Reclassification of the Leipzig Corpora Collection for Malay and Indonesian

Hiroki NOMOTO[⋆], Shiro AKASEGAWA[◇] and Asako SHIOHARA[⋆]

[⋆]Tokyo University of Foreign Studies and [◇]Lago Institute of Language
nomoto@tufs.ac.jp, lagoinst@gmail.com, asako@aa.tufs.ac.jp,

The Leipzig Corpora Collection (LCC), a collection of monolingual web corpora for 236 languages in the world, offers the largest openly available corpora of Malay and Indonesian. However, despite their unprecedented sizes, it is difficult to use them for accurate linguistic research and language resource development in their original forms, because data from one language are mixed with a non-negligible amount of data from the other language due to the striking similarities between the two languages. We thus have developed a simple decision tree-based language identifier specifically to distinguish between Malay and Indonesian, and reclassified the relevant LCC data. Of the Malay data, 4.1% has been reclassified, and 1.5% of the Indonesian data has also been reclassified. Our Malay-Indonesian language identifier was evaluated against 300 documents of a length similar to that of the LCC data.

1. Introduction[1]

With the advent of web corpora, for which data can be automatically collected from the web, it has become possible to build a large-size written language corpus in a short period of time even in languages with limited research resources. However, the quality of the resulting corpus depends very much on the way in which the data are collected. This is especially so when a language co-exists with other similar languages. A crosslinguistically applicable general language identification method will likely create a corpus of low quality in such languages, where a corpus of one language contains a substantial amount of data from another language, unless some measures to deal with similar languages are taken.

The Malay and Indonesian subcorpora of the Leipzig Corpora Collection (LCC; Goldhahn, Eckart & Quasthoff 2012) demonstrate this problem. Hence, despite their impressive scope, they are arguably unsuitable for linguistic research on the two languages. A corpus of Malay may lead to a flawed linguistic generalisation of Malay grammar because of the Indonesian data contained in it. It is also likely that one may not be able to come up with an otherwise possible generalisation due to the non-Malay data.

A similar problem is expected when the Malay and Indonesian subcorpora of LCC are used for language resource and tool development. In fact, LCC is the data source for Malay and Indonesian in the DSL corpus collection (Tan et al. 2014), which has been used in the Discriminating between similar languages (DSL) shared task at the past VarDial

[1] The research reported in this paper was conducted under the JSPS "Program for Advancing Strategic International Networks to Accelerate the Circulation of Talented Researchers" grant offered to Tokyo University of Foreign Studies entitled "A Collaborative Network for Usage-Based Research on Less-Studied Languages." We are grateful to JSPS and also to the University of Melbourne and Nanyang Technological University, Singapore, for supporting the first author's stay at the respective institutions. This study has been presented at the 21st International Symposium on Malay/Indonesian Linguistics (ISMIL), the 154th Meeting of the Linguistic Society of Japan and the 4th Atma Jaya Conference on Corpus Studies (ConCorps). We thank the two anonymous reviewers, the participants at these conferences and also Francis Bond and his computational linguistics lab members for their comments and criticisms. We would also like to thank Bali Ranaivo-Malancon for kindly sharing the test sentences described in section 4.2.

Hiroki NOMOTO, Shiro AKASEGAWA and Asako SHIOHARA, 2018. 'Reclassification of the Leipzig Corpora Collection for Malay and Indonesian'. NUSA 65: 47–66. [Permanent URL: http://repository.tufs.ac.jp/handle/10108/92899] [doi: https://doi.org/10.15026/92899]

(Workshop on NLP for Similar Languages, Varieties and Dialects) workshops. In such a situation, accomplishing high accuracy for distinguishing between Malay and Indonesian in the shared task means no more than accurately replicating the classification of LCC. Although this is still beneficial for investigating general models of language identification, it is desirable that any tool produced is developed with and for linguistic data which are accurately classified.

In order to improve LCC to become a more reliable linguistic resource, we have reclassified its Malay and Indonesian subcorpora by redoing language identification. The present paper reports the method and results of our reclassification project.

The paper is organised as follows. Section 2 explains the close relationship between the two languages discussed in this paper, i.e. Malay and Indonesian. Section 3 overviews LCC, particularly its Malay and Indonesian subcorpora, which we have reclassified. The method and results of our reclassification of LCC are presented in section 4. The Malay-Indonesian language identifier used in our reclassification is evaluated in section 5. Section 6 is the conclusion, where we also discuss some implications of the present study for linguistic research and language resource development involving similar languages.

2. Malay and Indonesian

As stated in the previous section, Malay and Indonesian are similar to each other. This section gives a brief overview of the relationship between the two languages.

The language name "Malay" has two uses. In its narrow sense, it refers to the language designated as the national language of Malaysia, Singapore and Brunei Darussalam. In this paper, the name "Malay" is used in this narrow sense. In addition, "Malay" is also used in a broader sense. "Malay" in the broad sense is a macrolanguage, characterised by common linguistic features, regardless of geopolitical borders. We refer to "Malay" in this sense as the "Malay macrolanguage" below. The Malay macrolanguage has spread widely throughout the above-mentioned three countries and their surrounding regions, particularly Indonesia. With roughly 300 million speakers across a wide geographic area, it has a number of regional and sociolinguistic varieties differing in phonology, vocabulary and grammar. Indonesian is one of these regional varieties of the Malay macrolanguage that is designated as the national language of the Republic of Indonesia. That is to say, the two languages discussed in this study are no more than two different regional varieties of the same language, namely the Malay macrolanguage.

When it comes to the formal registers for the standard varieties of Malay and Indonesian, i.e. Standard Formal Malay and Indonesian, the two languages show a striking similarity.[2] Standard Indonesian is closer to the standard varieties of Malaysia, Singapore and Brunei than to varieties of the Malay macrolanguage within Indonesia such as Ambon Malay and Kupang Malay. The differences between the two languages are mostly limited to phonology and vocabulary. The lexical difference is estimated to be about 10% (Asmah 2001). A few syntactic differences also exist (see, e.g., Nomoto & Kartini 2011). It is against

[2] By "standard variety" we do not mean a prescriptive normative language (i.e. standardised language) but a language variety that has emerged naturally as a common means of communication across the country that has few to no region-specific features or at least is deemed so by the speakers. The formation of a standard variety is influenced by the standardised language, but the two are distinct. It must be noted that some authors use the term Standard Indonesian to refer to the standardised language.

such a background that Malay and Indonesian are regarded as similar languages.

Although linguistically speaking, Malay and Indonesian are in a dialectal relationship, they are the national languages of distinct countries with their own identities, norms and standards. Malaysia, Singapore and Brunei each have their standard languages, but the differences between the three are trivial. Moreover, Singapore and Brunei look to Malaysia as a model for their standard languages (Asmah 1992). It is thus reasonable to treat the three standard languages as one. By contrast, greater differences exist between the standard languages of Malaysia, Singapore and Brunei on one hand and Indonesia on the other. This is because the region currently known as Indonesia was once a Dutch colony and national language development efforts took place separately from the other three countries, which were occupied by the British. The speakers are very sensitive to the differences between the two, especially those concerning phonology and vocabulary, and regard the two as different "languages." Therefore, Malay and Indonesian must be treated separately when developing linguistic resources and tools.

3. Leipzig Corpora Collection

LCC is a collection of web corpora, whose data are collected by automatic web crawling. LCC was developed by the NLP Group, Department of Computer Science, University of Leipzig (`http://corpora.uni-leipzig.de/`, accessed on 1 August 2017). It consists of monolingual corpora from up to 236 world's languages, including Malay and Indonesian. The corpus data can be downloaded in different sizes, up to 3 million sentences for free and without prior registration.

LCC's corpora of Malay and Indonesian are made up of 16 subcorpora, three for Malay and thirteen for Indonesian:

- Subcorpora of Malay: `msa_newscrawl_2011`, `msa_wikipedia_2016`, `ind-bn_web_2015`[3]
- Subcorpora of Indonesian: `ind_mixed_2012`, `ind_news_2008`, `ind_news_2009`, `ind_news_2010`, `ind_news_2011`, `ind_news_2012`, `ind_newscrawl_2011`, `ind_newscrawl_2012`, `ind_web_2011`, `ind_web_2012`, `ind_wikipedia_2016`, `ind-id_web_2013`, `ind-id_web_2015`

According to Goldhahn, Eckart & Quasthoff (2012), newspapers were crawled based on the information in *ABYZ News Links* (`http://www.abyznewslinks.com/`, accessed on 31 August 2018), which offers a list of online newspapers and the languages in which they are written for most countries in the world. Generic web pages were crawled using seeds tuples generated using language data from the translations of the Universal Declaration of Human Rights and *The Watchtower* (`https://www.jw.org/en/publications/magazines/`, accessed on 31 August 2018).

Table 1 shows the total sizes of Malay and Indonesian corpora in sentence and token counts.[4] The sizes of major existing corpora of Malay and Indonesian that are openly

[3] LCC miscategorises Brunei Malay as a dialect of Indonesian.

[4] The token counts here and elsewhere in the paper were calculated using WordPunctTokenizer of the Natural Language Toolkit (NLTK; Bird, Loper & Klein 2009). WordPunctTokenizer tokenizes punctuation marks and hyphens. Thus, a single word involving reduplication such as *kanak-kanak* 'child' is counted as

Table 1. The sizes of Malay and Indonesian corpora in LCC

Language	Sentence	Token
Malay	957,560	19,902,826
Indonesian	75,900,523	1,477,803,691
Total	76,858,083	1,497,706,517

Table 2. The sizes of existing corpora of Malay and Indonesian

Corpus	Variety	Size (words)
Malay Concordance Project	classical; all regions	5.8 million
Korpus DBP	modern; Malaysia	135 million
SEALang Malay	modern; Malaysia	2.5 million
SEALang Indonesian	modern; Indonesia	5 million

available online are given in Table 2 for the purpose of comparison. These corpora can be accessed from the following URLs (accessed on 4 August 2017):

- Malay Concordance Project: `http://mcp.anu.edu.au/`
- Korpus DBP: `http://sbmb.dbp.gov.my/korpusdbp/SelectUserCat.aspx`
- SEALang Corpus Malay: `http://sealang.net/malay/corpus.htm`
- SEALang Corpus Indonesian: `http://sealang.net/indonesia/corpus.htm`

Although the methods for calculating word counts may differ from corpus to corpus (see footnote 4), it is obvious that the existing corpora are far smaller than LCC, except for Korpus DBP. Korpus DBP's large size is due to the fact that it was built by the national language institute of Malaysia, Dewan Bahasa dan Pustaka (DBP). DBP has been publishing numerous books and magazines in Malay and was able to use these for their corpus data. Despite its impressive size, however, what one can do with Korpus DBP is extremely limited by its search interface, which does not allow access to multiple sentence strings, let alone the entire raw texts. Given the current situation, LCC can be said to be the best corpora of Malay and Indonesian in terms of its size.

However, as pointed out at the outset of this paper, LCC suffers from data quality problems. The most serious is the mixture of two languages: Indonesian data are found in the corpora of Malay and vice versa.[5] This problem arises because of errors in language identification during data collection. In fact, a higher error rate is anticipated for Malay and Indonesian than for other languages in LCC, given the closeness of the two languages. Distinguishing Indonesian from Malay is far more difficult than distinguishing English from Malay, not just for computers but also for humans, even more so when one tackles it with a single common strategy, as LCC presumably does.

The current study has corrected language identification errors in the Malay and Indonesian corpora of LCC by subjecting the entire data to a different language identification system

three tokens: *kanak*, -, *kanak*. Therefore, the token counts provided in this paper are bound to be larger than the corresponding word counts.

[5] Other problems include abundant spelling errors and lack of balance with regard to genres.

which we have created. The next section discusses our language identification system and the results of the reclassification of LCC by means of it.

4. Reclassification

4.1 Language categories

For our reclassification of LCC, we revised the original LCC language classification categories, as shown in Figure 1. The new categories reflect the linguistic facts discussed in section 2. The original classification has no category for the Malay macrolanguage, which encompasses all varieties of the Malay macrolanguage in the Malay Archipelago. Moreover, it miscategorises Standard Malay in Brunei as a dialect of Indonesian. In the new classification, the code **msa** is assigned to the Malay macrolanguage and accommodates data that cannot be identified either as Malay or Indonesian. A different code is used for Malay in the narrow sense, i.e. **zsm**. The code **ind** is reserved solely for Indonesian. These language categories are ISO693-3 languages codes, which are also used, for example, in *Ethnologue* (Simons & Fennig 2018).

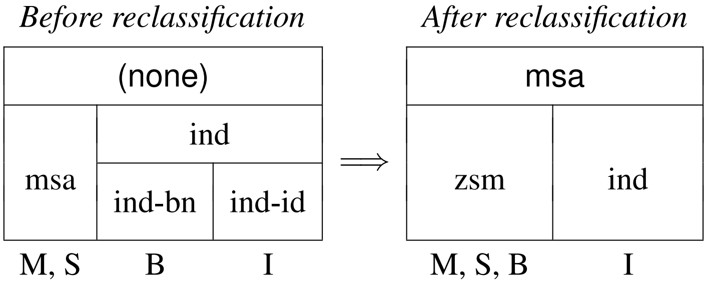

Figure 1. Language classification in LCC (M: Malaysia; S: Singapore; B: Brunei; I: Indonesia)

As pointed out by one of the reviewers, under the old rubric, it was possible to distinguish Brunei Malay, Malaysian/Singapore Malay and Indonesian, but this is impossible under the new rubric. This is admittedly a major loss of information. At present, it is difficult to identify whether a **zsm** page belongs to Malaysia/Singapore or Brunei for the following reasons. First, the only Brunei Malay subcorpus `ind-bn_web_2015` consists only of documents in the formal register. It thus contains no instance of *ani* 'this' and *atu* 'that' (cf. *ini* and *itu* in Malaysia and Singapore), which are quite common in the informal register. When it comes to the formal register, the standard languages of Malaysia, Singapore and Brunei are extremely similar (see section 2). There is a word that characterises Standard Formal Brunei Malay, namely *awda* 'you' (cf. *anda* in Malaysia and Singapore). But even this word occurs only 29 times in the entire subcorpus.

In fact, we believe that our language identification method discussed below, if adapted for distinguishing Brunei Malay, will identify Brunei Malay pages. This is because all documents in the Brunei subcorpus come from URLs with the Brunei country domain `.bn` and our language identifier utilises country domains as a last resort, which should be avoided if possible. However, it is feared that one would always have to appeal to this last resort. Hence, we will lump Brunei and Malaysia/Singapore together until a more sophisticated language identifier is developed.

4.2 Method

The unit of our final language identification is a web page or URL, assuming that every page has one main language and the amount of potential code-mixing is negligible. What

Table 3. The accuracy of Malay-Indonesian language identifiers at the sentence level

	Ranaivo-Malançon (2006)			This study		
	Correct	Error	Neutral	Correct	Error	Neutral
Malay	12	0	78	20	0	70
(%)	(13.3)	(0.0)	(85.6)	(22.2)	(0.0)	(76.7)
Indonesian	24	3	63	41	0	49
(%)	(26.7)	(3.3)	(68.9)	(44.4)	(0.0)	(54.4)

this means in practice is that all sentences from the same web page are assigned the same language code.

We do not adopt sentence as the unit, because it is known that sentence-level language identification is a challenging task in Malay and Indonesian. Ranaivo-Malançon (2006) tested her Malay-Indonesian Language Identifier against 90 sentences consisting of two to 10 words (10 sentences for each length). She employed a mechanism similar to our Malay-Indonesian Language Identifier to be discussed below but used smaller diagnostic lexical data. For the Malay test sentences, only 12 sentences (13.3%) were correctly identified as Malay. The accuracy is higher for the Indonesian test sentences, but it is still far from satisfactory, that is, 24 sentences (26.7%). We also examined the accuracy of our language identifier using the same data. The results were not satisfactory either. The details of results for Ranaivo-Malançon's and our language identifier are summarised in Table 3. "Neutral" in the table means "indeterminate" in the sense that no evidence exists to identify a sentence as either Malay or Indonesian. It belongs to the **msa** category in our classification categories discussed in section 4.1. It is feared that language identification with sentence as the final unit will leave too many sentences as indeterminate.

Nevertheless, we first identified the language of each sentence contained in a page because the LCC data are provided in sentences. Next, the results of sentence-level identification were integrated at page level. Consider a web page consisting of six sentences. Suppose that four of them were identified as **ind**, one as **zsm** and one as **msa**. Page-level integration is a process whereby the initial results for all six sentences are overwritten by the most frequent result, if any. The most frequent result in our current example is **ind**. Hence, the relevant page is identified as **ind**.

Our sentence-level language identification utilises two kinds of lexical data and the country domains in URLs. The first lexical data are two lists of 1,000 frequent words, one for Malay and another for Indonesian.[6] The lists were created using the LCC subcorpora

[6] As an afterthought, the frequency lists could have been made with bigrams or trigrams, though the use of *n*-grams is also known to have problems such as data sparsity. *N*-gram frequency lists can capture collocational differences between Malay and Indonesian (e.g., *selamat datang ke* 'welcome to' in Malay vs. *selamat datang di* 'welcome at' in Indonesian). Differences in word usage can also be captured by *n*-grams. For example, Indonesian uses *punya* as a possession verb as in *dia punya alasan* [3SG PUNYA excuse] 's/he has an excuse'. Malay, however, never uses *punya* in this way and the same sequence is interpreted only as a noun phrase meaning 'his/her excuse'. The *n*-grams containing *punya* should thus differ considerably.

`msa_news_crawl_2011` (Malay) and `ind_news_crawl_2011` (Indonesian). These two subcorpora consist of data from established newspapers, and hence the original language identification is reliable. The two lists contain no overlap. High frequency words such as the relativiser *yang* and the demonstrative *itu* 'that' are usually common to both languages. They do not enable two languages to be distinguished and are thus not included in the lists. Also excluded from the lists are proper names denoting local people and places.[7] This is because they can cause misidentification in writings about local people and places. For example, writings about Jakarta normally contain many instance of the word *Jakarta*, whether they are written in Malay or Indonesian. If *Jakarta* is included in the high frequency word list of Indonesian, texts about Jakarta written in Malay are very likely to be misidentified as Indonesian. The following shows the ten most frequent words in a subcorpus of Malay and Indonesian that occur in the high frequency word list of the respective language.

(1) Ten most frequent words in Malay (`msa_news_crawl_2011`)
peratus 'per cent', *iaitu* 'namely', *setiausaha* 'secretary', *aktiviti* 'activity', *kewangan* 'financial', *ehwal* 'affairs', *pingat* 'medal', *kakitangan* 'staff', *mesyuarat* 'meeting', *dijangka* 'to be expected'

(2) Ten most frequent words in Indonesian (`ind_news_crawl_2011`)
wib 'Indonesian Western Standard Time', *kasus* 'case', *partai* 'party', *uang* 'money', *miliar* 'billion', *maupun* 'and, nor, although', *bagian* 'part', *senin* 'Monday', *kecamatan* 'subdistrict', *dprd* 'Regional House of Representatives'

The second lexical data are a list of 753 words spelt differently in Malaysia and Indonesia (Nomoto, Yamashita & Osaka 2014).[8] This list is based on *Kamus Dewan* (fourth edition) and the online version of *Kamus Besar Bahasa Indonesia* (third edition), which are the most authoritative dictionaries in Malaysia and Indonesia, respectively. Table 4 shows 10 sample words from this list. Note that not all spelling differences are always reliable when dealing with web data, in which many texts are not edited or proofread, unlike traditional paper-based books and magazines. For example, many instances of the Malaysian spelling *bahawa* 'that (complementiser)' are found in otherwise Indonesian texts (the Indonesian spelling is *bahwa*).

The decision tree in Figure 2 shows the overall architecture of our language identifier. The specific processes involved in it are as follows:

[7] To be more specific, names of local people, places and ethnic groups were excluded only if they were not abbreviated. Thus, clippings like *Sulut* (from *Sulawesi Utara* 'North Sulawesi') were not excluded. This is because such abbreviations are peculiar to the relevant country and are not understood outside it. Names of local political parties, companies and sports teams are excluded, whether abbreviated or not.

[8] The original list in Nomoto, Yamashita & Osaka (2014) consists of 758 words. During the development of the language identifier, five words were excluded from it. This is because their frequencies were very low in general but happened to occur frequently in some development data, resulting in identification errors.

Table 4. List of words spelt differently in Malaysia (zsm) and Indonesia (ind)

zsm	**ind**	Meaning
Aidiladha	*Iduladha*	'Feast of the Sacrifice'
Aidilfitri	*Idulfitri*	'Feast of Breaking the Fast
ais	*es*	'ice'
akaun	*akun*	'account'
akauntan	*akuntan*	'accountant'
akordion	*akordeon*	'accordion'
aksiom	*aksioma*	'axiom'
aktiviti	*aktivitas*	'activity'
aktres	*aktris*	'actress'
alaihissalam	*alaihisalam*	'peace be upon him'

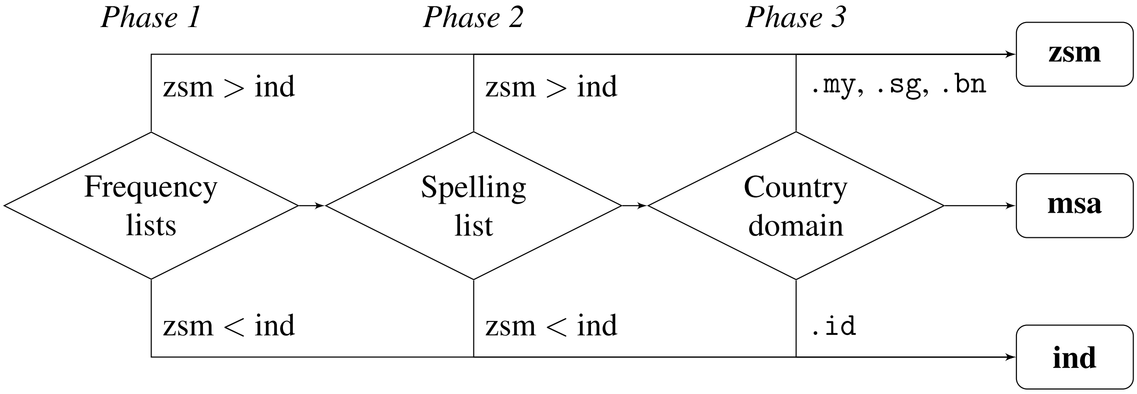

Figure 2. The overall architecture of our Malay-Indonesian language identifier

- Phase 1
  1. For each sentence, count the frequencies of the words included in the high frequency word lists.
  2. Assign to it the language code of the language with the higher frequency (**zsm** or **ind**).

     For example, a sentence in which words in the **zsm** list occur three times and words in the **ind** list occur once will be given the code **zsm**.
  3. Conduct page-level integration (see above).
  4. Proceed to Phase 2 if the page fails to be identified as either **zsm** or **ind**.
- Phase 2 (same process as Phase 1, but this time using the spelling difference list)
  1. For each sentence, count the frequencies of the words included in the spelling difference lists.
  2. Assign to it the language code of the language with the higher frequency (**zsm** or **ind**).
  3. Conduct page-level integration.
  4. Proceed to Phase 3 if the page fails to be identified as either **zsm** or **ind**.
- Phase 3
  1. Assign **zsm** if the country domain of the page's URL is `.my` (Malaysia), `.sg` (Singapore) or `.bn` (Brunei), and **ind** if the country domain is `.id` (Indonesia).
  2. If the page still fails to be identified, assign **msa**.

Table 5. The results of reclassification (Malay) (unit: page)

Subcorpus	Original	Malay (**zsm**)	Indonesian (**ind**)	Neutral (**msa**)	% of Indonesian
`msa_newscrawl_2011`	24,020	23,962	14	44	0.1
`ind-bn_web_2015`	456	455	1	0	0.2
`msa_wikipedia_2016`	104,444	55,575	5,228	43,641	5.0
Total	128,920	79,992	5,243	43,685	4.1

Table 6. The results of reclassification (Indonesian) (unit: page)

Subcorpus	Original	Malay (**zsm**)	Indonesian (**ind**)	Neutral (**msa**)	% of Malay
`ind_mixed_2012`	907,713	28,405	857,784	21,524	3.1
`ind_news_2008`, `ind_news_2009`, `ind_news_2010`, `ind_news_2011`, `ind_news_2012`, `ind_newscrawl_2011`, `ind_newscrawl_2012`	1,246,089	231	1,229,093	16,765	0.0
`ind_web_2011`, `ind_web_2012`, `ind-id_web_2013`, `ind-id_web_2015`	1,690,576	30,504	1,645,630	14,442	1.8
`ind_wikipedia_2016`	192,274	431	155,730	36,113	0.2
Total	4,036,652	59,571	3,888,237	88,844	1.5

Notice that the country domain information is used as the last resort because the country domain of a URL simply indicates the country where the URL is registered. Web pages written in Malay are often registered in Malaysia, Singapore or Brunei and those written in Indonesian are registered in Indonesia. However, this is not always the case. For instance, Indonesians living in Malaysia may write in Indonesian on pages registered in Malaysia, whose country domain is `.my`. Incidentally, after we had downloaded the subcorpora of Malay and Indonesian listed in section 3, LCC added three subcorpora with the Indian country domain `.in`. Two of them are categorised into Malay (`msa-in_web_2014` and `msa-in_web_2015`) and one into Indonesian (`ind-in_web_2015`). Because no large Malay- or Indonesian-speaking community is known in India, the real origin of the data in these subcorpora is not India but countries of the Malay Archipelago.

4.3 Results

Tables 5 and 6 show the results of reclassification for Malay and Indonesian, respectively. The second column shows the numbers of pages contained in the original LCC subcorpora. The pages have been reclassified as shown in the third through fifth columns.

The large proportion of the pages identified as "Neutral (**msa**)" may be accounted for by the following four factors, of which the first three are concerned with the nature of the data rather than our language identification methodology. The first factor is the resemblance between the two languages. A simple sentence such as *Saya makan nasi* [I eat rice] is equally acceptable and natural in both Malay and Indonesian. A page consisting of such sentences cannot and should not be categorised either as Malay or Indonesian. Instead, "Neutral " is the adequate classification for such a page.

Second, some pages are not long enough to give sufficient clues for language identification. Thus, although the average sentence count of the pages identified either as Malay (**zsm**) or Indonesian (**ind**) in `msa_wikipedia_2016` is 6.87 sentences, that of the pages identified as "Neutral (**msa**)" is only 4.64 sentences.

Third, the two *Wikipedia*-based subcorpora (i.e. `msa_wikipedia_2016` and `ind_wikipedia_2016`) have especially large numbers of "Neutral" pages, because the country domains in the URLs of *Wikipedia* pages are all `.org` and never `.my` or `.id`. Consequently, Phase 3 in our language identification system (see Figure 2), which makes reference to the country domain of a page's URL, has no effect.

The last factor has to do with the accuracy of our language identifier. It will be shown in the next section that our language identifier's ability to discriminate the exact language (i.e. **zsm** or **ind**) from the Malay macrolanguage (i.e. **msa**), which encompasses Malay and Indonesian, varies depending on genre, although its error rate (i.e. the rate of incorrectly identifying **zsm** texts as **ind** and vice versa) is consistently low, being 0.0%, except in one case (see Table 7).

Given the very low error rate of the language identifier, the pages that are reclassified into the other language (i.e. **ind** for Malay and **zsm** for Indonesian) are thought to be actually written in the other language. That is to say, 5.0% of the data in `msa_wikipedia_2016` is not Malay but Indonesian, and 3.1% of the data in `ind_mixed_2012` is not Indonesian but Malay. Indeed, manual inspection of a selection of `msa_wikipedia_2016` pages that were identified as Indonesian (**ind**) revealed that almost all were in fact written in Indonesian or were incomplete translations of originally Indonesian texts.

5. Evaluation

This section discusses the accuracy of our language identifier described in section 4.2. Furthermore, three options that could have been adopted in our language identification mechanism will be considered. The first option is concerned with the creation of the high frequency word lists. Recall that our high frequency word lists do not contain proper names denoting local people and places. What if those proper names were included in the lists? The second option has to do with the order of application for the two lexical lists. Is applying the high frequency word lists before the spelling difference list really better than the reverse order? It will be shown that these choices can affect the accuracy and that the options chosen in the current model are better than the alternatives. The third option is to combine the two lists into one.

5.1 Test data

Three sets of test data were prepared for each language. The three sets differ in their genres: news, Wiki and fiction. The news and Wiki components are similar to the data contained in LCC whereas fiction data are very limited. Each set consists of 50 files whose lengths are 458 tokens each. This number is based on the average length of the LCC data, i.e. 358 tokens/page. The additional 100 tokens are needed to examine changes in accuracy with different lengths.

News The news data consist of articles on the top page of the online version of a newspaper in Malaysia and Indonesia. The newspapers we chose are *Sinar Harian* for Malay (`http://www.sinarharian.com.my/`) and *Kompas* for Indonesian (`http://www.kompas.com/`). One hundred articles were extracted from both websites on 25 July 2017. For Malay, an additional 10 articles were extracted on 28 July 2017 because some pages in the existing data were shorter than 458 tokens and had to be combined with another.

Wiki We made use of the Asian Language Treebank Parallel Corpus (Riza et al. 2016; `http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/index.html`, downloaded on 19 July 2017) for the Wiki component of our test data. This corpus was built by translating English *Wikinews* articles into seven Asian languages, including Malay and Indonesian. The Wiki component differs from the news component above in that they are translations, which is a notable characteristic of *Wikipedia* pages in general. Although translations are somewhat artificial, as pointed out by one of the reviewers, the Wiki component has an advantage that the Malay and Indonesian data do not differ in terms of their semantic contents. Fifty URLs whose translations exceed 458 tokens in both languages were randomly selected.

Fiction 50 *cerpen*s or short stories longer than 458 tokens were collected from the following online *cerpen* collection sites on 26 July 2017: *Penulisan2u* (`http://www.penulisan2u.my/`) for Malay and *Cerpenmu* (`http://cerpenmu.com/`) for Indonesian.

For each component above, we examined the accuracies for different lengths by trimming the texts at a particular length. Below we will report the means for the lengths in the range of 358 ± 50 tokens.[9] The URL information of the sources was not taken into consideration. In other words, the evaluation is concerned only with the first two phases of our language identification algorithm, which rely on lexical data, but not the third phase, which uses the country domains in URLs (see Figure 2).

5.2 Results

Table 7 shows the results. In addition to the results for the three components, the results are also given for the combined category of the news and Wiki components, as it is this category that is the closest in composition to the LCC data. Note that the accuracy of our language identifier when applied to the LCC data are thought to be higher than the figures shown here because our test data lack the variability in the country domains present in the LCC data. Note also that presenting the results in terms of precision and recall is not useful here. This is because our test data do not contain any true negatives, i.e. Indonesian sentences in Malay test data and Malay sentences in Indonesian test data; the precision will always be 1.0.[10]

As can be seen from Table 7, our language identifier's ability to identify the exact language (i.e. the percentage of "Correct") varies considerably among different genres. The

[9] We omit the results for the lengths in the range of 358 ± 100 tokens, as they do not differ significantly from the figures reported here.

[10] As one of the reviewers correctly pointed out, we should have created an LCC-like, mixed language test set, which is a random selection of documents from the web, and hand-coded language categories. Then, we could have assessed the effect of URLs and presented the results in a conventional fashion with (meaningful) precisions, recalls and F-scores.

small percentages for the fiction component were anticipated, as a part of our language identification system is based on the news crawl subcorpora of LCC. Our language identifier performed better for the Wiki component than for the news component probably because the topics covered are wider in the latter, including entertainment and religion. Crucially, however, the error rate is consistently quite low overall: 0.0–2.0% for Malay and 0.0% for Indonesian. In short, our Malay-Indonesian language identifier is often uncertain about the decision between Malay and Indonesian, but it seldom makes a wrong judgement.

To make a comparison with an existing language identifier, we also tested one of the currently most popular language identifiers `langid.py` (Lui & Baldwin 2012) using the same test data. `langid.py` is an off-the-shelf language identification tool that has been pre-trained on 97 languages and is designed to function consistently well in a variety of domains. The results are shown in Table 8. "Error" in Malay means Indonesian (**ind**) and "Error" in Indonesian means Malay (**zsm**). The category "Neutral" is absent from the table because `langid.py` does not have a language category corresponding to our **msa**.[11] The most crucial difference between our Malay-Indonesian language identifier and `langid.py` is that the latter frequently makes wrong judgements. In Malay, the error rates are extremely high. Therefore, language classification using `langid.py` will unduly reduce the size of Malay data and expand the Indonesian data with non-Indonesian data. It must be noted that `langid.py` is a perfect tool to differentiate Malay and Indonesian (i.e. Malay macrolanguage) from other languages; it did not identify any of our test data as other languages such as Javanese. However, it needs to be supplemented by an additional Malay-Indonesian specific language identifier like ours when applied to actual tasks, where the distinction between Malay and Indonesian is indispensable.

[11] One might think that "Neutral" can be created using the confidence scores produced by `langid.py`, i.e. regarding as being indeterminate those results whose confidence scores are lower than a certain cut-off point. Such a manipulation is theoretically possible, but not in reality. This is because the confidence score is 1.0 in almost all cases, regardless of the accuracy of the identification.

Table 7. Mean accuracies tested on 50 files of different genres with lengths 358 ± 50 tokens

(a) Malay

	News			Wiki			News + Wiki			Fiction		
	Correct	Error	Neutral	Correct	Error	Neutral	Correct	Error	Neutral	Correct	Error	Neutral
M	27.4	0.0	22.6	49.0	1.0	0.0	76.4	1.0	22.6	0.0	0.0	50.0
SD	1.5	0.0	1.5	0.0	0.0	0.0	1.5	0.0	1.5	0.0	0.0	0.0
%	54.9	0.0	45.1	98.0	2.0	0.0	76.4	1.0	22.6	0.0	0.0	100.0

(b) Indonesian

	News			Wiki			News + Wiki			Fiction		
	Correct	Error	Neutral	Correct	Error	Neutral	Correct	Error	Neutral	Correct	Error	Neutral
M	35.0	0.0	15.0	50.0	0.0	0.0	85.0	0.0	15.0	7.6	0.0	42.4
SD	1.2	0.0	1.2	0.0	0.0	0.0	1.2	0.0	1.2	1.5	0.0	1.5
%	69.9	0.0	30.1	100.0	0.0	0.0	85.0	0.0	15.0	15.1	0.0	84.9

Table 8. Mean accuracies tested on 50 files of different genres with lengths 358 ± 50 tokens using `langid.py`

(a) Malay

	News		Wiki		News + Wiki		Fiction	
	Correct	Error	Correct	Error	Correct	Error	Correct	Error
M	28.5	21.5	15.8	34.2	44.3	55.7	15.3	34.7
SD	0.9	0.9	0.9	0.9	1.4	1.4	0.7	0.7
%	57.0	43.0	31.6	68.4	44.3	55.7	30.7	69.3

(b) Indonesian

	News		Wiki		News + Wiki		Fiction	
	Correct	Error	Correct	Error	Correct	Error	Correct	Error
M	49.8	0.2	49.0	1.0	98.8	1.2	49.6	0.4
SD	0.4	0.4	0.0	0.0	0.4	0.4	0.5	0.5
%	99.7	0.3	98.0	2.0	98.8	1.2	99.1	0.9

As an aside, we should note that correctly identifying Malay is more difficult than correctly identifying Indonesian. Notice that the percentages of "Correct" are lower in Malay than in Indonesian for all test data components in Tables 7 and 8. The same pattern is also found with Ranaivo-Malançon's (2006) Malay-Indonesian language identifier (Table 3). Furthermore, Lui (2014:214) presented a confusion matrix pointing to a similar and, in fact, even more radical pattern: although Indonesian is usually identified correctly as Indonesian (77.1%), Malay is more often confused as Indonesian (63.6%) than is correctly identified (24.2%).[12]

We conjecture that the different levels of difficulty in language identification between the two languages are due to the following factors. First, the two languages differ in lexical diversity (i.e. type-token ratio). Table 9 shows the lexical diversity values for our test data described in section 5.1. Malay has lower values than Indonesian for the news and Wiki components.[13] What this means is that if one creates a frequency list with the same length in both languages, the Malay list is more likely to include infrequent words than the Indonesian list. Thus, the frequency list in Malay is less useful as a language identification diagnostic than that in Indonesian.

The second point, which in fact is related to the first, has to do with preferred methods of new word formation. In our observation, which needs to be supported by an objective study in the future, when forming new words, Malay normally relies on compounding two or more existing words whereas Indonesian often creates totally new forms by clipping or changing the spelling of a foreign word so it conforms to the Indonesian phonotactics and

[12] The confusion matrix also contains other languages such as Javanese and Sundanese.

[13] The reverse pattern holds with the fiction component. It is not clear why the lexical diversity of Indonesian *cerpen*s is particularly low. One possibility is that *cerpen*s generally exhibit low lexical diversity values, as the Indonesian data indeed do, but Malay *cerpen*s are lexically more diverse because they contain a considerable number of English words as a result of frequent code-mixing.

Table 9. Lexical diversity in the test data

Component	Malay	Indonesian
News	0.193	0.204
Wiki	0.191	0.201
Fiction	0.193	0.182

Table 10. New word formation in Malay and Indonesian[14]

Malay	Indonesian		Meaning
lapangan terbang [field fly]	*bandara*	< *bandar udara* [port air]	‘airport’
Korea Utara [Korea north]	*Korut*	< *Korea Utara* [Korea north]	‘North Korea’
pilihan raya [selection grand]	*pemilu*	< *pemilihan umum* [selection general]	‘general election’
dalam talian [in line]	*daring*	< *dalam jaringan* [in net]	‘online’
telefon bimbit [telephone carry]	*ponsel*	< *telepon seluler* [telephone cellular]	‘mobile phone’
reka bentuk [invent form]	*desain*		‘design’
soal selidik [question research]	*kuesioner*		‘questionnaire’
separuh akhir [half final]	*semifinal*		‘semifinal’

orthography. Some examples are given in Table 10. Compounds whose component elements are separated by a white space will be ignored and treated as multiple tokens unless multi-word expression detection is incorporated into the tokenisation process involved in language identification (and lexical diversity calculation). An adequate treatment of compounds (multi-word expressions) is essential for improving the accuracy of Malay-Indonesian language identification.

Let us now turn to the evaluation of the three choices that we made in creating our language identifier: (i) proper names denoting local people and places were excluded from the high frequency word lists; (ii) the high frequency word lists were applied before the spelling difference list; (iii) the two kinds of lexical data were kept separate and applied one by one instead of combining the two into one.

With and without proper names Table 11 shows the results when the high frequency word lists of the original language identifier are replaced by those containing proper names denoting local people and places. Compare this table with Table 7 above. The results are mixed. In Malay, excluding proper names from the lists has negative effects

[14] For clippings, the longer forms are also used but less frequently.

Table 11. Mean accuracies tested on 50 files of different genres with lengths 358 ± 50 tokens using the high frequency word lists containing proper names

(a) Malay

	News			Wiki			Fiction		
	Correct	Error	Neut.	Correct	Error	Neut.	Correct	Error	Neut.
M	27.7	0.0	22.3	49.0	1.0	0.0	0.0	0.0	50.0
SD	1.8	0.0	1.8	0.0	0.0	0.0	0.0	0.0	0.0
%	55.5	0.0	44.5	98.0	2.0	0.0	0.0	0.0	100.0

(b) Indonesian

	News			Wiki			Fiction		
	Correct	Error	Neut.	Correct	Error	Neut.	Correct	Error	Neut.
M	34.4	0.0	15.6	50.0	0.0	0.0	5.2	0.0	44.8
SD	2.2	0.0	2.2	0.0	0.0	0.0	0.7	0.0	0.7
%	68.8	0.0	31.2	100.0	0.0	0.0	10.4	0.0	89.6

on the news component; fewer pages are identified correctly ($t = -6.5, p < .001, d = .18$) whereas more pages remain indeterminate ($t = 6.5, p < .001, d = .18$). By contrast, it has positive effects on two components in Indonesian, i.e. news and fiction; more pages are identified correctly (news: $t = 4.4, p < .001, d = .32$; fiction: $t = 20.7, p < .001, d = 2.0$) whereas fewer pages remain indeterminate (news: $t = -4.4, p < .001, d = .32$; fiction: $t = -20.7, p < .001, d = 2.0$). No statistically significant difference is found elsewhere. The original order is considered better because it has effects on more components in Indonesian, with larger effect sizes.

Different orders of lexical list application Table 12 shows the results when the order of lexical list application of the original language identifier is switched around, i.e. the spelling difference list is applied before the high frequency word lists. Compare Table 12 with Table 7 above. The original order performs better in the Indonesian fiction component in that it can identify more pages correctly ($t = 6.0, p < .001, d = .18$) and leaves fewer pages as indeterminate ($t = -6.0, p < .001, d = .18$). No statistically significant difference exists elsewhere.

Two separate lists and one combined list Table 13 shows the results when the high frequency word list and the spelling difference list are combined into one. Compare Table 13 with Table 7. The results are mixed, but very different between Malay and Indonesian. In Malay, keeping the two lists separate has negative effects on the news component; fewer pages are identified correctly (news: $t = -6.7, p < .001, d = .18$) and more pages remain indeterminate ($t = -6.7, p < .001, d = .18$). In Indonesian, on the other hand, the performance increases substantially if the two kinds of list are kept separate. The effects are found in all components, of which those on the news and fiction components are statistically significant; more pages are identified correctly (news: $t = 16.7, p < .001, d = 1.3$; fiction: $t = 22.6, p < .001, d = 2.6$) whereas fewer pages remain indeterminate (news: $t = -16.7, p < .001, d = 1.3$; fiction: $t = -20.6, p < .001, d = 2.6$). No statistically

Table 12. Mean accuracies tested on 50 files of different genres with lengths 358 ± 50 tokens with the reverse list application order

(a) Malay

	News			Wiki			Fiction		
	Correct	Error	Neut.	Correct	Error	Neut.	Correct	Error	Neut.
M	27.4	0.0	22.6	50.0	0.0	0.0	0.0	0.0	50.0
SD	1.5	0.0	1.5	0.0	0.0	0.0	0.0	0.0	0.0
%	54.9	0.0	45.1	100.0	0.0	0.0	0.0	0.0	100.0

(b) Indonesian

	News			Wiki			Fiction		
	Correct	Error	Neut.	Correct	Error	Neut.	Correct	Error	Neut.
M	35.0	0.0	15.0	49.0	0.0	1.0	7.3	0.0	42.7
SD	1.2	0.0	1.2	0.0	0.0	0.0	1.4	0.0	1.4
%	69.9	0.0	30.1	98.0	0.0	2.0	14.6	0.0	85.4

significant difference exists elsewhere. The original model in Figure 2, where the two lists are not combined but applied one after another, is considered better because it has effects on more components in Indonesian, with much larger effect sizes.

6. Conclusion

The Malay and Indonesian subcorpora of LCC have issues with language identification, where a corpus of one language is mixed with data from the other language. We have reclassified the data using our Malay-Indonesian language identifier. The reclassified data have already been sent to the Leipzig team and will be openly available on their website. Moreover, the reclassified data have been annotated for morphological information such as affixes and reduplication types using MALINDO Morph (Nomoto et al. 2018), and six subcorpora have been made searchable through the open online concordancer MALINDO Conc (Nomoto, Akasegawa & Shiohara 2018). Only the data reclassified as **zsm** or **ind** were used for these purposes.

This study has implications for web corpus building in similar languages in general. In data collection, it is important to ensure that every page contains a sufficient number of sentences to guarantee high accuracy in language identification (see section 4.3). Moreover, it was found that the presence or absence of proper names denoting local people and places in the diagnostic lexical data affects the accuracy of language identification (see section 5.2). Although positive and negative effects were observed in our test data, we believe that it is generally better to remove local proper names, especially for genres whose contents do not differ across languages such as *Wikipedia*. A comparison of the results obtained for the Malay *Wikipedia* subcorpus of LCC (`msa_wikipedia_2016`) with and without proper names supports this view. The percentage of Indonesian increases if proper names denoting local people and places are included in the high frequency word lists, as shown in Table 14. Manual examination of the increased pages reveals that they are texts on Indonesian themes written in Malay, hence the errors in language identification.

Table 13. Mean accuracies tested on 50 files of different genres with lengths 358 ± 50 tokens with one combined lexical list

(a) Malay

	News			Wiki			Fiction		
	Correct	Error	Neut.	Correct	Error	Neut.	Correct	Error	Neut.
M	27.7	0.0	22.3	49.0	1.0	1.0	0.0	0.0	50.0
SD	1.8	0.0	1.8	0.0	0.0	0.0	0.0	0.0	0.0
%	55.5	0.0	44.5	98.0	2.0	2.0	0.0	0.0	100.0

(b) Indonesian

	News			Wiki			Fiction		
	Correct	Error	Neut.	Correct	Error	Neut.	Correct	Error	Neut.
M	32.9	0.0	17.1	49.0	0.0	1.0	4.6	0.0	45.4
SD	2.0	0.0	2.0	0.0	0.0	0.0	0.6	0.0	0.6
%	65.8	0.0	34.2	98.0	0.0	2.0	9.2	0.0	90.8

Table 14. Effects of local proper names on language identification in `msa_wikipedia_2016` **(unit: page)**

Variables	Malay (**zsm**)	Indonesian (**ind**)	Neutral (**msa**)	% of Indonesian
No proper names (Table 5)	55,575	5,228	43,641	5.0
With proper names	55,413	5,395	43,636	5.2

The present study also has implications for the use of *Wikipedia* data in linguistic research and language resource development. As shown in Table 14, as much as 5% of Malay *Wikipedia* pages are in fact not written in Malay; they are either written in Indonesian or incomplete translations of Indonesian into Malay. The number is truly fatal at least in descriptive and theoretical linguistics, though it may not be in natural language processing; no serious linguist dares to claim that a dataset is from language A when s/he is aware that every one out of 20 examples is from language B. The lesson here is that if similar languages exist, the language classification in *Wikipedia* is not necessarily correct, and hence *Wikipedia* data must be handled with caution.

Abbreviations

3	third person	SD	standard deviation
M	mean	SG	singular

References

KBBI. 2001. *Kamus besar bahasa Indonesia*, edisi ketiga. Jakarta: Balai Pustaka.

KD. 2005. *Kamus Dewan*, edisi keempat. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Asmah Haji Omar. 1992. Malay as a pluricentric language. In Michael G. Clyne (ed.), *Pluricentric languages: Differing norms in different countries*, 401–420. Berlin: Mouton de Gruyter.

Asmah Haji Omar. 2001. The Malay language in Malaysia and Indonesia: From lingua franca to national language. *The Aseanists ASIA* II. `http://stateless.freehosting.net/AA2EMalay.htm` (Accessed on 13 May 2017).

Bird, Steven, Edward Loper & Ewan Klein. 2009. *Natural language processing with Python: Analyzing text with the Natural Language Toolkit.* Sebastopol, CA: O'Reilly Media Inc.

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)*,

Lui, Marco & Timothy Baldwin. 2012. `langid.py`: An off-the-shelf language identification tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, 25–30.

Lui, Marco Hoiyin. 2014. *Generalized language identification*. University of Melbourne dissertation.

Nomoto, Hiroki, Shiro Akasegawa & Asako Shiohara. 2018. Building an open online concordancer for Malay/Indonesian. Paper presented at the 22nd International Symposium on Malay/Indonesian Linguistics (ISMIL).

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 & Kartini Abd. Wahab. 2011. *Kena* passives in Indonesian: A Malaysian perspective. Paper presented at the 15th International Symposium on Malay/Indonesian Linguistics (ISMIL).

Nomoto, Hiroki, Nahoko Yamashita & Ayano Osaka. 2014. Senarai komprehensif perbezaan ejaan Malaysia dan ejaan Indonesia. In *Gogaku kenkyuujo ronshuu 19*, 21–31. Tokyo: Tokyo University of Foreign Studies.

Ranaivo-Malançon, Bali. 2006. Automatic identification of close languages—case study: Malay and Indonesian. *ECIT Transactions of Computers and Information Technology* (2). 126–134.

Riza, Hammam, Michael Purwoadi, Gunarso, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thai, Vichet Chea, Rapid Sun, Sethserey Sam, Sopheap Seng, Khin Mar Soe, Khin Thandar Nwet, Masao Utiyama & Chenchen Ding. 2016. Introduction of the Asian Language Treebank. In *Oriental COCOSDA*.

Simons, Gary F. & Charles D. Fennig (eds.). 2018. *Ethnologue: Languages of the world*, 21st edn. Dallas, TX: SIL International. `http://www.ethnologue.com` (Accessed on 31 August 2018).

Tan, Liling, Marcos Zampieri, Nikola Ljubešić & Jörg Tiedemann. 2014. Merging comparable data sources for the discrimination of similar languages: The DSL corpus collection. In *Proceedings of the Seventh Workshop on Building and Using Comparable Corpora (BUCC)*, 6–10.