Fossicking in dominant language teaching: Javanese and Indonesian 'low' varieties in language teaching resources

Zara Maxwell-Smith
The Australian National University / Canberra, ACT, Australia
Zara.Maxwell-Smith@anu.edu.au

Abstract

'Low' and 'high' varieties of Indonesian and other languages of Indonesia are poorly resourced for developing human language technologies. Many languages spoken in Indonesia, even those with very large speaker populations, such as Javanese (over 80 million), are thought to be threatened languages. The teaching of Indonesian language focuses on the prestige variety which forms part of the unusual diglossia found in many parts of Indonesia. We developed a publicly available pipeline to scrape and clean text from the PDFs of a classic Indonesian textbook, *The Indonesian Way*, creating a corpus. Using the corpus and curated wordlists from a number of lexicons I searched for instances of non-prestige varieties of Indonesian, finding that they play a limited, secondary role to formal Indonesian in this textbook. References to other languages used in Indonesia are usually made as a passing comment. These methods help to determine how text teaching resources relate to and influence the language politics of diglossia and the many languages of Indonesia.

1 Introduction

The teaching of Indonesian as a foreign language grapples with over 70 years of intense language politics and planning under presidents Sukarno and Suharto and during the *Reformasi* era after Suharto's fall (Heryanto, 1995; Sneddon, 2003b). Situated in a context of large-scale language shift and endangerment (Ravindranath and Cohn, 2014), in which estimates of 300 languages are dying or in trouble (Zein, 2020), Indonesian language teaching may have some influence on the future of these languages. In fact, Anderbeck noted the role of the educational domain in the spectacular spread of Indonesian at the expense of other languages (2015). As teaching contexts are complex and students' needs diverse, individual teachers are best placed to decide how to present the languages of Indonesia to their students. However, in order to make these decisions teachers need to be able to browse and understand the materials on offer; a process which can be improved through the use of technology.

Human language technologies offer methods to support linguistic analysis of teaching materials but 'low' and 'high' varieties of Indonesian and other languages of Indonesia are poorly resourced for developing those methods (Riza, 2019; Nomoto, 2020). For instance, publicly available lexicons are few in number and suffer from problems of high levels of 'noise' from English and other languages (Wilie et al., 2020). Technologies which combine qualitative and quantitative text analysis allow systematic work at scale (Andreotta et al., 2019) and can be useful in this context despite shortages in natural language processing resources for Indonesian and other languages used in Indonesia.

In this project, converting Indonesian language teaching resources to a searchable corpus enabled me to fossick, or 'sift for gold'; finding varieties of Indonesian and other languages of Indonesia. This paper describes the use of an innovative pipeline designed to help teachers assess teaching resources and engage with Indonesian language technologies and politics.

The paper presents a brief description of the endangerment of languages in Indonesia and the difficulties in choosing teaching resources for Indonesian teachers. It describes the methods used to convert the teaching resources into a corpus and presents an analysis of how a subset of important vocabulary items are presented in the *The Indonesian Way* (TIW). It demonstrates how extracting English and Standard Indonesian from the corpus creates a subset of data more amenable to human analysis, thereby revealing instances of other languages. Finally, it discusses the applicability of

Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages: Vol. 1 Papers, pages 24–32,
Online, March 2–3, 2021.

this approach to other teaching resources as well as text resources from other domains.

2 Background

Of the 707 languages spoken in Indonesia at least 300 are dying or 'in trouble' (Zein, 2020, p. 130 for detailed discussion). Ravindrath and Cohn found that even languages with very large speaker populations are threatened:

> "In terms of language endangerment then it seems there is no such thing as "too big to fail"." (2014, p. 73)

Thus, while ethnologue.com lists Javanese in categories *Large*[1] and *Institutional (EGIDS 0-4)*[2] (2020), there are indications that the necessary inter-generational transmission may not be occurring (2009; 2013).

'Indonesian' is usually thought to be behind the language shift endangering other languages of Indonesia (Anderbeck, 2015). However, the diglossic nature of Indonesian itself is complex. Multiple non-standard *'low'* (L) varieties are used in different geographical regions while a *'high'* (H) or 'standard' variety dominates education and formal events (Sneddon, 2003a). Further, a mid-diglossic may be evolving in mass media and everyday speech, all of which should be considered by teachers (Nataprawira, 2018).

Teaching Indonesian as a foreign language requires reference to the so-called 'native' speakers of the language. The development of Indonesia as an imagined community (Anderson, 1991), creates an imagined community of language speakers (Norton, 2001) whose linguistic status is often inferred from their membership of the nation-state. In foreign language teaching, the diversity of Indonesian varieties often collapse into the H/standardised variety, or 'bahasa baku'. There is "a quite common assumption that 'Indonesian' refers solely to the formal language" (2003a); the H variety is positively evaluated, to the detriment of L varieties. I take the position that official lines which denigrate non-standard Indonesian are inter-woven with negative perceptions of other languages (i.e., some Javanese speakers describe their language as "old fashioned" and speakers are "poor and village-like" (Setiawan, 2013)).

[1]The language has more than 1,000,000 users

[2]The language has been developed to the point that it is used and sustained by institutions beyond the home and community

Teaching the Indonesian H variety can be a careful decision; a convenient simplification for beginning students; something simply not considered when designing materials. Regardless of the reason for this common decision, when choosing resources it is extremely difficult for teachers to assess how each resource relates to language diversity, or how the imagined 'native' speaker will be described in the materials they use. Computationally assessing the presence of other languages and Indonesian varieties within teaching resources allows teachers, students and researchers to do more than flick through a resource to get a sense of the language in it.

While one approach could be to search teaching resources for instances of endangered languages, the poor computational resourcing of many of these languages would compromise the outcome, not to mention the difficulties of sourcing the hundreds of possible languages which could be present. Although it is well-documented that languages in Indonesia, including Indonesian, share and borrow from each other (Haspelmath and Tadmor, 2009), subtracting computationally well-supported English lexicons and some basic Standard Indonesian lexicons can produce a data subset which is richer in the information sought by this study. The pipeline described in the next section presents this method for stakeholders to learn and think about the way in which Indonesian is taught.

3 Materials and Methods

This project uses a pipeline developed in partnership with Australian technology company Appen, to scrape and clean text from the PDFs and other documents. It is publicly available via the Co-EDL Github text-helpers repository. We used our pipeline to process the classic Indonesian textbook, TIW (Quinn and Kozok, 2016), creating a corpus. TIW is commonly used as a first year introductory textbook in tertiary education.

The pipeline (Figure 1) unites and normalizes data from the 9 PDF files of TIW. It then processes and cleans data with bespoke Indonesian and teaching-genre scripts, producing lists of corpus types for human inspection and annotation. Human annotations, along with curated English, Indonesian, Javanese, and Sundanese lexicons normalized by the pipeline were used for further linguistic processing and cleaning to create a searchable corpus. These same lexicons were then used in operations

Wordlist	Source	Size (in words)
`\words`	System File (Unix)	235,886
CMU English	CMU pronunciation dictionary (Rudnicky, 2015)	134,429
Australian English	Australian English Lexicon (Anderson, 2017)	128,913
Manual English List	Created during data cleaning steps	808
KOIN Indonesian	Korpus Indonesia (KOIN) (Kwary, 2019)	17,254
Wordnet Indonesian	Wordnet Bahasa (Bond et al., 2014)	107,224
Colloquial Indonesian	Kamus Alay (Salsabila et al., 2018)	4,332
Standard Indonesian	Kamus Alay (Salsabila et al., 2018)	2,004
Javanese Lexicon	(Google, 2018a)	53,893
Sundanese Lexicon	(Google, 2018b)	42,855

Table 1: Wordlists

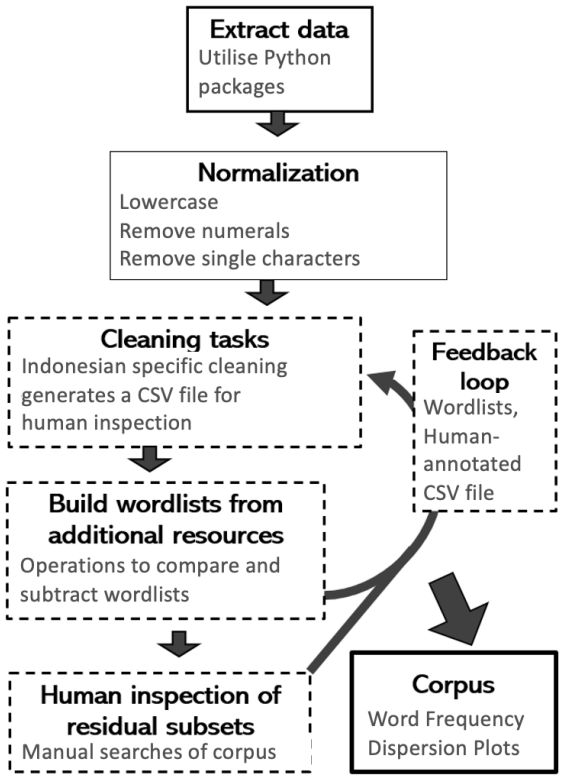

Figure 1: Pipeline.

to compare, search and strip the teaching resource as reported in the results section. In particular, I subtracted the Standard Indonesian and English lexicons to create a residual subset suitable for human analysis.

Cleaning artefacts particular to this text genre required steps to strip punctuation and remove various strings of special characters, numerals and letters that represented exercise numbers in the textbook modules. I noted language specific cleaning included difficulties processing hyphens. In this particular dataset, strings included four different Unicode hyphens, some of which would best be used to split a token, while others were contained in valid Indonesian lexemes (strings such as "anak-anak" (*children*) and "kehijau-hijauan" (*greenish*)) which are frequently occurring tokens.

Curated Indonesian and English wordlists were built to investigate the properties of the language from different perspectives. The best coverage of words in the corpus (see Chapter 2 in Nation (2016) for discussion of types, words, and word families) was achieved by combining three English lexicons created from the Unix system file /words (Unix) (235,886 words), the CMU pronunciation dictionary (Rudnicky, 2015) (134,429 words) and an Australian English Lexicon (Anderson, 2017) (128,913 words). A manual list of English words in TIW which were not found in the lexicons above was created during the cleaning process (808 words).

Indonesian wordlists were built from the Indonesian Corpus/Korpus Indonesia (KOIN) (Kwary, 2019) (17,254 words) and Wordnet Bahasa (Bond et al., 2014) (107,224). The Colloquial Indonesian Lexicon/Kamus Alay (Salsabila et al., 2018), was used to build two lexicons. Firstly a lexicon of colloquial language which included various 'online' slang and word spellings (4,332 words) as well as lexicon of the standard forms paired with slang in Kamus Alay (2,004 words). Javanese and Sundanese wordlists were built from Google's languages resources with minor modifications (Google, 2018a,b).

By subtracting the wordlists of English and Standard Indonesian, the pipeline created a dataset small enough to be 'picked over' manually. This subset, *'Residual Types'* is equal to TIW lexicon minus the wordslists: *CMU English*, *Australian*

English and *Manual English List* as well as *KOIN Indonesian*, *Wordnet Indonesian* and *Standard Indonesian* (see Table 1).

Given that the languages I look at in this project borrow prolifically from each other (Haspelmath and Tadmor, 2009), and likely have thousands of words in common, I did not try to distinguish between them in a broad sense. Instead I estimated that words in common with lists of Standard Indonesian are very likely to be being used in a Standard Indonesian sense in these teaching resources. For example, while 'puncak' (*mountain peak*) is present in Standard Indonesian and Javanese, it is much more likely being used in its Standard Indonesian sense in a teaching resource such as this. Removing the Standard Indonesian word lists thereby produced a list of residual types more suitable for examination. Individual words from this list were then assessed for their provenance and contextual use.

4 Results

TIW includes approximately 335,165 tokens and 13,385 individual types (unique strings). In general, TIW focuses on the H variety of Indonesian. The highest frequency Indonesian words in TIW (Table 2) include only the H variety of Indonesian[3]. Consistent with a note in the middle of Lesson 4 which asserts that the textbook is mostly 'formal Indonesian' and at times includes some "informal (or slangy) usage" (2016, Page 35, Module 1 of 8), the searchable corpus allowed me to find subsections of informal Indonesian (see Figure 5). Meanwhile, Figure 2 visualises the dispersion of a set of informal vocabulary; noting that these informal forms are characteristic of Jakartan Indonesian (JI) as described by Sneddon 2006.

Further evidence of the usage and dispersion of Jakartan Indonesian can be seen in the representation of negators (Figure 3). The H Indonesian negator "tidak" (*not*) has strong presence throughout the resource, while JI "nggak" (*not*) is concentrated in the earlier sections of the resource (this is also true of the JI in Figure 2). Negators from regional 'L' Indonesian varieties and other languages of Indonesia are included for comparison. The negators "kagak" (Betawi) and "ndak" (Minangkabau) occurred as part of a brief description of Indonesian linguistic systems.

From these plots it appears that JI in TIW does not build as a target of language acquisition as students progress. In this way, word dispersion plots are very useful for teachers when designing assessment and planning their classes. For instance, the word "begini" (*so/like this*), which only occurs towards the end of the resource (with one exception), would need to be excluded from any early assessments. Likewise, given the early focus and then drop away of JI in TIW, it would appear that students are indeed "invited to explore" Jakartan Indonesian (Quinn and Kozok, 2016), but that it is not presented as a serious target for language acquisition nor suitable for assessment in the final stages of the course.

This finding is moderated by the findings on Indonesian pronouns presented in Figure 4. Frequent and broad inclusion of informal Indonesian pronouns "aku" (*I*) and "kamu" (*you*) alongside the more formal "saya" (*I*) and "Anda" (*you*) indicate that TIW does not perpetuate what Sneddon describes as a:

> "failure to recognize anything but the most formal variety as Indonesian [which] has frequently led to its being stigmatized as a 'soulless' and alienating language." (Sneddon, 2003a)

The use and teaching of Indonesian pronouns is a prominent aspect of the literature on Indonesian (see Morgan 2011 and Djenar 2006) and requires teachers to take a particular stance. Information about how resources present various options for the first and second person pronoun is (or should be) of central concern to teachers. Detailed information about how these pronouns occur in resources has been incredibly opaque, perhaps even to authors of textbooks. Computational tools which allow large-scale analysis of such important questions for language teachers are useful for all stakeholders.

To detect other languages of Indonesia in TIW I compared lexicons of Javanese and Sundanese to the TIW lexicon. There were 2,181 types and 83,176 tokens in common with the Google Javanese wordlist once English was removed from the corpus. However, human inspection revealed that Google's lexicon not only included a large number of Indonesian words which might be considered loaned into Javanese, but also words that would usually be regarded as Indonesian rather

[3]Noting that "apa" has both a formal (*what*) and informal form (*closed question marker/or*) in the textbook and this would contribute to its presence in the top 20 most frequent words.

Word		Frequency	Word		Frequency
di	*(in/at/on)*	5,391	dengan	*(with)*	1,195
saya	*(I/me/my)*	4,074	apakah	*(question marker)*	1,182
yang	*(that/which)*	2,598	ibu	*(mum, Mrs)*	950
tidak	*(not)*	1,926	rumah	*(house)*	928
dan	*(and)*	1,912	dari	*(from)*	902
latihan	*(exercise)*	1,742	apa	*(what)*	874
anda	*(you - Formal)*	1,678	ini	*(this)*	874
itu	*(that)*	1,426	suka	*(like)*	830
ada	*(there is/exists)*	1,361	hari	*(day)*	788
ke	*((go) to)*	1,333	sekali	*(very)*	751

Table 2: The most frequent Indonesian words in TIW.

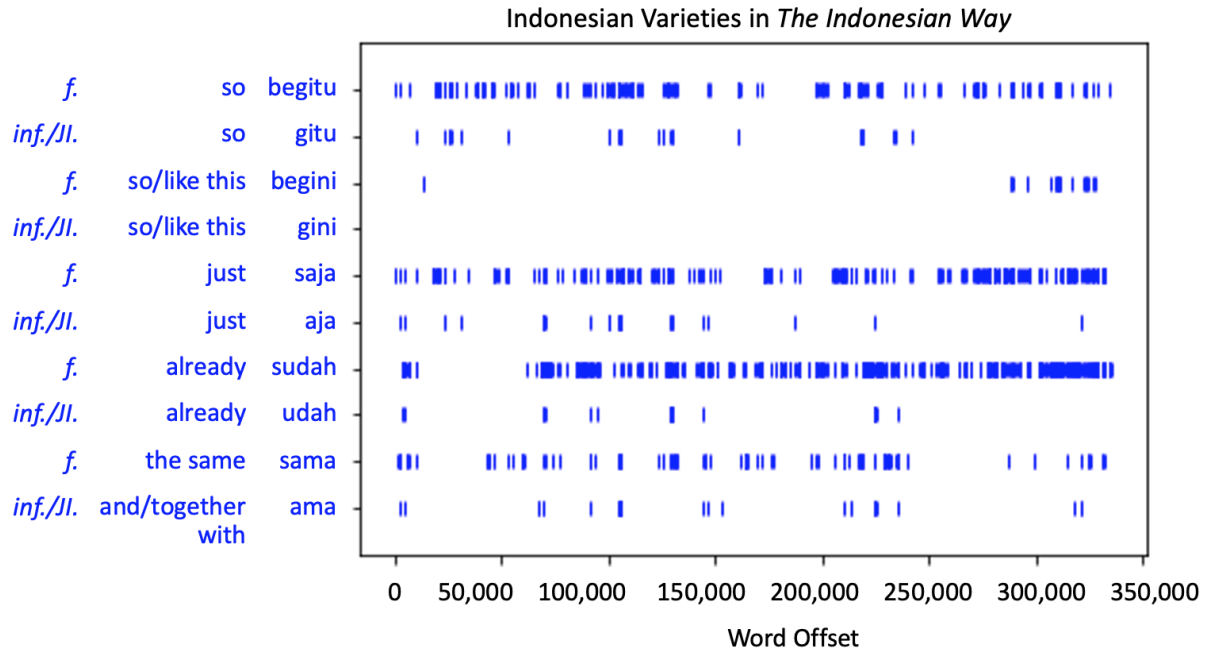

Figure 2: Lexical dispersion of standard and colloquial varieties. The standard variety form is listed first followed by the informal or Jakartan Indonesian form. [*f. formal, inf. informal*]

than Javanese[4]. It should be noted that Google's lexicon also contains a large number of English words (there were 10,946 types in common with a compilation of my 4 English lists). Similar doubts about the quality of the Sundanese lexicon arise with 1,903 types and 52,638 tokens in common between TIW and the Google Sundanese wordlist. These wordlists do not appear useful in determining the presence of Javanese nor Sundanese in this resource; a different approach was more successful.

The subset *Residual Types* (see Methods and Materials) contained 1,656 words, including a mix of noise such as typographical errors and proper nouns. However, at that length it was possible for a human to scan it and pick out terms for further examination. This process brought forth a number of sections in which different ethnic groups and their languages were mentioned (albeit in passing). A search for specific Javanese tokens from *Residual Types* unearthed numerous references to Javanese culture. For example, searching for "mripat" (*eyes*) unearthed:

> "in the Javanese language a special, compulsory vocabulary exists for referring to other people's bodies. You refer to your own eyes as your *mripat* but to someone else's as their *tingal* or *sotya*." [italics in original]

[4]For example the lexicon includes 'ketika' (*when*) which is 'nalika' in Javanese and 'dua' (*two*) which is 'loro'.

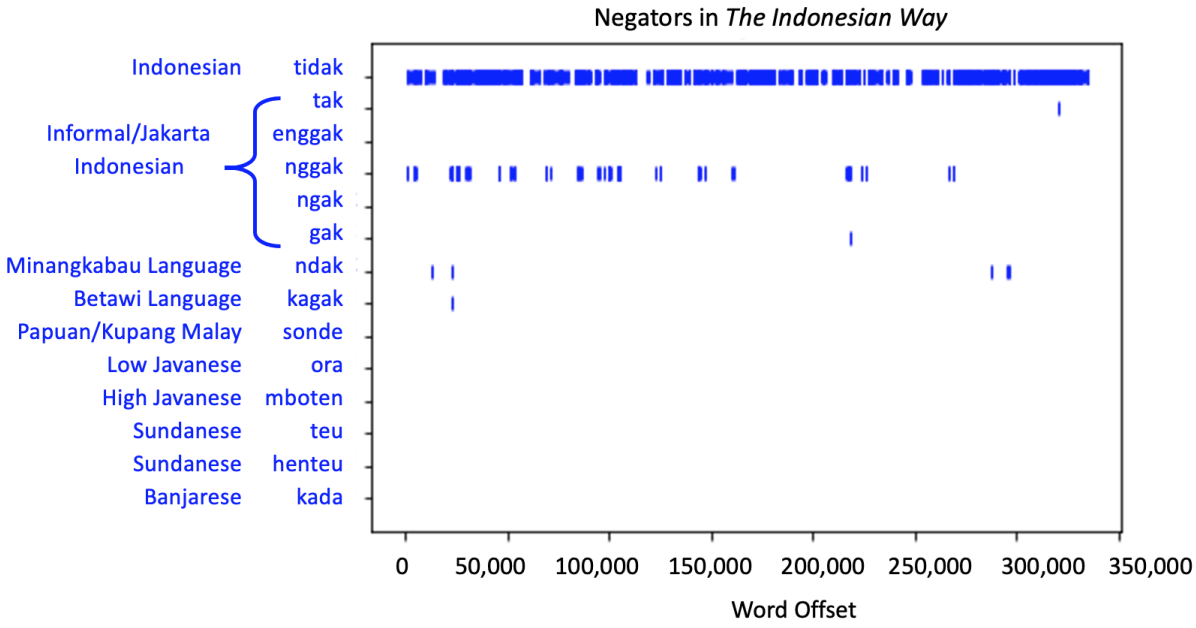

Figure 3: Lexical Dispersion of Negators

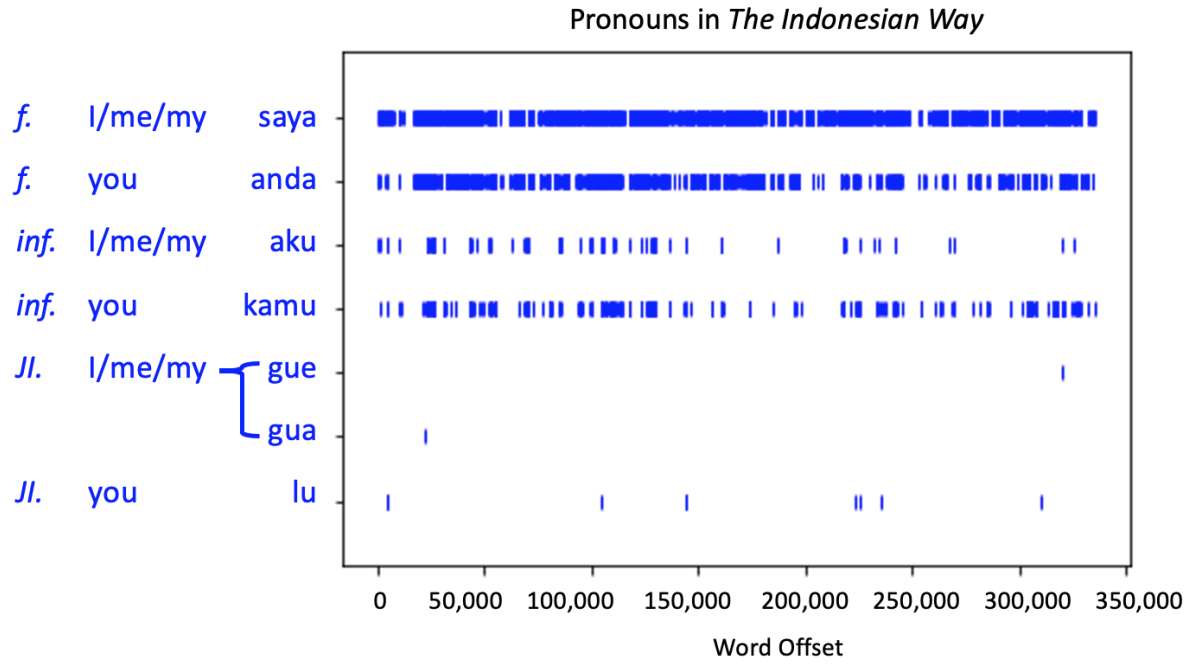

Figure 4: Lexical Dispersion of Pronouns [*f. formal, inf. informal*]

While mentions of Indonesia's other languages are made in passing, using residual tokens to search the corpus, proved a fruitful method of discovering how and where these mentions are made. This process provides an extremely useful method for educators to investigate how a resource approaches cultural and linguistic diversity.

5 Discussion

This pipeline offers a new insight into TIW which can inform stakeholders about its approach to varieties of Indonesian (such as its approach to informal styles of Indonesian — see Figure 5) and other languages as well as issues of central concern such as pronoun use (Morgan, 2011). As a publicly available pipeline (Apache 2.0 license), the pipeline offers stakeholders with suitable skill

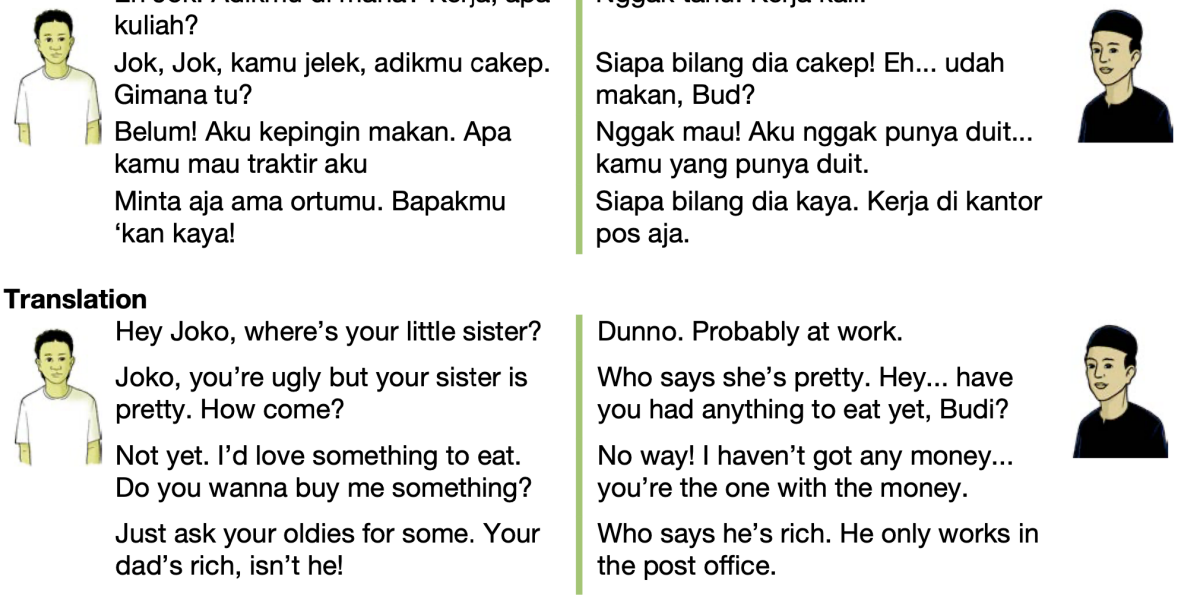

Figure 5: Excerpt of sample 'informal style' dialogue from Lesson 22, Module 2 of The Indonesian Way

sets and resources the opportunity to analyze other Indonesian teaching resources, as well as other genres of Indonesian-English text.

As the first analysis of its kind (as far as I am aware), it is not yet possible to offer comparison of TIW to other textbooks. However, with some adaptions these methods could be applied to other Indonesian teaching resources. The pipeline also has potential in the analysis of school resources used in Indonesia; resources which have a major impact on language planning and diversity in Indonesia (Zein, 2020).

With further adaptions, the pipeline may be suitable for use with other languages/language pairs. Parts of the pipeline are useful for researchers looking at low-resource languages in other parts of the world, particularly when examining how these languages are represented alongside and within dominant language texts such as newspapers.

I noted some vulnerabilities in the cleaning process relating to the conversion of PDF files to the corpus, especially in relation to the use of multiple fonts within a single word. There were also unresolved abnormalities due to page breaks in the PDFs which resulted in word strings being concatenated. While I estimated these to be infrequent occurrences in this dataset, I recommend careful attention to such issues in other uses of the pipeline. This is especially important with teaching resources which are often characterised by varied fonts and text which help human readers' comprehension. I also note that any Indonesian words which are spelled identically in English would likely be removed by English lexicons and as such, calculating the percentage of the text in either English or Indonesian requires a more sophisticated pipeline. Tools described by Uliniansyah et al (2013) and Amalia et al (2019) may provide some solutions to these issues, but are not publicly available.

This work also contributes to the development and scrutiny of Indonesian NLP resources (Nomoto, 2020) which require significant investment given the population of Indonesian exceeds 270 million (WorldBank, 2019). It shows how using the same computational tools regularly used in industry can be helpful to advance the education sector (Maxwell-Smith et al., 2020), even for languages with scarce resources.

6 Conclusion

Computational resources for Indonesian are blossoming (Wilie et al., 2020), though it is still an under-resourced language, especially given its size. Alone these resources are not particularly helpful to time-poor teachers. Conventional applications used to access resources do not provide the means to assess the usage of language/s in a teaching resource. The pipeline this project used to analyze and provide insight into TIW, can illustrate how teaching resources relate to language politics with empirical data about the representation or lack of

representation of different varieties of Indonesian, as well as other languages of Indonesia.

Beyond the teaching setting, a pipeline of this nature can assist researchers to understand how poorly resourced languages are represented in various text sources. With further development, it has potential to inform our understanding of language change and usage and thereby assist revitalisation efforts.

Acknowledgments

I gratefully acknowledge my partnership with Appen in the development of this pipeline, in particular I am grateful to Romi Hill for her dedication and patience. I would like to thank anonymous reviewers, my supervisors Hanna Suominen and Danielle Barth, and other colleagues for their contributions. I am also very grateful to the authors of *The Indonesian Way* for their enthusiasm for this work.

References

Amalia Amalia, Opim Salim Sitompul, Erna Budhiarti Nababan, Maya Silvia Lydia, and Nadia Rahmatunnisa. 2019. Bahasa Indonesia Text Corpus Generation Using Web Corpora Approaches. *Journal of Theoretical and Applied Information Technology*, 97(24).

Karl Anderbeck. 2015. Portraits of language vitality in the languages of Indonesia. *Language documentation and cultural practices in the Austronesian world: Papers from 12-ICAL*, 4:19–47.

Benedict Anderson. 1991. *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso.

Tom Anderson. 2017. *australian-english-lexicon.txt*. The English dictionaries come directly from SCOWL and is thus under the same copyright of SCOWL. The affix file is a heavily modified version of the original english.aff file which was released as part of Geoff Kuenning's Ispell and as such is covered by his BSD license. Part of SCOWL is also based on Ispell thus the Ispell copyright is included with the SCOWL copyright. The collective work is Copyright 2000-2016 by Kevin Atkinson as well as any of the copyrights mentioned below: Copyright 2000-2016 by Kevin Atkinson.

Matthew Andreotta, Robertus Nugroho, Mark J. Hurlstone, Fabio Boschetti, Simon Farrell, Iain Walker, and Cecile Paris. 2019. Analyzing social media data: A mixed-methods framework combining computational and qualitative text analysis. *Behavior Research Methods*, 51:1766–1781.

Francis Bond, Lian Tze Lim, Enya Kong Tang, and Hammam Riza. 2014. The combined Wordnet Bahasa. *NUSA: Linguistic studies of languages in and around Indonesia*, 57:83–100.

Dwi Noverini Djenar. 2006. Patterns and variation of address terms in colloquial indonesian. *Australian Review of Applied Linguistics*, 29(2):22–22.

Ethnologue. 2020. Size and vitality of javanese.

Google. 2018a. Javanese lexicon.

Google. 2018b. Sundanese lexicon.

Martin Haspelmath and Uri Tadmor. 2009. *Loanwords in the World's Languages: A Comparative Handbook*. De Gruyter Mouton, Berlin, Boston.

Ariel Heryanto. 1995. *Language of development and development of language : the case of Indonesia*. Dept. of Linguistics, Research School of Pacific Studies, The Australian National University.

Deny A. Kwary. 2019. A corpus platform of indonesian academic language. *SoftwareX*, 9:102 – 106.

Zara Maxwell-Smith, Simón González Ochoa, Ben Foley, and Hanna Suominen. 2020. Applications of Natural Language Processing in Bilingual Language Teaching: An Indonesian-English Case Study. In *Proceedings of the 15th Workshop on Innovative Use of NLP for Building Educational Applications*, page 124–134.

Anne-Marie Morgan. 2011. Me, myself, I: exploring conceptions of self and others in Indonesian names and pronouns with early learners. *Babel*, 45(2-3):26.

Halim Nataprawira. 2018. *Recognising the Sociolinguistic Reality of Spoken Indonesian: A Corpus and Usage Analysis of A Middle Diglossic Variant*. Ph.D. thesis, University of the Sunshine Coast, Queensland.

Paul Nation. 2016. *Making and Using Word Lists for Language Learning and Testing*. John Benjamins.

Hiroki Nomoto. 2020. Towards genuine stemming and lemmatization in malay/indonesian. *Proceedings of the 26th Annual Meeting of the Natural Language Processing Society (March 2020)*.

Bonny Norton. 2001. Non-participation, imagined communities and the language classroom. In M Breen, editor, *Learner contributions to language learning: New directions in research*, pages 159–171. Cascadilla Press.

George Quinn and Uli Kozok. 2016. *The Indonesian Way: Modules 1-8*. The development of "The Indonesian Way" was sponsored by grant P017A090375-10 from the US Department of Education, International Research and Studies Program. The development of the print version was made possible by a grant received from the University of Tasmania.

Maya Ravindranath and Abigail C. Cohn. 2014. Can a language with millions of speakers be endangered? *Journal of the Southeast Asian Linguistics Society (JSEALS) 7 (2014): 64-75*.

Hamam Riza. 2019. Oriental-OCOSDA 2019 Indonesia country report. *2019 22nd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–7.

Alexander Rudnicky. 2015. *CMUdict (the Carnegie Mellon Pronouncing Dictionary) - cmudict-0.7b*. Copyright (c) 2015, Alexander Rudnicky.

Nikmatun Aliyah Salsabila, Yosef Ardhito Winatmoko, Ali Akbar Septiandri, and Ade Jamal. 2018. Colloquial Indonesian Lexicon. *2018 International Conference on Asian Language Processing (IALP)*, pages 226–229.

Slamet Setiawan. 2013. *Children's Language in a Bilingual Community in East Java*. Ph.D. thesis, The University of Western Australia, School of Social Sciences.

Nancy J. Smith-Hefner. 2009. Language shift, gender, and ideologies of modernity in central java, indonesia. *Journal of Linguistic Anthropology*, 19(1):57–77.

James Neil Sneddon. 2003a. Diglossia in Indonesian. *Bijdragen tot de taal-, land-en volkenkunde/Journal of the Humanities and Social Sciences of Southeast Asia*, 159(4):519–549.

James Neil Sneddon. 2003b. *The Indonesian Language: Its History and Role in Modern Society*. UNSW Press.

James Neil Sneddon. 2006. *Colloquial Jakartan Indonesian*, volume 581. Pacific Linguistics, Research School of Pacific and Asian Studies, Australian National University, Canberra.

Teduh Uliniansyah, Hammam Riza, and Oskar Riandi. 2013. Developing corpus management system for Bahasa Indonesia the "Perisalah" project. *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, pages 1–4.

Unix. dict\words. In *Unix system file, usually stored in /usr/share/dict/words or /usr/dict/words*.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, and Ayu Purwarianti. 2020. IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding. *The 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 9th International Joint Conference on Natural Language Processing (AACL-IJCNLP 2020)*.

WorldBank. 2019. Indonesia data.

Subhan Zein. 2020. *Language Policy in Superdiverse Indonesia*. Taylor Francis Group.