomoto, Akasegawa & Shiohara 2018a, b)

This corpus was tagged using two approaches. For the first approach, it was lemmatized or stemmed using the morphology analyzer provided by Tan et al. (2017), and later tagged with the Malay NLP tool provided by colleagues in the Department of Artificial Intelligence, University of Malaya.

For the second approach, it was tagged using the MorphInd POS tagger online (Larasati, Kuboň & Zeman 2011).[16] The following examples in (2) show the output from different resources for the same two sentences:

(2) **Original sentence**

Mungkin kemelesetan ekonomi tahun lalu membuatkan banyak pasangan mengambil keputusan untuk tidak mempunyai anak.

'Maybe the economy recession last year caused many couples to decide not to have a child.'

[16] The tagsets of Larasati, Kuboň & Zeman (2011) are slightly different from the online version (http://septinalarasati.com/MorphInd/). We used the online tagsets in this study.

(a) **Malay NLP tool**[17]

(i) MaxEnt version:

Mungkin_nn kemelesetan_nn ekonomi_nnu tahun_nnc lalu_jj membuatkan_vbt banyak_jj pasangan_nnc mengambil_nn keputusan_nn untuk_in tidak_neg mempunyai_nn anak_nnc ._.

(ii) Frequency-based version:

Mungkin_RB kemelesetan_NN ekonomi_NNU tahun_NNC lalu_JJ membuatkan_NN banyak_JJ pasangan_NNC mengambil_NN keputusan_NN untuk_IN tidak_NEG mempunyai_NN anak_NNC

(b) **Morphology analyzer by Tan et al. (2017)**

Mungkin ke+meleset+an ekonomi tahun lalu mem+buat+kan banyak pasang+an meng+ambil ke+putus+an untuk tidak mem+punya+i anak.

(c) **MorphInd POS tagger**

mungkin<f>_F--$ ^ kemelesetan<x>_X--$ ^ekonomi<n>_NSD$
^tahun<n>_NSD$ ^lalu<a>_ASP$ ^meN+buat<v>+kan_VSA$
^banyak<a>_ASP$ ^pasang<v>+an_NSD$ ^meN+ambil<v>_VSA$
^ke+putus<a>+an_NSD$ ^untuk<r>_R--$ ^tidak<g>_G--$
^meN+punya<v>+i_VSA$ ^anak.<f>_F--$

From the results, we could see that we needed to process the data using two separate programs (2a and 2b) in the first approach, while there was only one program (2c) needed for the second approach. Although combining the two programs in the first approach was possible, we used the Indonesian MorphInd POS tagger in the end. The advantages and disadvantages of the two approaches are summarized in Table 2 below:

Table 2. Comparison of the two annotation approaches used in the study

Morphology Analyzer by Tan et al. (2017) + Malay NLP Tagger	**MorphInd POS Tagger (Larasati, Kuboň & Zeman 2011)**
Advantages: • Tools based on Malaysian Malay • Fewer unknown words	**Advantages:** • Morphology and POS are processed at the same time • Higher accuracy of POS tagging • The POS for roots is provided (e.g., ^peN+sokong<v>_NSD$)
Disadvantages: • Morphology and POS are processed in separate programs • Higher technical cost of combining the two results • High cost of human corrections of incorrect tags	**Disadvantages:** • Many unknown words (Malaysian words not in Indonesian, e.g., ^parlimen<x>_X--$) • Proper nouns are not recognized (e.g., ^kuala<f>_F--$ ^lumpur<n>_NSD$)

Due to the high cost of computing and the need to manually check all parts-of-speech in the first approach, we chose the second approach. Although MorphInd is an Indonesian

[17] "This is the adapted Malaysian Malay version of the MaxEnt POS tagger from the Pan Localization project." http://engine.malaynlp.com/pos_maxent.zul (accessed September 9, 2019).

POS tagger, its stemmers and tagsets are considerably accurate; thus, it was more economical to have both morphology analyzing and tagging at the same time. These were the main reasons for selecting MorphInd. The following section will present an evaluation of the corpus data after it was completely tagged by running MorphInd.

4. Results: Evaluation of applying MorphInd taggers to Malaysian Malay

From the data collected, we successfully tagged 35,767 files, amounting to 13,979,859 words (delimited by space). In order to understand the tagsets of MorphInd, we will first provide their coding principle below (cited from http://septinalarasati.com/MorphInd, accessed February 25, 2019), and the MorphInd tagset list shown in Table 3 below was taken verbatim from the documentation on the website (accessed September 9, 2019):

> MorphInd has a fine-grained tagset which was inspired by the PENN Treebank tagset and adapted accordingly for Indonesian morphology. The tagset also adopts the concept of positional tags of the Prague Dependency Treebank tagset to cover most of the language behaviors that occur simultaneously in a surface word. Given in the table below is the complete MorphInd tagset.

The tagsets have feminine (F) and masculine (M) tags for nouns ending with -*wan* and -*wati* from Sanskrit.

From the test results in (2c) earlier, in addition to the POS tag, there was also the root tag, called the "lemma tag," in lower-case font. In (3) below, the lemma tags are respectively verb <v> for *buat* 'make/do' and adjective <a> for *putus* 'break'. The POS tag for *mem-buat-kan* is a "singular active verb" (VSA), and for *ke-putus-an* it is a "non-specified singular noun" (NSD).

(3)	*mem-buat-kan* ^meN+buat<v>+kan_VSA$	*ke-putus-an* ^ke+putus<a>+an_NSD$

From the above two words, it can be seen that the lemma tag is different from the overall tag of the word—*putus* is an adjective <a> but *ke-putus-an* is a noun (NSD). Most words started with the symbol "^" and ended with the symbol "$." For instance, the relative marker *yang*, which topped the wordlist as the most frequent word, was tagged as ^yang<s>_--S$ in the corpus, which marked the "subordinating conjunction" as both the lemma "<s>" and the whole word "--S" (if only one upper-case tag was needed, dashes were placed in the remaining two slots).The complete lemma tags are shown in Table 4 below. They are rather consistent with the POS tags in Table 3, but they are in lower-case font and appear after the root of the word—*buat*<v> and *putus*<a>—as shown above.

We used MorphInd to tag our entire corpus. For the results that follow, we used a frame to run the statistics. We wrote a Python program assuming a structured frame of four parts for each Malay word—an optional first prefix, an optional second prefix, an optional lemma (including a lemma name, a lemma tag, an optional suffix, and a POS tag), and a required lemma (also including a lemma name, a lemma tag, an optional suffix, and a POS tag)—as demonstrated in (4) below:[18]

[18] The Python matching pattern was re.search(pattern=r'\^' + prefix_pattern % 1 + prefix_pattern % 2 + lemma1_tag1_suffix1_pos1_pattern + lemma2_tag2_suffix2_pos2_pattern, string=word).

Table 3. MorphInd part-of-speech tagsets

1st Position		2nd Position		3rd Position	
N	Noun	**P**	Plural	**F**	Feminine
		S	Singular	**M**	Masculine
				D	Non-Specified
P	Personal Pronoun	**P**	Plural	**1**	First Person
		S	Singular	**2**	Second Person
				3	Third Person
V	Verb	**P**	Plural	**A**	Active Voice
		S	Singular	**P**	Passive Voice
C	Numeral	**C**	Cardinal Numeral		
		O	Ordinal Numeral		
		D	Collective Numeral		
A	Adjective	**P**	Plural	**P**	Positive
		S	Singular	**S**	Superlative
H	Coordinating Conjunction				
S	Subordinating Conjunction				
F	Foreign Word				
R	Preposition				
M	Modal				
B	Determiner				
D	Adverb				
T	Particle				
G	Negation				
I	Interjection				
O	Copula				
W	Question				
X	Unknown				
Z	Punctuation				

Table 4. Lemma tagsets[19]

n	Noun	**h**	Coordinating Conjunction	**b**	Determiner	**o**	Copula
p	Personal Pronoun	**s**	Subordinating Conjunction	**d**	Adverb	**w**	Question
v	Verb	**f**	Foreign Word	**t**	Particle	**x**	Unknown
c	Numeral	**r**	Preposition	**g**	Negation	**z**	Punctuation
a	Adjective	**m**	Modal	**i**	Interjection		

[19] "Adjective" in the table on their website was given the tag "q" but in the script it was given the tag "a." We thus changed "q" to "a" in this table.

(4) A structured frame of four parts for each Malay word

1	**2**	**3**				**4**			
first prefix	**(second prefix)**	**(lemma-name)**	**(lemma tag)**	**(suffix)**	**(POS)**	**lemma name**	**lemma tag**	**(suffix)**	**POS**
meN+	ber+	None	None	None	None	*henti*	<a>	+kan	_VSA$
ber+	ke+	None	None	None	None	*mungkin*	<d>	+an	_VSA$
ber+	None	None	None	None	None	*surai*	<v>	None	_VSA$
None	None	None	None	None	None	*tuduh*	<v>	+an	_NSD$

If a Malay word had only one prefix, it was identified as the first prefix in our frame, and the value of the optional second prefix was given "None", as illustrated by the third row for the word ^ber+surai<v>_VSA$. In the results, the devised frame successfully matched 11,821,108 words among the 13,979,859 words (about 85%) in our news corpus. The remaining 15% that did not fall in this frame were patterns undetected by this frame. These will not be further analyzed in this paper but will be refined in a later version of this corpus; thus, we will not report the results of these 11,821,108 words in our corpus at this time.

Table 5 below shows all the lemma tags we captured with this frame. From the results, we found that nouns topped the lemma list, indicating that most often the morphological roots were types of nouns. The second highest was punctuation. Verbs ranked fourth and adjectives fifth. The unknown lemmas, as exemplified in Table 5, were words not recognized by the system. At 12.28%, this meant that there were about this proportion of Malay vocabulary not recognized by the Indonesian program, which was an additional result we found in this study—differences between Malay and Indonesian. Among these were Malaysian words not stored in the Indonesian lexicon (e.g., *kerusi* 'chair' in Malaysian Malay is *kursi* in Indonesian), or terminology used in Malaysian Malay, such as a large amount of borrowing (e.g., *moden* for 'modern', *kes* for 'case', and *kondominium* for 'condominium'), or words not used in Indonesian, such as *penguatkuasaan* in Malay.

As for the POS in the whole corpus, the results are given in Table 6 below. As can be seen, the majority of POS were also types of nouns (22%), followed by punctuation. Basically, the top four types were similar to the lemma tags in Table 5. However, the POS tags divided nouns and verbs into several groups, so the actual percentages were higher than those shown in Table 6. The fifth highest frequency in Table 6 is prepositions. Examples are also given in Table 6. As for adjectives (ASP), we could not determine why the label was "positive singular adjective" as some negative adjectives such as *gagal* 'fail' were found. Maybe the word "positive" had a different meaning, but no explanation was found for this in Larasati, Kuboň & Zeman (2011).

Table 5. All lemma tags in the corpus

Lemma	Lemma Name	Example	English Gloss	Frequency	Percentage
<n>	Noun	suara**<n>**_NSD$	sound	2,463,158	20.84%
<z>	Punctuation	^,**<z>**_Z--$	[comma]	1,488,378	12.59%
<x>	Unknown	^penguatkuasaan**<x>**_X--$	execution [pen-in.execution-an]	1,451,904	12.28%
<v>	Verb	^ber+kait**<v>**+an_VSA$	ber-relate-an ‘related’	1,345,415	11.38%
<a>	Adjective	^ke+penting**<a>**+an_NSD$	ke-important-an ‘importance’	992,674	8.40%
<r>	Preposition	^dengan**<r>**_R--$	with	870,164	7.36%
<f>	Foreign Word	^tv**<f>**_F--$	TV	656,353	5.55%
<s>	Subordinating Conjunction	^sebelum**<s>**_S--$	before	486,644	4.12%
<b>	Determiner	^itu**<b>**_B--$	that	413,663	3.50%
<p>	Personal Pronoun	^mereka**<p>**_PP3$	they	412,957	3.49%
<d>	Adverb	^memang**<d>**_D--$	indeed	339,943	2.88%
<c>	Numeral	^tujuh**<c>**_CC-$	seven	317,644	2.69%
<h>	Coordinating Conjunction	^atau**<h>**_H--$	or	270,921	2.29%
<g>	Negation	^tidak**<g>**_G--$	no(t)	106,242	0.90%
<m>	Modal	^akan**<m>**_M--$	will	89,445	0.76%
<o>	Copula	^adalah**<o>**_O--$	be	48,439	0.41%
<t>	Particle	^pun**<t>**_T--$	PUN	35,076	0.30%
<w>	Question	^mana**<w>**_W--$	where	27,039	0.23%
<i>	Interjection	^astaga**<i>**_I--$	gosh	5,049	0.04%
			Total	**11,821,108**	**100.00%**

Table 6. All POS tags in the corpus

POS	Name of POS	Examples	English Gloss	Frequency	Percentage
NSD	Non-Specified Singular Noun	^suara<n>_NSD$ ^ke+penting<a>+an_NSD$	sound ke+important<a>+an 'importance'	2,653,190	22.44%
Z	Punctuation	^"<z>_Z--$	[double quotation marks]	1,488,378	12.59%
X	Un-known	^moden<x>_X--$ ^teruk<x>_X--$ ^bahawa<x>_X--$ ^keputeraan<x>_X--$ ^segamat<x>_X--$	modern terrible that birthday of a prince Segamat [name of place]	1,451,887	12.28%
VSA	Active Singular Verb	^meN+ubah<v>_VSA$ ^ter+fikir<v>_VSP$ ^meN+jalan<v>+i_VSA	meN+change 'to change' ter+think 'suddenly.think.of.something' meN+walk<v>+i 'to walk'	1,206,688	10.21%
R	Preposition	^dari<r>_R--$ ^sebagai<r>_R--$ ^di<r>_R--$	from as at	870,280	7.36%
ASP	Positive Singular Adjec-tive	^penting<a>_ASP$ ^gagal<a>_ASP$	important fail	662,888	5.61%
F	Foreign Word	^malaysia<f>_F--$ ^datuk<f>_F--$ ^sultan<f>_F--$ ^hospital<f>_F--$	Malaysia Datuk [title] Sultan hospital	628,350	5.32%
S	Subordinating Conjunction	^tetapi<s>_S--$ ^yang<s>_S--$	but REL	486,448	4.12%
B	Determiner	^itu<b>_B--$ ^ini<b>_B--$ ^beberapa<b>_B--$	that this several	414,656	3.51%
D	Adverb	^sampai<d>_D--$ ^sebenarnya<d>_D--$ ^masih<d>_D--$	until actually still	339,638	2.87%
PS	Singular Personal Pronoun	^sirip<n>_NSD+dia<p>_PS3$ ^beliau<p>_PP3$ ^saya<p>_PS1$	fin+GEN.3S his [respect] I	316,685	2.68%
VSP	Passive Singular Verb	^di+dengar<v>_VSP$	di+hear 'be heard'	286,414	2.42%
CC	Cardinal Numeral	^satu<c>_CC-$ ^2004<c>_CC-$	one 2004	276,455	2.34%
H	Coordinating Conjunction	^dan<h>_H--$ ^maupun<h>_H--$	and although	270,418	2.29%
G	Negation	^bukan<g>_G--$ ^tak<g>_G--$ ^belum<g>_G--$	no(t) no(t) [colloquial] yet	106,242	0.90%

POS	Name of POS	Examples	English Gloss	Frequency	Percentage
PP	Plural Personal Pronoun	^mereka<p>_PP3$ ^kita<p>_PP1+lah<t>_T--$	they we [inclusive]	96,272	0.81%
M	Modal	^akan<m>_M--$	will	89,445	0.76%
O	Copula	^adalah<o>_O--$	be	48,439	0.41%
T	Particle	^terus<a>+kan_VSA+**lah<t>_T--$**	continue+kan+**lah** ‘continue-LAH’	35,076	0.30%
W	Question	^apa<w>_W--$	what	27,039	0.23%
CO	Ordinal Number	^ke+sembilan<c>_CO-+dia<p>_PS3$ ^ke+se+puluh<c>_CO-	ke+seven-GEN.3S ‘its ninth’ ke+se+ten ‘tenth’	27,209	0.23%
ASS	Superlative Singular Adjective	^ter+sempit<a>_ASS$ ^ter+mulia<a>_ASS$	ter+narrow ‘the narrowest’ ter+noble ‘the noblest’	24,847	0.21%
I	Interjection	^wah<i>_I--$ ^aduhai<i>_I--$	‘wah’ ‘Goodness’	7,662	0.06%
NSM	Masculine Singular Noun	^jelita<a>+wan_NSM$ ^harta<n>+wan_NSM$	pretty+wan ‘beautiful girl’ property+wan ‘rich man’	3,784	0.03%
VPA	Active Plural Verb	^ber+tempiar<v>+an_VPA$ ^ber+gilir<v>_VPA$	ber+scatter+an ‘scattered all around’ ber+take.turn ‘to take turn’	1,485	0.01%
CD	Collective Numeral	^ber+ribu<c>_CD-$ ^ber+puluh<c>_CD-$	ber+thousand ‘thousands’ ber+ten ‘tens’	1,146	0.01%
NSF	Feminine Singular Noun	^seni<n>+wati_NSF$ ^olahraga<n>+wati_NSF$	art+wati ‘actress’ sports+wati ‘female athlete’	86	0.00%
NPD	Non-Specified Singular Noun	^gerigi<n>_NPD$	saw	1	0.00%
			Total	**11,821,108**	**100.00%**

We also calculated all the affixes in the corpus. The distribution of suffixes is shown in Table 7 below. “None” means that a word had no suffix but may have had a prefix, and these constituted the majority of the total instances (89.6932%).[20] Words with the suffix *-an* ranked second, followed by *-kan*.

[20] Results up till four decimals were shown due to the low frequency of some of the percentages.

Table 7. All suffixes in the corpus

Suffix	Example	English Gloss	Frequency	Percentage
None	^ter+paksa<v>_VSP$ ^se+lepas<a>_ASP$	ter+force 'be forced' se+after 'after'	10,602,729	89.6932%
-an	^per+dagang<v>**+an**_NSD$ ^hias<v>**+an**_NSD$	per+trade+an 'trading/business' decorate+an 'decoration'	717,931	6.0733%
-kan	^meN+kait<v>**+kan**_VSA$ ^meN+bahas<v>**+kan**_VSA$	meN+relate+kan 'to relate' meN+debate+kan 'to debate'	356,930	3.0194%
-i	^meN+hadir<v>**+i**_VSA$ ^di+dengar<v>**+i** _VSP$	meN+attend+i 'to attend' di+hear+i 'be heard'	140,085	1.1850%
-wan	^usaha<n>**+wan**_NSM$ ^jelita<a>**+wan**_NSM$	work+wan 'entrepreneur' pretty+wan 'good looking man'	3,347[21]	0.0283%
-wati	^seni<n>**+wati**_NSF$ ^olahraga<n>**+wati**_NSF$	arts+wati 'actress' sports+wati 'female athlete'	86	0.0007%
		Total	**11,821,108**	**100.0000%**

Table 8 below shows the prefixes in the corpus. Words with no prefix (but may have had a suffix) also topped the list, followed by *meN+*, which constituted 4.84% of the total corpus, followed by *ber-* as the second most frequent and *di+* as the third most frequent.

Regarding all the morphological combinations, Table 9 below shows the most frequent (per million) pattern under each prefix. For example, among all the occurrences of *ber-*, the single use of *ber-* without a second prefix or any suffix was the most frequent form (21,569.81 per million). Comparatively, [*di+ +kan*] was the most frequent pattern of *di-* (8,524.41 per million). Similarly, we found that [*ke+ +an*], [*meN+*], [*peN+*], [*se+*], and [*ter+*] were the top combinations in their respective categories. Based on these results, we could predict the pattern of each morphological combination in the corpus. For instance, we could see that [*di+*] was slightly less frequent than [*di+ +kan*]. Sneddon et al. (2010: 261) commented that *-kan* with *di-* is sometimes obligatory for verbs such as *ajar* 'teach' and *beri* 'give', but obligatory for other verbs such as *tinggal* 'leave' and *maksud* 'mean'. Given this result, we analyzed the kinds of roots (lemmas) that appeared with each combination and further examined whether the postulation by Sneddon et al. (2010) was also proven in the corpus.

[21] A total of 437 instances of *tai-wan* and one instance of incorrectly parsed *ter-uk-wan* [*terlampau teruk wan mahyuddin berkata...*] were removed manually from the suffix *-wan*. They were added back into 'None' after manual examination.

Table 8. All prefixes in the corpus

Prefix	Example	English Gloss	Frequency	Percentage
None	^jalan<n>_NSD$	road	9,995,832	84.56
	^masyarakat<n>_NSD$	society		
	^hias<v>+an_NSD$	decorate+an ‘decoration’		
meN+	^meN+cetus<v>+kan_VSA$	meN+outburst+kan ‘to cause an outburst’	572,184	4.84
	^meN+hadir<v>+i_VSA$	meN+attend+i ‘to attend’		
ber+	^ber+ikut<v>+an_VSA$	ber+follow+an ‘to follow’	303,731	2.57
	^ber+banding<v>_VSA$	ber+compare ‘to compare’		
	^ber+tukar<v>_VSA$	ber+change ‘to change’		
di+	^di+dapat<v>+i_VSP$	di+get+i ‘be found’	242,070	2.05
	^di+gantung<v>_VSP$	di+hang ‘be hung’		
peN+	^peN+lancar<a>+an_NSD$	peN+smooth+an ‘to launch’	225,869	1.91
	^peN+sokong<v>_NSD$	peN+support ‘supporter’		
ke+	^ke+bakar<v>+an_NSD$	ke+burn+an ‘a fire’	207,014	1.75
	^ke+lahir<v>+an_NSD$	ke+birth+an ‘birth’		
	^ke+dua<c>_CO-$	ke+two ‘second’		
per+	^per+himpun<v>+an_NSD$	per+gather+an ‘gathering’	103,396	0.87
	^per+hubung<v>+an_NSD$	per+contact+an ‘connection’		
	^per+tanya<v>+an_NSD$	per+ask+an ‘question’		
se+	^se+lain <a>_ASP$	se+other ‘other than’	97,913	0.83
	^se+lepas<a>_ASP$	se+after ‘after’		
	^se+baik <a>_ASP$	se+good ‘as soon/good as’		
ter+	^ter+kejut<v>_VSP$	ter+shock ‘shocked’	73,090	0.62
	^ter+besar<a>_ASS$	ter+big ‘biggest’		
pe+	^pe+cinta<n>_NSD$	pe+love ‘lover’	9	0.00
		Total	**11,821,108**	**100.00%**

Table 9. All morphological combinations in the corpus

Prefix1+	Prefix2+	+Suffix1	+Suffix2	Tokens	Per Million
ber+	--	--	--	254,979	21569.81
ber+	--	--	*+an*	36,691	3103.85
ber+	--	--	*+kan*	7,107	601.21
ber+	*ke+*	--	*+an*	2,694	227.90
ber+	*peN+*	--	*+an*	1,000	84.59
ber+	*peN+*	--	--	935	79.10
ber+	*se+*	--	--	116	9.81
ber+	*per+*	--	*+an*	70	5.92
ber+	*se+*	--	*+an*	65	5.50
ber+	--	*+an*	--	33	2.79
ber+	--	--	*+i*	29	2.45
ber+	--	*+kan*	--	8	0.68
ber+	*ke+*	*+an*	--	2	0.17
ber+	*peN+*	--	*+kan*	2	0.17
di+	--	--	*+kan*	100,768	8524.41
di+	--	--	--	99,263	8397.10
di+	--	--	*+i*	33,869	2865.13
di+	*per+*	--	*+kan*	4,337	366.89
di+	*per+*	--	--	1,559	131.88
di+	--	*+i*	--	847	71.65
di+	--	*+kan*	--	768	64.97
di+	*per+*	--	*+i*	628	53.13
di+	*per+*	*+kan*	--	27	2.28
di+	--	--	*+an*	2	0.17
di+	*per+*	*+i*	--	2	0.17
ke+	--	--	*+an*	182,539	15441.78
ke+	--	--	--	18,273	1545.79
ke+	--	*+an*	--	6,051	511.88
ke+	*se+*	--	--	151	12.77
meN+	--	--	--	271,972	23007.32
meN+	--	--	*+kan*	211,260	17871.42
meN+	--	--	*+i*	74,813	6328.76
meN+	*per+*	--	*+kan*	7,365	623.04
meN+	--	*+kan*	--	3,858	326.37
meN+	--	*+i*	--	1,376	116.40
meN+	*per+*	--	*+i*	807	68.27
meN+	*per+*	--	--	481	40.69
meN+	*ber+*	--	*+kan*	124	10.49
meN+	*per+*	*+kan*	--	100	8.46
meN+	*per+*	*+i*	--	28	2.37

Prefix1+	Prefix2+	+Suffix1	+Suffix2	Tokens	Per Million
peN+	--	--	--	118,401	10016.07
peN+	--	--	*+an*	105,041	8885.88
per+	--	--	*+an*	96,821	8190.52
per+	--	--	--	4,997	422.72
peN+	--	*+an*	--	2,427	205.31
per+	--	*+an*	--	1,578	133.49
pe+	--	--	--	9	0.76
se+	--	--	--	97,907	8282.39
se+	*per+*	--	--	6	0.51
ter+	--	--	--	72,009	6091.56
ter+	*peN+*	--	--	952	80.53
ter+	--	--	*+kan*	70	5.92
ter+	--	--	*+i*	34	2.88
ter+	*ke+*	--	*+an*	13	1.10
ter+	--	--	*+an*	10	0.85
ter+	*peN+*	--	*+an*	1	0.08
--	--	--	*+an*	292,984	24784.82
--	--	--	*+i*	29,905	2529.80
--	--	--	*+kan*	25,897	2190.74
--	--	*+an*	--	6,004	507.91
--	--	--	*+wan*	3,347	283.14
--	--	*+i*	--	650	54.99
--	--	*+kan*	--	554	46.87
--	--	--	*+wati*	86	7.28
--	--	*+wan*	--	14	1.18
--	--	--	--	9,636,392	815185.18
			Total	**11,821,108**	

In addition to the above morphological combinations, we also found lists of the roots without affixes, lists of foreign words (“F”s), and lists of unknown words (“X”s) in the corpus. Table 10 below shows the top ten combinations.

The lists in Table 10 show the following: (a) Like many other languages, function words are the most frequent words in Malay; (b) Many of the foreign words were proper nouns commonly found in Malaysia. These words were not collected as part of the vocabulary in the MorphInd dictionary. This also shows that MorphInd could be improved in terms of its proper noun lexicon, and a country name such as “Malaysia” and a title such as “Dr.” could be treated better than “foreign words”; and (c) Most of the “X”s were words not found in Indonesian. From Table 6 previously, we knew that these words constituted about 12.28% of the total corpus, yet in the top ten “X”s, we still found two—*sahaja* ‘only’ and *pelbagai* ‘various kinds’—that should exist in Indonesian (though the former has a different use) but were still tagged as “X”s by MorphInd. This is one of the limitations of MorphInd, which will be discussed further in the conclusion.

In what follows, we will present a demonstration of prefix analysis using *ber-* as an example.

Table 10. Top 10 roots with no affixes, "F"s, and "X"s in the corpus

Top 10 Roots	Frequency	Top 10 "F"s	Frequency	Top 10 "X"s	Term in Indonesian	Frequency
yang 'REL'	268,626	*malaysia* 'Malaysia'	33,174	*kerana* 'because'	*karena*	35,484
dan 'and'	230,602	*datuk* [title]	23,635	*mohd* [proper Name]	*moh*	13,458
di 'at'	189,582	*kuala* 'estuary' [part of a proper name]	13,471	*polis* 'police'	*polisi*	13,055
dia '3s'	167,544	*abdul* [name of a person]	10,034	*majlis* 'council'	*majelis*	12,933
itu 'that'	132,437	*dr* 'Dr.'	7,924	*syarikat* 'company'	*perusahaan*	12,384
ini 'this'	122,646	*com* [url]	7,710	*mahu* 'want'	*mau*	11,799
dalam 'in(side)'	92,401	*al* [Arabic term]	6,826	*iaitu* 'that is'	*yaitu*	11,443
untuk 'for'	88,630	*ahmad* [proper name]	6,594	*sahaja* 'only'[22]	*saja*	11,195
dengan 'with'	86,327	*mac* [March]	6,240	*pelbagai* 'various kinds'	*pelbagai*	10,658
tidak 'no(not)'	78,407	*anwar* [proper name]	6,137	*bahawa* 'that'	*bahwa*	10,490

5. A demonstration of the prefix analysis of *ber-*

As mentioned, the single use of *ber-* without a second prefix or any suffix was the most frequently seen form of *ber-*. In this section, we will show a search for *ber-* using our annotated corpus. Given that there were 303,731 instances of *ber-*, it was necessary to select a smaller number of instances for detailed analysis. For this purpose, we looked at different types of *ber-* (instead of tokens). Using a smaller portion of the data, consisting of one sixth of the total newspaper articles, we analyzed 5,740 randomly selected texts and found 662 types of *ber-* words in total, amounting to 12,668 tokens. The types and tokens were collected using the AntConc 3.5.7 Concordancer (Anthony, 2005) (see Figure 9), from which we generated a wordlist for *ber-*. This final list excluded examples such as *ber-nama*, which is the name of the newswire BERNAMA, and therefore these instances were not counted. A tagged corpus enables precise search results with morphological information. By using a tagged corpus, we avoided noise such as *beri* 'give', *berita* 'news', etc.

Based on the instances above, we analyzed the lemma tags of *ber-*, as shown in Table 11 below. From the results, it can be seen that *ber-* combined with words that were mainly nouns (e.g., *ber-kesan*), followed by verbs (e.g., *ber-main*) and adjectives (e.g., *ber-*

[22] *Sahaja* and *pelbagai* were both found in the *Kamus Besar Bahasa Indonesia* (*KBBI*), but they were not tagged using MorphInd. *Sahaja* has a slightly different use in Indonesian. (https://kbbi.kemdikbud.go.id/entri/sahaja; https://kbbi.kemdikbud.go.id/entri/pelbagai; accessed September 20, 2019). We thank David Moeljadi for providing some of the terms in Indonesian.

sabar). “Types” refer to the types of words that formed combinations with *ber*-, while “tokens” refer to the number of instances found for the particular lemma tag.

Table 11. Tags for the lemmas of *ber*-

Lemma Tags	Types	%	Tokens	%
Noun	369	55.74	6,354	50.16
Verb	178	26.89	3,588	28.32
Adjective	102	15.41	2,533	20.00
Numeral	7	1.06	150	1.18
Adverb	2	0.30	33	0.26
Subordinating/Coordinating	2	0.30	8	0.06
Determiner	1	0.15	1	0.01
Foreign Word	1	0.15	1	0.01
Total	**662**	**100.00**	**12,668**	**100.00**

Table 11 above shows that a tagged corpus enabled the calculation of morphology information for the formation of *ber*- words. This kind of information cannot be obtained with a corpus of running texts. Table 12 below shows the kinds of tags provided for whole *ber*- words. These were the tags provided for each occurrence of *ber*- in the sentence, and the tags were given based on the uses of *ber*- in different sentences.

Table 12. Tags for *ber*- words

Word Tags	Gloss	Types	%	Tokens	%
VSA	Verb-Singular-Active	647	97.73	12 523	98.86
VPA	Verb-Plural-Active	5	0.76	73	0.58
CD	Numeral-Collective	7	1.06	59	0.47
D	Adverb	1	0.15	10	0.08
ASP	Adjective-Singular-Positive	1	0.15	2	0.02
B	Determiner	1	0.15	1	0.01
Total		**662**	**100.00**	**12,668**	**100.00**

Table 12 shows that a majority of *ber*- words were active singular verbs (*ber-malam*, *ber-tutur*), while a small number were active plural verbs (*ber-kempen*, *ber-kembar*) and numeral collective words (*ber-dua*, *be(r)-ratus*, etc.). There was only one adverb (*ber-ikut-nya*), singular positive adjective (*ber-sahaja*), and determiner (*ber-bagai*). These tags were produced directly from the tagger, but were later checked manually. Minimal corrections were carried out for these tags.

With these results, we moved on to more sophisticated analysis, such as semantic categorization. *Ber*- has several different uses, among which are “mental events” (*Ali sedang ber-fikir* ‘Ali PROG ber-think’ ‘Ali is thinking’) and “reciprocal/collective action” (*Mereka masih ber-runding* ‘3PL still ber-negotiate’ ‘They are still negotiating’) (Wee, 1995: 93). In order to move on to a more detailed description of *ber*-, further semantic annotation was necessary. This was carried out by adding semantic information to the tags provided by the MorphInd POS tagger. For instance, if we found an instance of *ber-runding* in our data, the following annotation in (5) was carried out:

(5) ^ber+runding<v>_VSA$ ==>^ber+runding<v>_VSA**rec**$

In (5), a semantic tag was added to the end of the POS tag so that additional information could be added without affecting the original structure of the data. In this way, a prefix could be analyzed not only in terms of its roots and POS but also its semantic information.

6. Limitations and conclusion

This study highlighted the need for a tagged corpus of Malaysian Malay. To achieve this, 35,767 texts from a Malaysian news corpus were tagged using two approaches. A final version using the second approach, the MorphInd POS tagger, was eventually adopted and resulted in a corpus with morphology (stemmed morphemes and affixes) and POS tag information. POS tags were produced for the whole word and for the lemmas. With this information, the calculation of morphological combinations and POS was carried out.

Despite all the advantages we found using MorphInd, there were still some limitations. The problems we found with MorphInd are as follows. First, as mentioned, its dictionary for proper names needs improvement. Names such as *Sulaiman* were given an NSD tag ^sulaiman<n>_NSD$, but names such as Ahmad received a foreign word tag (cf. Table 10). Capital city names such as Kuala Lumpur were treated separately as ^kuala<f>_F--$ ^lumpur<n>_NSD$, with the former as a foreign word but the latter as an NSD. Second, some words, for some unknown reason, were not successfully tagged, such as ^difahamkan<x>_X--$ and ^terpancut<x>_X--$. A check in the *Kamus Besar Bahasa Indonesia* (*KBBI*) dictionary showed that *faham* 'understand' is a non-standard use in Indonesian, while the standard use is *paham*.[23] *Pancut* 'to eject' is a word collected in the *KBBI*, but was not tagged by MorphInd. Therefore, some unsuccessfully tagged words were due to vocabulary differences, while the reason for others was unknown. Third, some errors were detected, such as *ter-selaras* 'ter-same.level' 'coordinated'. Although it is well understood that MorphInd treats some lexicalized words as one word (thus, *ter-selaras* instead of the double prefixes *ter-se-laras*), the assigned tag was incorrect in ^ter+selaras<a>_ASS$. ASS refers to a superlative singular adjective such as *ter-tinggi* 'ter-tall' 'highest', but *ter-selaras* is not an ASS. Similarly, we also found *ter-selamat* 'ter-save' 'be saved' as ^ter+selamat<a>_ASS and ^ter+putus<a>_ASS$, both of which were errors. Following the above, the fourth limitation was the decisions made a priori regarding lexicalized words such as ^terkini<a>_ASP$ 'most recent', ^adalah<o>_O--$ 'be', ^secara<r>_R--$ 'in way of', and ^sebagai<r>_R--$ 'as', but not for words such as ^se+lepas<a>_ASP$ 'after' and ^se+paruh<n>_ASP$ 'half'. To further illustrate this phenomenon, in the lexicalized word ^kasihan<i>_I--+dia<p>_PS3$, *kasihan* 'pitiful' was treated as one word instead of *kasih-an*, its derived form. Other examples included ^meN+kemuka<v>+kan_VSA$, whereby the root was *muka* instead of *kemuka*, and ^sebenarnya<d>_D--$, which could be further lemmatized into *se-benar-nya*.

In other words, the dictionary used in MorphInd needs to be expanded and checked for consistency.[24] Given the above, a great deal of improvements are still needed to attest for the precision and practicality of the tagger. Despite all the aspects that need improvement, we produced an end-product—the annotated Malaysian Malay corpus—which will allow impossible works in the past to be carried out more easily.

[23] https://kbbi.kemdikbud.go.id/entri/paham (accessed September 23).

[24] The tagger came with a default mapping dictionary file that aimed to produce correct tag results for certain words when the tags were assigned. The word list in the mapping dictionary should be further discussed and expanded to produce more consistent tagging results when these tags are assigned.

Abbreviations

A	adjective	NSF	feminine singular noun
ASP	positive singular adjective	NSM	masculine singular noun
ASS	superlative singular adjective	O	copula
B	determiner	P	personal pronoun
C	numeral	PP	plural personal pronoun
CC	cardinal numeral	PS	singular personal pronoun
CD	collective numeral	R	preposition
CO	ordinal number	S	subordinating conjunction
D	adverb	T	particle
F	foreign word	V	verb
G	negation	VPA	active plural verb
H	coordinating conjunction	VSA	active singular verb
I	interjection	VSP	passive singular verb
M	modal	W	question
N	noun	X	unknown

References

Anthony, Laurence. 2005. AntConc: Design and development of a freeware corpus analysis toolkit for the technical writing classroom. In *Proceedings of the International Professional Communication Conference*, 729−737.

Chu, Benjamin Min Xian, Mohamed Lubani, Kwei Ping Liew, Khalil Bouzekri, Rohana Mahmud & Dickson Lukose. 2016. Benchmarking Mi-POS: Malay Part-of-Speech Tagger. *International Journal of Knowledge Engineering* 2(3). 115−121.

Chung, Siaw-Fong. 2010. Numeral classifier *buah* in Malay: A corpus-based study. *Language and Linguistics* 11(3). 553−577.

Chung, Siaw-Fong. 2011. Uses of *ter-* in Malay: A corpus-based study. *Journal of Pragmatics* 43(3). 799−813.

Chung, Siaw-Fong. 2013. Investigating *di*, *dalam*, and *di dalam* 'in(side)' in Standard Malaysian Malay using corpus data. *Language and Linguistics* 14(6).1009−1034.

Chung, Siaw-Fong. 2014. A corpus-based investigation of Malay antonymous Prepositions *bawah* : *atas* and *dalam* : *luar*. In Siaw-Fong Chung & Hiroki Nomoto (eds.), *Current trends in Malay linguistics*, 51−82. *NUSA* 57. Jakarta and Tokyo: Universitas Katolik Indonesia Atma Jaya and Tokyo University of Foreign Studies.

Chung, Siaw-Fong. 2019. *Lagi* in Standard Malaysian Malay: Its meaning conceptualization. *Concentric: Studies in Linguistics* 45(1). 82−111.

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 8th International Language Resources and Evaluation (LREC'12)*, 759−765.

Imran Ho Abdullah, Zaharani Ahmad, Rusdi Abdul Ghani & Nor Hashimah Jalaluddin. 2004. A practical grammar of Malay—A corpus based approach to the description of Malay: Extending the possibilities for endless and lifelong language learning. In *Proceedings of the First COLLA Regional Workshop*, 90−99.

Kilgarriff, Adam, Siva Reddy, Jan Pomikálek & Avinesh PVS. 2010, May. A corpus factory for many languages. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*, 904–910.

Kilgarriff, Adam & David Tugwell. 2002. Sketching words. In Marie-Hélène Corréard (ed.), *Lexicography and natural language processing: A Festschrift in honour of B. T. S. Atkins*. Euralex, 125–137.

Knowles, Gerald O. & Zuraidah Mohd Don. 2003. Tagging a corpus of Malay texts, and coping with "syntactic drift". In *Proceedings of the Corpus Linguistics 2003 Conference*. University of Lancaster, Centre for Computer Corpus Research on Language, 422−428.

Knowles, Gerald O. & Zuraidah Mohd Don. 2006. *Word class in Malay: A corpus-based approach*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Larasati, Septina Dian, Vladislav Kuboň & Daniel Zeman. 2011. Indonesian morphology tool (MorphInd): Towards an Indonesian corpus. In Mahlow Cerstin & Michael Piotrowski (eds.), *Systems and frameworks for computational morphology (Proceedings of the Second International Workshop, SFCM 2011),* 119−129

Lee, Lay Wah & Hui Min Low. 2011. Developing an online Malay language word corpus for primary schools. *International Journal of Education & Development Using Information & Communication Technology* 7(3). 96–101.

Mohd Hanafi Ahmad Hijazi, Lyndia Libin, Alfred Rayner & Frans Coenen. 2016. Bias aware lexicon-based sentiment analysis of Malay dialect on social media data: A study on the Sabah language. In *Proceedings of the 2016 2nd International Conference on Science in Information Technology (ICSITech),* Balikpapan, 356−361. doi: 10.1109/ICSITech.2016.7852662

Nomoto, Hiroki, Hannah Choi, David Moeljadi & Francis Bond. 2018. 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, Shiro Akasegawa & Asako Shiohara. 2018a. Building an open online concordancer for Malay/Indonesian. Paper presented at the 22nd International Symposium on Malay/Indonesian Linguistics (ISMIL). University of California, Los Angeles, USA.

Nomoto, Hiroki, Shiro Akasegawa & Asako Shiohara. 2018b. Reclassification of the Leipzig Corpora Collection for Malay and Indonesian. *NUSA* 65. 47−66.

Norshuhani Zamin, Alan Oxley, Zainab Abu Bakar & Syed Ahmad Farhan. 2012. A lazy man's way to part-of-speech tagging. In Deborah Richards & Byeong Ho Kang (eds.), *Knowledge management and acquisition for intelligent systems (Proceedings of the 12th Pacific Rim Knowledge Acquisition Workshop, PKAW 2012),* 7457, 106–117. Heidelberg: Springer.

Rayner, Alfred, Adam Mujat & Joe Henry Obit. 2013. A ruled-based part of speech (RPOS) tagger for Malay text articles. In Ali Selamat, Ngoc Thanh Nguyen, & Habibollah Haron (eds.), *Intelligent information and database systems (Proceedings of the 5th Asian Conference),* Part II, 50−59. Heidelberg: Springer.

Scannell, Kevin P. 2007. The Crúbadán project: Corpus building for under-resourced languages. In *Building and exploring web corpora: Proceedings of the 3rd Web as Corpus Workshop*, 4, 5–15. Louvain-la-Neuve, Belgium.

Sneddon, James N., Alexander Adelaar, Dwi N. Djenar & Michael C. Ewing. 2010. *Indonesian: A comprehensive grammar*, 2nd edn. London and New York: Routledge.

Tan, Tien-Ping, Bali Ranaivo-Malançon, Laurent Besacier, Yin-Lai Yeong, Keng Hoon Gan & Enya Kong Tang. 2017. Evaluating LSTM networks, HMM and WFST in Malay part-of-speech tagging. *Technology Transforming Lives I* 9(2-9). 79–83.

Yap, Melvin J., Susan J. Rickard Liow, Sajlia Binte Jalil & Siti Syuhada Binte Faizal. 2010. The Malay lexicon project: A database of lexical statistics for 9,592 words. *Behavior Research Methods* 42(4). 992–1003.

Wee, Hock Ann Lionel. 1995. *Cognition in grammar: The problem of verbal prefixation in Malay*. University of California, Berkeley dissertation.

Zuraidah Mohd Don. 2010. Processing natural Malay texts: A data-driven approach. *Trames: Journal of the Humanities and Social Sciences* 14(1). 90–103.