One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia

Anonymous ACL submission

Abstract

NLP research is impeded by a lack of resources and awareness of the challenges presented by under-represented languages and dialects. Focusing on the languages spoken in Indonesia, the second most linguistically diverse and the fourth most populous nation of the world, we provide an overview of the current state of NLP research for Indonesia's 700+ languages. We highlight challenges in Indonesian NLP and how these affect the performance of current NLP systems. Finally, we provide general recommendations to help develop NLP technology not only for languages of Indonesia, but also other underrepresented languages.

1 Introduction

Research in natural language processing (NLP) has traditionally focused on developing models for English and a small set of other languages with large amounts of data (see Figure 1, bottom right). While the lack of data is generally cited as the key reason for the lack of progress in NLP for underrepresented languages (Hu et al., 2020; Joshi et al., 2020), we argue that another factor relates to the diversity and the lack of understanding of the linguistic characteristics of such languages. Through the lens of the languages spoken in Indonesia, the world's second-most linguistically diverse country, we seek to illustrate the challenges in applying NLP technology to a such diverse pool of languages.

Indonesia is the 4th most populous nation globally with 273 million people spread over 17,508 islands. There are more than 700 languages spoken in Indonesia, equal to 10% of the world's languages, second only to Papua New Guinea (Eberhard et al., 2021). However, most of these languages are not well documented in the literature; many are not formally taught and no established standard exists across speakers (Novitasari et al., 2020). Many of them are decreasing in use, as Indonesian (*Bahasa Indonesia*), the national language, is more frequently used as the primary language across the country. This process may ultimately result in a monolingual society (Cohn and Ravindranath, 2014). One study finds that among 98 Indonesian local languages, 36 are considered safe, 51 are endangered, and 11 are already extinct (Anindyatri and Mufidah, 2020).

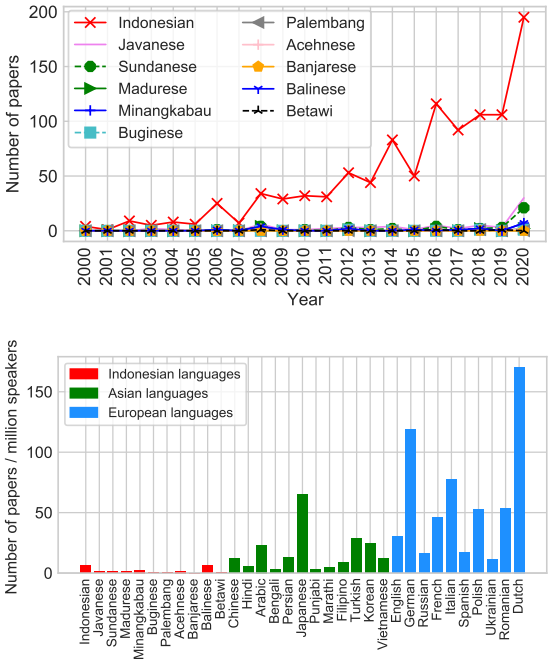

Figure 1: Following Joshi et al. (2020), we compile ACL Anthology (main conferences and workshops) to count the distribution of published works that mention Indonesian languages. **Top:** Distribution of papers in 20 years. **Bottom:** Distribution per million speakers, compared to other Asian and European languages.

Table 1 shows the 10 Indonesian local languages with the most speakers, along with Indonesian for comparison (Eberhard et al., 2021). Javanese and Sundanese are at the top with 84M and 34M speakers respectively, while Madura, Minangkabau, and Buginese have around 6M speakers. Despite their large speaker populations, these local languages are poorly represented in the NLP literature. Compared to Indonesian, the number of research papers

Language	ISO	# Speakers
Indonesian	id	198 M
Javanese	jav	84 M
Sundanese / Sunda	su	34 M
Madurese / Madura	mad	7 M
Minangkabau	min	6 M
Buginese	bug	6 M
Betawi	bew	5 M
Acehnese / Aceh	ace	4 M
Banjar	bjn	4 M
Balinese	ban	3 M
Palembang Malay (musi)	mus	3 M

Table 1: The 10 most spoken Indonesian local languages according to Ethnologue (Eberhard et al., 2021).

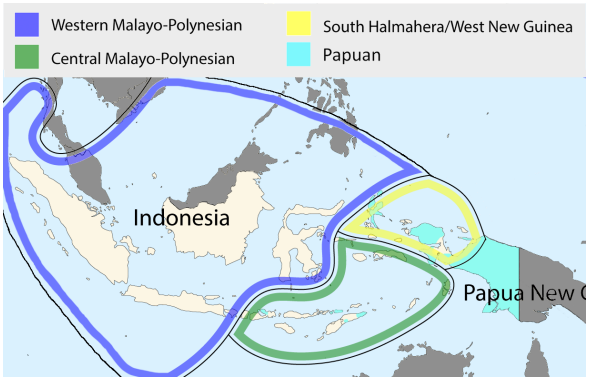

Figure 2: Map of Austronesian and Papuan languages in Indonesia

mentioning these languages has barely increased over the past 20 years (Figure 1, top). Furthermore, compared to their European counterparts, Indonesian languages are drastically understudied (Figure 1, bottom). This is true even for Indonesian, which has nearly 200M speakers.

Language technology should be accessible to everyone in their native languages (European Language Resources Association, 2019), including Indonesians. In the context of Indonesia, language technology research offers some benefits. First, language technology is one of the potential peacemaker tools in a multi-ethnic country, helping Indonesians understand each other better and avoid the ethnic conflicts of the past (Bertrand, 2004). On a larger scale, language technology promotes language use (European Language Resources Association, 2019) and helps language preservation. Despite these benefits, following Bird (2020), we recommend a careful assessment of individual usage scenarios of language technology so they are implemented for the good of the local population.

For language technology to be useful in the Indonesian context, it additionally has to account for the dialects of local languages. Dialects in Indonesia are influenced by the geographical location and regional culture of their speakers (Vander Klok, 2015) and thus often differ substantially in morphology and vocabulary, posing challenges for NLP systems. In this paper, we provide an overview of the current state of NLP for Indonesian languages. We then discuss the challenges presented by those languages and demonstrate how they affect state-of-the-art systems in NLP. We finally provide recommendations for developing better NLP technology not only for languages in Indonesia but also other under-represented languages.

2 Background and Related Work

2.1 History and Taxonomy

Indonesia is one of the richest countries in the world in terms of linguistic diversity. More than 400 of its languages belong to the Austronesian language family, while the others are Papuan languages spoken in the eastern part of the country. As shown in Figure 2, the Austronesian languages in Indonesia belong to three main groups: Western-Malayo-Polynesian (WMP), Central-Malayo-Polynesian (CMP), and South-Halmahera-West-New-Guinea (SHWNG) (Blust, 1980). WMP languages are Malay, Indonesian, Javanese, Sundanese, Balinese, and Minangkabau, among others. All languages mentioned in Table 1 are in this group. Languages belonging to CMP are languages of the Lesser Sunda Islands from East Sumbawa (with Bimanese) onwards to the east, and languages of the central and southern Moluccas (including the Aru Islands and the Sula Archipelago). The SHWNG group consists of languages of Halmahera and Cenderawasih Bay, as far as the Mamberamo River, and of the Raja Ampat Islands. The Papuan languages, meanwhile, are mainly spoken in Papua, such as Dani, Asmat, Maybrat, and Sentani. Some Papuan languages are also spoken in Halmahera, Timor, and the Alor Archipelago (Palmer, 2018; Ross, 2005).

Most Austronesian linguists and archaeologists agree that the original 'homeland' of Austronesian languages must be sought in Taiwan and, prior to Taiwan, in coastal South China (Adelaar, 2005; Bellwood and Dizon, 2008; Bellwood et al., 2011). The Austronesian people moved from Taiwan to the Philippines in the second millennium CE. From the Philippines, they moved southward to Borneo and

Sulawesi. From Borneo, they migrated to Sumatra, the Malay Peninsula, Java, and even to Madagascar. From Sulawesi, they moved southward to the CMP area and eastward to the SHWNG area. From there, they migrated to Oceania and Polynesia, as far as New Zealand, Easter Island, and Hawaii (Gray and Jordan, 2000). The people that lived in insular Southeast Asia, such as in the Philippines and Indonesia, before the arrival of Austronesians were Australo-Melanesians (Bellwood, 1997). Gradual assimilation with Austronesians occurred although some pre-Austronesian groups still survive such as Melanesian people in eastern Indonesia (Ross, 2005; Coupe and Kratochvíl, 2020).

At the time of the arrival of the first Europeans, Malay had become the major language (lingua franca) of interethnic communication in Southeast Asia and beyond (Steinhauer, 2005; Coupe and Kratochvíl, 2020). It functioned as the language of trade and the language of Islam because Muslim merchants from India and the Middle East were the first to introduce the religion into the harbor towns of Indonesia. After the arrival of Europeans, Malay was used by the Portuguese and Dutch to spread Catholicism and Protestantism. When the Dutch extended their rule over areas outside Java in the nineteenth century, the importance of Malay increased, and thus, the first standardization of the spelling and grammar occurred in 1901, based on Classical Malay (Abas, 1987; Sneddon, 2003). In 1928, the participants of the Second National Youth Congress proclaimed Malay (henceforth called Indonesian) as the unifying language of Indonesia. During World War II, the Japanese occupying forces forbade all use of Dutch in favor of Indonesian, which from then onward effectively became the new national language. After independence until the present, Indonesian has functioned as the main language in education, mass media, and government activities. Many local language speakers are increasingly using Indonesian with their children because they believe it will help them toward a better education and career (Klamer, 2018).

2.2 Efforts in Multilingual Research

Recently, pretrained multilingual language models such as mBERT (Devlin et al., 2019), mBART (Liu et al., 2020), and mT5 (Xue et al., 2021b) were proposed. Their coverage, however, focuses on high-resource languages. Among them, only mBERT and mT5 include Indonesian local languages, i.e., Javanese, Sundanese, and Minangkabau, but with comparatively little pretraining data.

Some multilingual datasets for question answering (TyDi QA; Clark et al., 2020), dialogue (XPersona; Lin et al., 2021), passage ranking (mMARCO; Bonifacio et al., 2021), cross-lingual visual question answering (xGQA; Pfeiffer et al., 2021), common sense reasoning (XCOPA; Ponti et al., 2020), abstractive summarization (Hasan et al., 2021), language and vision reasoning (MaRVL; Liu et al., 2021), and machine translation (FLORES-101; Goyal et al., 2021) include Indonesian but most others do not, and very few include Indonesian local languages. An exception is the weakly supervised named entity recognition dataset, WikiAnn (Pan et al., 2017), which covers several Indonesian local languages, namely Acehnese, Javanese, Minangkabau, and Sundanese.

Parallel corpora including Indonesian local languages are i) CommonCrawl; ii) Wikipedia parallel corpora like MediaWiki Translations[1] and WikiMatrix (Schwenk et al., 2021); iii) the Leipzig corpora (Goldhahn et al., 2012), which include Indonesian, Javanese, Sundanese, Minangkabau, Madurese, Acehnese, Buginese, Banjar, and Balinese; and iv) JW-300 (Agić and Vulić, 2019), which includes dozens of Indonesian local languages, e.g., Batak language groups, Javanese, Dayak language groups, and several languages in Nusa Tenggara.[2]

2.3 Progress in Indonesian NLP

Most NLP research on Indonesian has been done across multiple topics, such as sentiment analysis (Naradhipa and Purwarianti, 2011; Lunando and Purwarianti, 2013), hate speech detection (Alfina et al., 2017; Ibrohim and Budi, 2019; Sutejo and Lestari, 2018), morphological analysis (Pisceldo et al., 2008), POS tagging (Wicaksono and Purwarianti, 2010; Dinakaramani et al., 2014; Kurniawan and Aji, 2018), named entity recognition (Budi et al., 2005; Gunawan et al., 2018), question answering (Mahendra et al., 2008; Fikri and Purwarianti, 2012), machine translation (Yulianti et al., 2011), and speech recognition (Lestari et al., 2006; Baskoro and Adriani, 2008; Zahra et al., 2009). However, many of these studies either kept the data private or used non-standardized resources with a lack of documentation and open-sourced code,

[1] https://mediawiki.org/wiki/Content_translation

[2] Recent studies (Caswell et al., 2021), however, have raised concerns regarding the quality of such multilingual corpora for under-represented languages.

which makes them extremely difficult to reproduce. Recently, Wilie et al. (2020), Koto et al. (2020b), and Cahyawijaya et al. (2021) collect Indonesian NLP resources as benchmark data. Others have also begun to create standardized labelled data for Indonesian NLP, e.g. the works of Kurniawan and Louvan (2018), Guntara et al. (2020), Mahendra et al. (2021), Koto et al. (2021), and Artari et al. (2021).

On the other hand, a handful of NLP research explore the local languages. Suryani et al. (2015) study machine translation in Sundanese by using prior POS tagging information, while Suryani et al. (2018) develop a word stemmer for Sundanese. Koto and Koto (2020) explore sentiment analysis and machine translation for Minangkabau. Safitri et al. (2016) work on spoken data language identification in three Indonesian local languages, i.e., Minangkabau, Sundanese and Javanese. Azizah et al. (2020) develop end-to-end neural text-to-speech model for Indonesian, Sundanese, and Javanese. Recently, Cahyawijaya et al. (2021) established a machine translation benchmark in Sundanese and Javanese using Bible data. Wibowo et al. (2021) studied a family of colloquial Indonesian, which is influenced by some local languages via morphological transformation, and Putri et al. (2021) worked on abusive language and hate speech detection on Twitter for five local languages, namely Javanese, Sundanese, Madurese, Minangkabau, and Musi.

3 Challenges for Indonesian NLP

3.1 Limited Resources

Monolingual Data Unlabelled corpora are crucial for building large language models, such as GPT-2 (Radford et al., 2019) or BERT (Devlin et al., 2019). Available unlabelled corpora such as Indo4B (Wilie et al., 2020) and Indo4B-Plus (Cahyawijaya et al., 2021) mainly include data in Indonesian, with the latter containing ≈10% of data in Javanese and Sundanese (see Appendix C). In comparison, in multilingual corpora such as CC-100 (Conneau et al., 2020), Javanese and Sundanese data accounts for only 0.001% and 0.002% of the corpus size while in mC4 (Xue et al., 2021b), there are only 0.6M Javanese and 0.3M Sundanese tokens out of a total of 6.3T tokens.

In addition, we measure data availability in Wikipedia, compared to the number of speakers as in Figure 3.[3] Among highly spoken local languages, much fewer data is available for Indonesian languages, compared to European languages with similar numbers of speakers. For example, Wikipedia contains more than 3 GB of Italian articles but less than 50 MB of Javanese articles, despite both languages having a comparable number of speakers. Similarly, Sundanese has less than 25 MB of articles, whereas languages of comparable speakers have more than 1.5 GB of articles. Similar trends hold for most other Asian languages.[4]

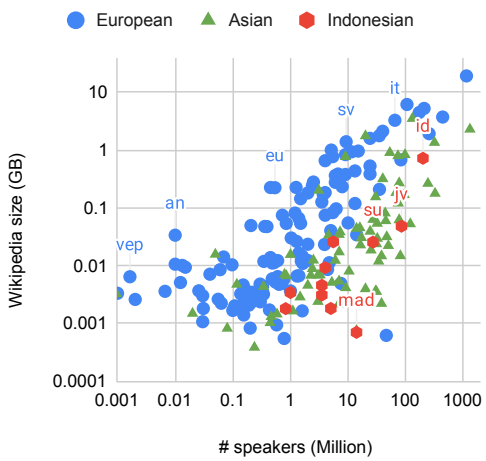

Figure 3: Wikipedia data size (in GB) compared to the number of speakers.

Beyond the most spoken local languages, other Indonesian local languages do not have Wikipedia instances, in contrast to European languages with few speakers. However, it is very difficult to find alternative sources for high-quality text data for other local languages of Indonesia (such as news websites), as most such sources are written in Indonesian. Resources in tail languages are even more lacking, due to a very low number of speakers. Moreover, most of the languages in the long tail are mainly used in a spoken context, making text data difficult to obtain.

These statistics demonstrate that collecting unlabelled corpora for Indonesian local languages is extremely difficult. This makes it impractical to develop strong pretrained language models for these languages, which have been the foundation for many recent state-of-the-art NLP systems.

Labelled Data Most work on Indonesian NLP (see §2) did not publicly release their data or mod-

[3]The number of speakers is collected from Wikidata (Vrandečić and Krötzsch, 2014), from the `number of speakers (P1098)` property as of Nov 7th 2021, while the size is collected from the 20211101 Wikipedia dump.

[4]Other continents such as Africa are even more underrepresented in terms of Wikipedia data (see Appendix D).

English	Mudung Laut	Dusun Teluk	Mersam	Suo Suo	Teluk Kuali	Lubuk Telau	Bunga Tanjung	Pulau Aro
I/me	sayo	aku	awaʔ	sayo	kito, awaʔ	am^{b}o	ambo	ambo
You	kau, kamu	kau	ka^{d}n	kamu	kaan	kamu	aŋ, kau, kayo	baʔaŋ
he/she	dioʔ	dioʔ, ɲo	ɲo	kau	ɲo	ɲo	ɲo	iɲo
if	kalu	jiko, kalu	kalu	bilao	kalu	jiko	koʔ	kalu
one	satu	sekoʔ	sekoʔ	sekoʔ	ciɜʔ	sekoʔ	sekoʔ, so	sekoʔ

Table 2: Lexical variation of Jambi Malay across different villages in Jambi, collected from Anderbeck (2008).

English	**Context**	**Ngoko**			**Krama**
		Western	Central	Eastern	Eastern
I/me	I like to eat fried rice.	inyong, enyong	aku	aku	kulo
You	Where will you go?	rika, kowe, ko	kowe, siro, sampeyan	koen, awakmu, sampeyan	panjenengan
How	How do I read this?	priwe	piye	yo'opo	pripun
Why	Why is this door broken?	ngapa	ngopo	opo'o	punapa
Will	Where will you go?	arep	arep	kate, ate	badhe
Not/no	The calculation is not correct.	ora	ora	gak	mboten

Table 3: Lexical variation of Javanese dialects and styles across different regions of Java island. Native speakers were asked to translate the words, given the context.

els, thus limiting reproducibility. Although recent Indonesian NLP benchmarks are addressing this issue, they mostly focus on the Indonesian language (see Appendix C). Some widely spoken local languages such as Javanese, Sundanese, or Minangkabau have extremely small labelled datasets compared to Indonesian, while others have barely any.

The lack of such datasets makes NLP development for the local languages difficult. However, constructing new labelled datasets is still challenging due to (1) the lack of speakers of some languages, (2) the vast continuum of dialectical variation (see §3.2.1), and (3) the absence of writing standard in most local languages (see §3.3).

3.2 Language Diversity

The diversity of Indonesian languages is not only due to the large number of local languages, but also the large number of dialects of these languages (§3.2.1). Speakers of local languages also often mix languages in conversation, which makes colloquial Indonesian more diverse (§3.2.2). In addition, some local languages are more commonly used in conversational contexts, so they do not have consistent writing forms in written media (§3.3).

3.2.1 Regional Dialects and Style Differences

Indonesian local languages often have multiple dialects, depending on the geographical location. Local languages of Indonesian spoken in different locations might be different (have some lexical variation) to one another, despite still being categorized as the same language (Fauzi and Puspitorini, 2018). For example, Anderbeck (2008) shows that villages across the Jambi province use different dialects of Jambi Malay. Similarly, Kartikasari et al. (2018) show that Javanese between different cities in central and eastern Java can have more than 50% lexical variation, while Purwaningsih (2017) shows that Javanese in different districts in the Lamongan Regency has up to 13% lexical variation. Similar studies have been conducted on other languages, such as Balinese (Maharani and Candra, 2018) and Sasak (Sarwadi et al., 2019).

Moreover, Indonesian and its local languages have multiple styles, even within the same dialect. One factor that affects style is the level of politeness and formality—similar to Japanese and other Asian languages (Bond and Baldwin, 2016). More polite language is used when speaking to a person with a higher social position, especially to elders, seniors, and sometimes strangers. Different politeness levels manifest in the use of different honorifics and even different lexical terms.

To illustrate the distinctions between regional dialects and styles, we highlight common words and utterances across dialects and styles in Jambi Malay and Javanese in Tables 2 and 3. For Jambi Malay, we sample the result from Anderbeck (2008). For Javanese, we ask native speakers to translate basic words across three regional dialects: Western, Central, and Eastern Javanese, and two different styles: *Ngoko* (standard, daily-use Javanese) and *Krama* (polite Javanese, used to communicate to elders and those with higher social status). However, since contemporary *Krama* Javanese is not very different between regions, we only consider *Krama* from Eastern speaker's perspective.

Style	Region	Model				
		Langid.py		FastText		CLD3
		Top-1	Top-3	Top-1	Top-3	Top-1
Ngoko	Western	0.241	0.621	0.069	0.379	0.759
Ngoko	Central	0.345	0.690	0.379	0.724	0.828
Ngoko	Eastern	0.276	0.552	0.103	0.379	0.552
Krama	Eastern	0.345	0.759	0.379	0.586	0.897

Table 4: Language identification accuracy based on different Javanese dialects and styles. Systems do not perform equally well across dialects and styles.

Colloquial Indonesian	Translation
Ada yang **ngetag** foto **lawas** di FB	Someone is tagging old photos in FB
Quotenya Andrew Ng ini relevan **banget**	This Andrew Ng quote is very relevant
Bilo kita pergi main lagi?	When will we go play again?
Ini **teh** aksara jawa kenapa susah **banget**?	Why is this Javanese script very difficult?

Table 5: Colloquial Indonesian code-mixing examples from social media. Color code: English, Betawinese, Javanese, Minangkabau, Sundanese, Indonesian.

Jambi Malay is not widely spoken (1M speakers), but has many dialects across villages. As shown in Table 2, many common words are spoken differently across dialects and styles. Similarly, Javanese is also different across regions. Not every Javanese speaker understands *Krama*, since its usage is very limited. Moreover, the number of Javanese speakers who can use *Krama* is declining (Cohn and Ravindranath, 2014).[5] Examples from other languages are shown in Appendix E.

Case Study in Javanese

Dialectical and style differences pose a challenge to NLP systems. To explore the extent of this challenge, we conduct an experiment to test the robustness of NLP systems to variations in Javanese dialects. We ask native speakers[6] to translate 29 simple sentences into Javanese according to the specified dialect and style. We then evaluate several language identification systems on those instances. Language identification is a core part of multilingual NLP and a necessary step for collecting textual data in a language. Despite its importance, it is an open research area, particularly for under-represented languages (Caswell et al., 2020).

We compare Langid.py (Lui and Baldwin, 2012), FastText (Joulin et al., 2017), and CLD3.[7] The results can be seen in Table 4. In general, the language identification systems are more accurate in detecting Javanese texts in the *Ngoko*-Central dialect, or *Krama*, since the systems were trained on Javanese Wikipedia data, which is written in either the *Ngoko*-Central or *Krama* dialects and styles. If an NLP system can only detect certain dialects, then this information should be conveyed explicitly. Problems arise if we assume that the model works equally well across dialects. For example, in the case of language identification, if we use the model to collect datasets automatically, then Javanese datasets with poor-performing dialects will be under represented in the data.

3.2.2 Code-Mixing

Code-mixing is an occurrence where a person speaks alternately in two or more languages in a conversation (Poplack, 1980; Winata et al., 2019, 2021a). This phenomenon is common in Indonesian conversations (Barik et al., 2019; Wibowo et al., 2020, 2021). In a conversational context, people sometimes mix their local languages with standard Indonesian, resulting in colloquial Indonesian (Siregar et al., 2014). This colloquial-style Indonesian is used daily in speech and conversation, and is common on social media (Sutrisno and Ariesta, 2019). Some frequently used code-mixed words (especially on social media) are even intelligible to people that do not speak the original local languages. Interestingly, code-mixing can also occur in border areas where people are exposed to multiple languages, therefore mixing them together. For example, people in Jember (a regency district in East Java) combine Javanese and Madurese in their daily conversation (Haryono, 2012).

Indonesian code-mixing not only occurs at the word level but also at the morpheme level (Winata, 2021). For example, *quotenya* ('his/her quote', see Table 5) combines the English word 'quote' and the Indonesian suffix *-nya*, which denotes possession; similarly, *ngetag* combines the Betawinese prefix *nge-* and the English word 'tag'. More examples can be found in Table 5.

[5] *Krama* is used to speak formally (e.g., with older or respected people). Nowadays, however, people prefer to use Indonesian more in formal situation. People who move from sub-urban areas to bigger cities tend to continue to use *Ngoko* and thus also pass *Ngoko* on to their children.

[6] Our annotators are based in Banyumas for Western Javanese, Jogjakarta for Central Javanese, and Jember for Eastern Javanese. Using dialects from different cities might result in a slightly different result.

[7] https://github.com/google/cld3

Language	Meaning	Written Variation	IPA
Javanese (Eastern–*Ngoko*)	what	apa / opo	/ɔpɔ/
	there is	ana / ono / onok	/ɔnɔʔ/
	you	kon / koen	/kɔn/
Balinese (Alus–*Singgih*)	yes	inggih / nggih	/ʔŋgih/
	I / me	tiang / tyang	/tiaŋ/
	<greeting>	swastyastu / swastiastu	/swastiastu/
Sundanese (Badui–*Loma*)	please / sorry	punten / punteun	/puntən/
	red	beureum / berem	/bərim/
	salivating	ngacai / ngacay	/ŋacaɪ/

Table 6: Written form variations in several local languages, confirmed by native speakers.

3.3 Orthography Variation

Many Indonesian local languages are mainly used in spoken settings and have no established standard orthography system. Some local languages do originally have their own archaic writing systems that derive from the Jawi alphabet or Kawi script, and even though standard transliteration into the Roman alphabet exists for some (e.g., Javanese and Sundanese), they are not widely known and practiced (Soeparno, 2015). Hence, some words have multiple romanized writings that are mutually intelligible by speakers, as they are pronounced the same. Some examples can be seen in Table 6. Such variety of the written form is common in many local languages in Indonesia. This variation leads to a significantly larger vocabulary size, especially for NLP systems that use word-based representations, and results in words being spelled differently despite referring to the same word, a challenge for subword-based models.

3.4 Societal Challenges

Language evolves together with the speakers. A more widely used language may have a larger digital presence, which fosters a more written form of communication while languages that are used only within small communities may emphasize the spoken form. There are also languages that are declining, where the speakers prefer to use Indonesian rather than their local language. In contrast, there are isolated residents that use the local language daily and are less proficient in Indonesian (Nurjanah et al., 2018; Jahang and Meirina, 2021). These variations give rise to different requirements and there is no single solution for all.

Technology and education is not well-distributed within the nation. Internet penetration in Indonesia is 73.7% in 2020, but is mainly concentrated on the Java island. Among the non-Internet users, 39% explain that they do not understand the technology, while 15% state that they do not have the device to access the internet.[8] In some areas where Internet is not seen as a basic need, imposing NLP technology on them may not necessarily be relevant. At the same time, general NLP development within the nation faces difficulties due to the lack of funding especially in universities outside of Java. GPU servers are still scarce, even on Java.[9]

The dynamics of population movement in Indonesia also need to be taken into consideration. For example, there are urban communities who transmigrate to remote areas for social purposes, such as teaching or becoming doctors for underdeveloped villages. Each of these situations might call for various new NLP technologies to be developed to facilitate better communication.

4 Opportunities

Based on the challenges for Indonesian NLP highlighted in the previous section, we formulate proposals for improving the state of Indonesian NLP research, as well as of other under-represented languages. Our proposals cover several aspects including metadata documentation; potential research directions; and engagement with communities.

4.1 Better Documentation

In line with studies promoting proper data documentation for NLP research (Bender and Friedman, 2018; Rogers et al., 2021; Alyafeai et al., 2021), we recommend the following considerations.

Regional Dialect Metadata We have shown that the same languages can have a large variation depending on region and dialect. Therefore, we suggest adding regional dialect metadata to NLP datasets and models, not only for Indonesian but for other languages as well. This is particularly important for languages with large dialectical differences. Regional dialect metadata is also important to clearly communicate NLP capabilities to stakeholders and end users as it will help set an expectation of what types of dialects systems can handle. Additionally, regional metadata can indirectly inform the topics of the data, especially for crawled data sources.

[8] The Indonesian Internet Providers Association (APJII) survey: https://apjii.or.id/survei2019x

[9] For instance, we estimate the whole computer science faculty of the nation's top university owns 8 V100 GPUs.

Style and Register Metadata Similarly, we also suggest adding style and register metadata. This metadata can capture the politeness level of the text, not only for Indonesian but also other languages. In addition, this metadata can be used to document the formality level of the text, so may be useful for research on modeling style or style transfer.

4.2 Potential Research Direction

In Indonesia, there are only few widely spoken languages that have been investigated in NLP, while the rest remain unstudied. Mitigating this limitation, we suggest future research to focus more on under-represented and unexplored languages.

Data-Efficient NLP Pretrained language models, which have taken NLP world by storm, require an abundant amount of monolingual data. However, data collection has been a long-standing problem for low-resource languages. Therefore, we recommend more exploration into designing data-efficient approaches such as adaptation methods (Artetxe et al., 2020; Aji et al., 2020; Gururangan et al., 2020; Koto et al., 2021), few-shot learning (Winata et al., 2021b; Madotto et al., 2021; Le Scao and Rush, 2021), and learning from related languages (Khanuja et al., 2021; Khemchandani et al., 2021). The goal of these methods is effective resource utilization, that is, to minimize the financial costs for computation and data collection as advocated by Schwartz et al. (2020), Cahyawijaya (2021), and Nityasya et al. (2021).

Data Generation Data collection efforts need to be commenced as soon as possible, despite all the challenges (§3.1). Here, we suggest collecting parallel data between Indonesian and each of the local languages due to several reasons. First, a lot of Indonesians are bilingual (Koto and Koto, 2020), that is, they speak both Indonesian and their local language, which facilitates data collection. Moreover, the fact that the local languages have some vocabulary overlap with Indonesian (See Table 7 in Appendix) might help building translation systems using relatively fewer parallel data (Nguyen and Chiang, 2017). Finally, having such parallel data, we can build translation systems for synthetic data generation. In line with this approach, the effectiveness of models trained on synthetic translated dataset can be explored.

Robustness to Code-mixing and Non-Standard Orthography Languages in Indonesia are prone to variations due to code-mixing and non-standard orthography, which occurs on the morpheme or even grapheme level. Models that are applied to Indonesian code-mixed data need to be able to learn morphologically faithful representations. Therefore, we recommend more exploration on methods derived from subword tokenization (Gage, 1994; Kudo, 2018) and token-free models (Gillick et al., 2016; Tay et al., 2021; Xue et al., 2021a) to deal with this problem.

NLP Beyond Text For many Indonesian local languages that are rarely if ever written, speech is a more natural communication format. We thus recommend more attention on less text-focused research, such as spoken language understanding (SLU) (Chung et al., 2021; Serdyuk et al., 2018), speech recognition (Besacier et al., 2014; Winata et al., 2020), and multimodality (Dai et al., 2020, 2021) in order to progress NLP in such languages.

4.3 Engage with Communities

As discussed in §3.4, it is difficult to generalize a solution across local languages. We thus encourage the NLP community to work more closely with native speakers and local communities (Nekoto et al., 2020). This is necessary to provide solutions and resources that support use cases benefiting the native speakers and communities of under-represented languages. We advise the involvement of linguists, for example to aid the language documentation process (Anastasopoulos et al., 2020). As GPU access can be a challenge for Indonesian research institutions, we suggest to engage with academic communities. We support open-science movements such as BigScience[10] or ICLR CoSubmitting Summer[11], which help to start collaborations and to reduce the entry barrier to NLP research.

5 Conclusion

In this paper, we have highlighted challenges in Indonesian NLP. Indonesia is one of the most populous country and the second-most linguistically diverse, with over 700 local languages, yet Indonesian NLP is under represented and under explored. Based on the observed challenges, we have also presented recommendations to improve the situation, not only for Indonesian, but for other under represented languages as well.

[10] https://bigscience.huggingface.co/

[11] https://blog.iclr.cc/2021/08/10/broadening-our-call-for-participation-to-iclr-2022/

References

Husen Abas. 1987. *Indonesian as a unifying language of wider communication: a historical and sociolinguistic perspective*. Number 73 in Pacific Linguistics Series D. Canberra.

Alexander Adelaar. 2005. The Austronesian languages of Asia and Madagascar: A historical perspective. In Alexander Adelaar and Nikolaus P. Himmelmann, editors, *The Austronesian Languages of Asia and Madagascar*, chapter 1, pages 1–42. Routledge, Oxon.

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Alham Fikri Aji, Nikolay Bogoychev, Kenneth Heafield, and Rico Sennrich. 2020. In neural machine translation, what does transfer learning transfer? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7701–7710.

Ika Alfina, Ruli Manurung, and Mohamad Ivan Fanany. 2016. DBpedia entities expansion in automatically building dataset for Indonesian NER. In *2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 335–340.

Ika Alfina, Rio Mulia, Mohamad Ivan Fanany, and Yudo Ekanata. 2017. Hate speech detection in the indonesian language: A dataset and preliminary study. In *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 233–238. IEEE.

Zaid Alyafeai, Maraim Masoud, Mustafa Ghaleb, and Maged S Al-shaibani. 2021. Masader: Metadata sourcing for arabic text and speech data resources. *arXiv preprint arXiv:2110.06744*.

Antonios Anastasopoulos, Christopher Cox, Graham Neubig, and Hilaria Cruz. 2020. Endangered languages meet Modern NLP. In *Proceedings of the 28th International Conference on Computational Linguistics: Tutorial Abstracts*, pages 39–45, Barcelona, Spain (Online). International Committee for Computational Linguistics.

Karl Ronald Anderbeck. 2008. *Malay dialects of the Batanghari river basin (Jambi, Sumatra)*. SIL International.

Anisya O. Anindyatri and Imarotul Mufidah. 2020. *Gambaran Kondisi Vitalitas Bahasa Daerah di Indonesia*. Kementrian Pendidikan dan Kebudayaan Pusat Data dan Teknologi Informasi, Tangerang Selatan, Indonesia.

Valentina Kania Prameswara Artari, Rahmad Mahendra, Meganingrum Arista Jiwanggi, Adityo Anggraito, and Indra Budi. 2021. Multi-pass sieve coreference resolution for Indonesian. In *Proceedings of Recent Advances in Natural Language Processing*, pages 84–90.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637.

Jessica Naraiswari Arwidarasti, Ika Alfina, and Adila Alfa Krisnadhi. 2019. Converting an Indonesian constituency treebank to the Penn treebank format. In *2019 International Conference on Asian Language Processing (IALP)*, pages 331–336.

Annisa Nurul Azhar, Masayu Leylia Khodra, and Arie Pratama Sutiono. 2019. Multi-label aspect categorization with convolutional neural networks and extreme gradient boosting. In *2019 International Conference on Electrical Engineering and Informatics (ICEEI)*, pages 35–40. IEEE.

Kurniawati Azizah, Mirna Adriani, and Wisnu Jatmiko. 2020. Hierarchical transfer learning for multilingual, multi-speaker, and style transfer dnn-based tts on low-resource languages. *IEEE Access*, 8:179798–179812.

Anab Maulana Barik, Rahmad Mahendra, and Mirna Adriani. 2019. Normalization of indonesian-english code-mixed twitter data. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 417–424.

Sadar Baskoro and Mirna Adriani. 2008. Developing an indonesian speech recognition system. In *Second MALINDO Workshop. Selangor, Malaysia*.

Peter Bellwood. 1997. *Prehistory of the Indo-Malaysian Archipelago*. University of Hawaii Press, Honolulu.

Peter Bellwood, Geoffrey Chambers, Malcolm Ross, and Hsiao-chun Hung. 2011. Are 'cultures' inherited? Multidisciplinary perspectives on the origins and migrations of Austronesian-speaking peoples prior to 1000 BC. In *Investigating archaeological cultures*, pages 321–354. Springer.

Peter Bellwood and Eusebio Dizon. 2008. Austronesian cultural origins : out of Taiwan, via the Batanes Islands, and onwards to western Polynesia. In Alicia Sanchez-Mazas, editor, *Past human migrations in East Asia: matching archaeology, linguistics and genetics*. Routledge, London.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Jacques Bertrand. 2004. *Nationalism and ethnic conflict in Indonesia*. Cambridge University Press.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech communication*, 56:85–100.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Robert Blust. 1980. Austronesian etymologies. *Oceanic Linguistics*, 19:1–181.

Francis Bond and Timothy Baldwin. 2016. *Introduction to Japanese Computational Linguistics*.

Luiz Henrique Bonifacio, Israel Campiotti, Roberto Lotufo, and Rodrigo Nogueira. 2021. mmarco: A multilingual version of ms marco passage ranking dataset.

Indra Budi, Stéphane Bressan, Gatot Wahyudi, Zainal A. Hasibuan, and Bobby A. A. Nazief. 2005. Named entity recognition for the Indonesian language: Combining contextual, morphological and part-of-speech features into a knowledge engineering approach. In *Discovery Science*, pages 57–69, Berlin, Heidelberg. Springer Berlin Heidelberg.

Samuel Cahyawijaya. 2021. Greenformers: Improving computation and memory efficiency in transformer models via low-rank approximation.

Samuel Cahyawijaya, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kuncoro, Sebastian Ruder, Zhi Yuan Lim, Syafri Bahar, Masayu Leylia Khodra, et al. 2021. Indonlg: Benchmark and resources for evaluating Indonesian natural language generation. *arXiv preprint arXiv:2104.08200*.

Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. 2020. Language ID in the wild: Unexpected challenges on the path to a thousand-language web text corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6588–6608, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Isaac Caswell, Julia Kreutzer, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. 2021. Quality at a glance: An audit of web-crawled multilingual datasets. *arXiv preprint arXiv:2103.12028*.

Yu-An Chung, Chenguang Zhu, and Michael Zeng. 2021. SPLAT: Speech-language joint pre-training for spoken language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1897–1907, Online. Association for Computational Linguistics.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Abigail C Cohn and Maya Ravindranath. 2014. Local languages in indonesia: Language maintenance or language shift. *Linguistik Indonesia*, 32(2):131–148.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexander R. Coupe and František Kratochvíl. 2020. *Asia before English*, pages 15–48. John Wiley & Sons, Inc.

Wenliang Dai, Samuel Cahyawijaya, Zihan Liu, and Pascale Fung. 2021. Multimodal end-to-end sparse model for emotion recognition. In *NAACL*.

Wenliang Dai, Zihan Liu, Tiezheng Yu, and Pascale Fung. 2020. Modality-transferable emotion embeddings for low-resource multimodal emotion recognition. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 269–280, Suzhou, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Arawinda Dinakaramani, Fam Rashel, Andry Luthfi, and Ruli Manurung. 2014. Designing an Indonesian part of speech tagset and manually tagged Indonesian corpus. In *2014 International Conference on Asian Language Processing (IALP)*, pages 66–69.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2021. *Ethnologue: Languages of the World. Twenty-fourth edition.* Dallas, Texas: SIL International.

European Language Resources Association. 2019. BLT4All: Language Technologies for All. https://lt4all.elra.info/en/. [Online; accessed Dec. 2019.].

Muhammad Fachri. 2014. Named entity recognition for Indonesian text using hidden markov model. Undergraduate Thesis.

Andri Imam Fauzi and Dwi Puspitorini. 2018. Dialect and identity: A case study of javanese use in whatsapp and line. In *IOP Conference Series: Earth and Environmental Science*, volume 175, page 012111. IOP Publishing.

Abdurrisyad Fikri and Ayu Purwarianti. 2012. Case based Indonesian closed domain question answering system with real world questions. In *2012 7th International Conference on Telecommunication Systems, Services, and Applications (TSSA)*, pages 181–186.

Philip Gage. 1994. A new algorithm for data compression. *C Users J.*, 12(2):23–38.

Dan Gillick, Cliff Brunk, Oriol Vinyals, and Amarnag Subramanya. 2016. Multilingual language processing from bytes. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1296–1306, San Diego, California. Association for Computational Linguistics.

Dirk Goldhahn, Thomas Eckart, and 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)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

Russell D. Gray and Fiona M. Jordan. 2000. Language trees support the express-train sequence of Austronesian expansion. *Nature*, 405:1052–1055.

Yohanes Gultom and Wahyu Catur Wibowo. 2017. Automatic open domain information extraction from Indonesian text. In *2017 International Workshop on Big Data and Information Security (IWBIS)*, pages 23–30. IEEE.

William Gunawan, Derwin Suhartono, Fredy Purnomo, and Andrew Ongko. 2018. Named-entity recognition for indonesian language using bidirectional lstm-cnns. *Procedia Computer Science*, 135:425–432.

Tri Wahyu Guntara, Alham Fikri Aji, and Radityo Eko Prasojo. 2020. Benchmarking multidomain English-Indonesian machine translation. In *Proceedings of the 13th Workshop on Building and Using Comparable Corpora*, pages 35–43, Marseille, France. European Language Resources Association.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

Akhmad Haryono. 2012. *Perubahan dan perkembangan bahasa: Tinjauan historis dan sosiolinguistik*. Ph.D. thesis, Udayana University.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Devin Hoesen and Ayu Purwarianti. 2018. Investigating bi-lstm and crf with pos tag embedding for Indonesian named entity tagger. In *2018 International Conference on Asian Language Processing (IALP)*, pages 35–38. IEEE.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization. In *Proceedings of ICML 2020*.

Muhammad Okky Ibrohim and Indra Budi. 2019. Multi-label hate speech and abusive language detection in Indonesian Twitter. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 46–57, Florence, Italy. Association for Computational Linguistics.

Arfinda Ilmania, Abdurrahman, Samuel Cahyawijaya, and Ayu Purwarianti. 2018. Aspect detection and sentiment classification using deep neural network for Indonesian aspect-based sentiment analysis. In *2018 International Conference on Asian Language Processing (IALP)*, pages 62–67.

Benediktus Sridin Sulu Jahang and Zita Meirina. 2021. 1,3 juta anak di ntt belum bisa berbahasa indonesia. Last accessed on 05/10/2021.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of ACL 2020*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics.

Erlin Kartikasari, Kisyani Laksono, Dian Savitri Agusniar, and Diah Yovita Suryarini. 2018. A study of dialectology on Javanese "Ngoko" in Banyuwangi, Surabaya, Magetan, and Solo. *Humaniora*, 30(2):128.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. MuRIL: Multilingual Representations for Indian Languages. *arXiv preprint arXiv:2103.10730*.

Yash Khemchandani, Sarvesh Mehtani, Vaidehi Patil, Abhijeet Awasthi, Partha Talukdar, and Sunita Sarawagi. 2021. Exploiting Language Relatedness for Low Web-Resource Language Model Adaptation: An Indic Languages Study. In *Proceedings of ACL 2021*.

Marian Klamer. 2018. Documenting the linguistic diversity of indonesia: Time is running out. In *Proceedings of 'Revitalization of local languages as the pillar of pluralism'*, pages 1–10. APBL (Asosiasi Peneliti Bahasa-bahasa Lokal) and Nusa Cendana University, Kupang, Satya Wacana Press.

Fajri Koto and Ikhwan Koto. 2020. Towards computational linguistics in Minangkabau language: Studies on sentiment analysis and machine translation. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 138–148, Hanoi, Vietnam. Association for Computational Linguistics.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2020a. Liputan6: A large-scale Indonesian dataset for text summarization. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 598–608, Suzhou, China. Association for Computational Linguistics.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. IndoBERTweet: A pretrained language model for Indonesian twitter with effective domain-specific vocabulary initialization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. 2020b. IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 757–770, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Kemal Kurniawan and Alham Fikri Aji. 2018. Toward a standardized and more accurate Indonesian part-of-speech tagging. In *2018 International Conference on Asian Language Processing (IALP)*, pages 303–307. IEEE.

Kemal Kurniawan and Samuel Louvan. 2018. Indosum: A new benchmark dataset for Indonesian text summarization. In *2018 International Conference on Asian Language Processing (IALP)*, pages 215–220. IEEE.

Teven Le Scao and Alexander M Rush. 2021. How many data points is a prompt worth? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2627–2636.

Dessi Puji Lestari, Koji Iwano, and Sadaoki Furui. 2006. A large vocabulary continuous speech recognition system for indonesian language. In *15th Indonesian Scientific Conference in Japan Proceedings*, pages 17–22.

Zhaojiang Lin, Zihan Liu, Genta Indra Winata, Samuel Cahyawijaya, Andrea Madotto, Yejin Bang, Etsuko Ishii, and Pascale Fung. 2021. XPersona: Evaluating multilingual personalized chatbot. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 102–112, Online. Association for Computational Linguistics.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Marco Lui and Timothy Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30.

Edwin Lunando and Ayu Purwarianti. 2013. Indonesian social media sentiment analysis with sarcasm detection. In *2013 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 195–198.

Andrea Madotto, Zhaojiang Lin, Genta Indra Winata, and Pascale Fung. 2021. Few-shot bot: Prompt-based learning for dialogue systems. *arXiv preprint arXiv:2110.08118*.

Putu Devi Maharani and Komang Dian Puspita Candra. 2018. Variasi leksikal bahasa Bali dialek kuta selatan. *Mudra Jurnal Seni Budaya*, 33(1):76–84.

Rahmad Mahendra, Alham Fikri Aji, Samuel Louvan, Fahrurrozi Rahman, and Clara Vania. 2021. IndoNLI: A natural language inference dataset for indonesian. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Rahmad Mahendra, Septina Dian Larasati, and Ruli Manurung. 2008. Extending an Indonesian semantic analysis-based question answering system with linguistic and world knowledge axioms. In *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation*, pages 262–271, The University of the Philippines Visayas Cebu College, Cebu City, Philippines. De La Salle University, Manila, Philippines.

Rahmad Mahendra, Heninggar Septiantri, Haryo Akbarianto Wibowo, Ruli Manurung, and Mirna Adriani. 2018. Cross-lingual and supervised learning approach for Indonesian word sense disambiguation task. In *Proceedings of the 9th Global Wordnet Conference*, pages 245–250, Nanyang Technological University (NTU), Singapore. Global Wordnet Association.

Miftahul Mahfuzh, Sidik Soleman, and Ayu Purwarianti. 2019. Improving joint layer rnn based keyphrase extraction by using syntactical features. In *2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, pages 1–6. IEEE.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics.

Aqsath Rasyid Naradhipa and Ayu Purwarianti. 2011. Sentiment classification for Indonesian message in social media. In *Proceedings of the 2011 International Conference on Electrical Engineering and Informatics*, pages 1–4.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Toan Q Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301.

Made Nindyatama Nityasya, Haryo Akbarianto Wibowo, Radityo Eko Prasojo, and Alham Fikri Aji. 2021. Costs to consider in adopting nlp for your business.

Sashi Novitasari, Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2020. Cross-lingual machine speech chain for Javanese, Sundanese, Balinese, and Bataks speech recognition and synthesis. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 131–138, Marseille, France. European Language Resources association.

Fajrin Nurjanah et al. 2018. Pengembangan kemampuan berbahasa indonesia siswa sekolah dasar desa terpencil melalui metode karyawisata berbasis potensi lokal. *FKIP e-PROCEEDING*, pages 167–176.

Bill Palmer. 2018. *The languages and linguistics of the New Guinea area: A comprehensive guide*, volume 4. Walter de Gruyter GmbH.

Valantino Ateng Pamolango. 2012. Geografi dialek bahasa Saluan. *PARAFRASE: Jurnal Kajian Kebahasaan & Kesastraan*, 12(02).

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin O. Steitz, Stefan Roth, Ivan Vulic, and Iryna

Gurevych. 2021. xgqa: Cross-lingual visual question answering. *ArXiv*, abs/2109.06082.

Femphy Pisceldo, Rahmad Mahendra, Ruli Manurung, and I Wayan Arka. 2008. A two-level morphological analyser for the indonesian language. In *Proceedings of the Australasian Language Technology Association Workshop 2008*, pages 142–150.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Shana Poplack. 1980. Sometimes i'll start a sentence in spanish y termino en espanol: toward a typology of code-switching1.

Apriyani Purwaningsih. 2017. Geografi dialek bahasa jawa pesisiran di desa paciran kabupaten lamongan. In *Proceeding of International Conference on Art, Language, and Culture*, pages 594–605.

Ayu Purwarianti and Ida Ayu Putu Ari Crisdayanti. 2019. Improving bi-lstm performance for Indonesian sentiment analysis using paragraph vector. In *2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, pages 1–5. IEEE.

Ayu Purwarianti, Masatoshi Tsuchiya, and Seiichi Nakagawa. 2007. A machine learning approach for Indonesian question answering system. In *Artificial Intelligence and Applications*, pages 573–578.

Shofianina Dwi Ananda Putri, Muhammad Okky Ibrohim, and Indra Budi. 2021. Abusive language and hate speech detection for indonesian-local language in social media text. In *Recent Advances in Information and Communication Technology 2021*, pages 88–98, Cham. Springer International Publishing.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Anna Rogers, Tim Baldwin, and Kobi Leins. 2021. Just what do you think you're doing, dave?'a checklist for responsible data use in nlp. *arXiv preprint arXiv:2109.06598*.

Malcolm Ross. 2005. Pronouns as a preliminary diagnostic for grouping Papuan languages. *Papuan pasts: Cultural, linguistic and biological histories of Papuan-speaking peoples*, 572:15–65.

Nur Endah Safitri, Amalia Zahra, and Mirna Adriani. 2016. Spoken language identification with phonotactics methods on minangkabau, sundanese, and javanese languages. *Procedia Computer Science*, 81:182–187.

Mei Silviana Saputri, Rahmad Mahendra, and Mirna Adriani. 2018. Emotion classification on Indonesian twitter dataset. In *2018 International Conference on Asian Language Processing (IALP)*, pages 90–95. IEEE.

Gita Sarwadi, Mahsun Mahsun, and Burhanuddin Burhanuddin. 2019. Lexical variation of Sasak Kuto-Kute dialect in North Lombok district. *Jurnal Kata: Penelitian tentang Ilmu Bahasa dan Sastra*, 3(1):155–169.

Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. 2020. Green ai. *Communications of the ACM*, 63(12):54–63.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Dmitriy Serdyuk, Yongqiang Wang, Christian Fuegen, Anuj Kumar, Baiyang Liu, and Yoshua Bengio. 2018. Towards end-to-end spoken language understanding. IEEE.

Ken Nabila Setya and Rahmad Mahendra. 2018. Semi-supervised textual entailment on Indonesian wikipedia data. In *2018 International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*.

Masitowarni Siregar, Syamsul Bahri, Dedi Sanjaya, et al. 2014. Code switching and code mixing in indonesia: Study in sociolinguistics. *English Language and Literature Studies*, 4(1):77–92.

James Neil Sneddon. 2003. *The Indonesian language: Its history and role in modern society*. UNSW Press, Sydney.

Soeparno. 2015. Kerancuan fono-ortografis dan ortofonologis bahasa Indonesia ragam lisan dan tulis. *Diksi*, 12(2).

Hein Steinhauer. 2005. Colonial history and language policy in insular southeast asia and madagascar. In Alexander Adelaar and Nikolaus P. Himmelmann, editors, *The Austronesian Languages of Asia and Madagascar*, chapter 3, pages 65–86. Routledge, Oxon.

Arie Ardiyanti Suryani, Dwi Hendratmo Widyantoro, Ayu Purwarianti, and Yayat Sudaryat. 2015. Experiment on a phrase-based statistical machine translation using pos tag information for sundanese into indonesian. In *2015 International Conference on Information Technology Systems and Innovation (ICITSI)*, pages 1–6.

Arie Ardiyanti Suryani, Dwi Hendratmo Widyantoro, Ayu Purwarianti, and Yayat Sudaryat. 2018. The rule-based sundanese stemmer. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 17(4).

Taufic Leonardo Sutejo and Dessi Puji Lestari. 2018. Indonesia hate speech detection using deep learning. In *2018 International Conference on Asian Language Processing (IALP)*, pages 39–43. IEEE.

Bejo Sutrisno and Yessika Ariesta. 2019. Beyond the use of code mixing by social media influencers in instagram. *Advances in Language and Literary Studies*, 10(6):143–151.

Yi Tay, Vinh Q Tran, Sebastian Ruder, Jai Gupta, Hyung Won Chung, Dara Bahri, Zhen Qin, Simon Baumgartner, Cong Yu, and Donald Metzler. 2021. Charformer: Fast character transformers via gradient-based subword tokenization. *arXiv preprint arXiv:2106.12672*.

Jozina Vander Klok. 2015. The dichotomy of auxiliaries in javanese: Evidence from two dialects. *Australian Journal of Linguistics*, 35(2):142–167.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Haryo Akbarianto Wibowo, Made Nindyatama Nityasya, Afra Feyza Akyürek, Suci Fitriany, Alham Fikri Aji, Radityo Eko Prasojo, and Derry Tanti Wijaya. 2021. IndoCollex: A testbed for morphological transformation of Indonesian word colloquialism. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3170–3183, Online. Association for Computational Linguistics.

Haryo Akbarianto Wibowo, Tatag Aziz Prawiro, Muhammad Ihsan, Alham Fikri Aji, Radityo Eko Prasojo, Rahmad Mahendra, and Suci Fitriany. 2020. Semi-supervised low-resource style transfer of Indonesian informal to formal language with iterative forward-translation. In *2020 International Conference on Asian Language Processing (IALP)*, pages 310–315. IEEE.

Alfan Farizki Wicaksono and Ayu Purwarianti. 2010. Hmm based part-of-speech tagger for bahasa indonesia. In *4th International MALINDO (Malaysian-Indonesian Language) Workshop*.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, et al. 2020. Indonlu: Benchmark and resources for evaluating Indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 843–857.

Genta Indra Winata. 2021. Multilingual transfer learning for code-switched language and speech neural modeling. *arXiv preprint arXiv:2104.06268*.

Genta Indra Winata, Samuel Cahyawijaya, Zhaojiang Lin, Zihan Liu, and Pascale Fung. 2020. Lightweight and efficient end-to-end speech recognition using low-rank transformer. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6144–6148.

Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2021a. Are multilingual models effective in code-switching? In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 142–153.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021b. Language models are few-shot multilingual learners. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 1–15.

Genta Indra Winata, Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2019. Code-switched language models using neural based synthetic data from parallel sentences. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 271–280.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2021a. Byt5: Towards a token-free future with pre-trained byte-to-byte models.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021b. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Evi Yulianti, Indra Budi, Achmad N. Hidayanto, Hisar M. Manurung, and Mirna Adriani. 2011. Developing Indonesian-English hybrid machine translation system. In *2011 International Conference on Advanced Computer Science and Information Systems*, pages 265–270.

Amalia Zahra, Sadar Baskoro, and Mirna Adriani. 2009. Building a pronunciation dictionary for indonesian speech recognition system. In *The TCAST workshop, Singapore*.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared*

Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.

A Language Statistics

In Figure 4, we contrast the distribution of publication with Indonesian language compared to European languages. Despite number of Indonesian speakers is much larger compared to some European languages, number of published research in Indonesian is still comparatively lower.

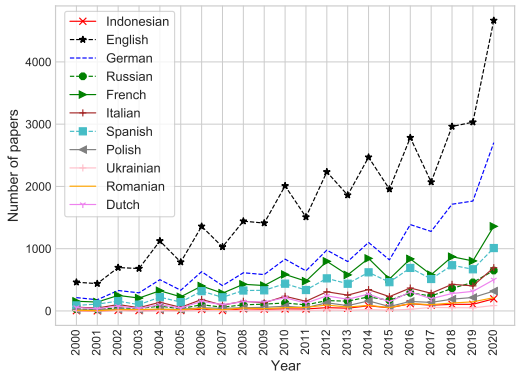

Figure 4: Number of published research works per year in Indonesian and European languages from 2000 to 2020.

B Wikipedia Vocabulary Overlap

Lang	# Vocab		
	All (k)	Top 1% ∩ KBBI (%)	Top 100 ∩ KBBI (%)
id	2023	59.3	**96**
jav	435	46.8	43
su	286	44.3	47
min	252	30.3	41
bug	23	35.7	27
map-bms	14	76.7	79
gor	12	40.5	49
ace	12	37.6	46
ban	10	43.3	46
bjn	4	62.9	69
nia	1	25.9	30
mad	1	26.9	24

Table 7: Vocabulary of Indonesian languages in Wikipedia, filtered with KBBI third edition [12]

In Table 7, we present vocabulary statistics of Indonesian languages in Wikipedia. Due to the noisy nature of Wikipedia, we use "*Kamus Besar Bahasa Indonesia*" (KBBI) third edition,[13] the official dictionary for the Indonesian language to filter the top 1% and top-100 most frequent words. As expected, the top 1% words are less reliable, with only 59.3% of the vocabulary overlap between `id` and KBBI. In the top-100 words, there is a 96% word overlap with KBBI, making this set more reliable. Previous work on Minangkabau by Koto and Koto (2020) also showed that `id-min` words have a 55% overlap in a manually curated bilingual dictionary, closer to the top-100 value for `min` in Table 7.

[12] KBBI is the official Indonesian dictionary.

[13] https://github.com/geovedi/indonesian-wordlist

C Indonesian NLP Resources

On Table 8, we list statistics of Indonesian language corpora for different tasks, including sentiment analysis, part-of-speech tagging, summarization, NLI, and discourse. Although most datasets are in Indonesian and only a few are in Minangkabau (Koto and Koto, 2020), Javanese and Sundanese (Cahyawijaya et al., 2021), these resource collections are arguably beneficial for constructing resources in other local languages. This is because 1) Indonesian can be used as a pivot language with regard to local languages due to the large vocabulary overlap (see Table 7), and 2) most Indonesians are bilingual, speaking both Indonesian and their local language (Koto and Koto, 2020).

D Wikipedia Availability

In Figure 5 we compare Wikipedia size (in GB file size) compared to the number of speakers across various languages. We show that some African languages are even more under-resourced.

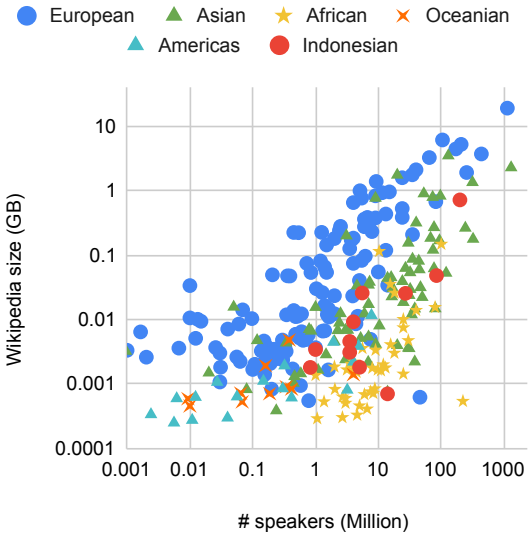

Figure 5: Wikipedia data size (in GB) compared to the number of speakers.

Name	Task Type	Size	% Indo	% Local
	Labelled Datasets			
POSP (Hoesen and Purwarianti, 2018)	PoS Tagging	8k	100%	0%
BaPOS (Dinakaramani et al., 2014)	PoS Tagging	10k	100%	0%
NERGrit (Wilie et al., 2020)	Named Entity Recognition	2k	100%	0%
NERP (Hoesen and Purwarianti, 2018)	Named Entity Recognition	8k	100%	0%
(Gultom and Wibowo, 2017)	Named Entity Recognition	2k	100%	0%
Singgalang (Alfina et al., 2016)	Named Entity Recognition	48K	100%	0%
(Fachri, 2014)	Named Entity Recognition	2K	100%	0%
KEPS (Mahfuzh et al., 2019)	Keyphrase Extraction	1k	100%	0%
FacQA (Purwarianti et al., 2007)	Question Answering	3k	100%	0%
WReTE (Setya and Mahendra, 2018)	Natural Language Inference	0.5k	100%	0%
IndoNLI (Mahendra et al., 2021)	Natural Language Inference	18k	100%	0%
CASA (Ilmania et al., 2018)	Sentiment Analysis	1k	100%	0%
SmSA (Purwarianti and Crisdayanti, 2019)	Sentiment Analysis	13k	100%	0%
HoASA (Azhar et al., 2019)	Sentiment Analysis	3k	100%	0%
SA in IndoLEM (Koto et al., 2020b)	Sentiment Analysis	5k	100%	0%
SA in MinangNLP (Koto and Koto, 2020)	Sentiment Analysis	5K	0%	100%
(Saputri et al., 2018)	Emotion Classification	4k	100%	0%
(Ibrohim and Budi, 2019)	Hate Speech Detection	13k	100%	0%
TED En-Id (Guntara et al., 2020)	Machine Translation	93k	100%	0%
News En-Id (Guntara et al., 2020)	Machine Translation	42k	100%	0%
Religion En-Id (Guntara et al., 2020)	Machine Translation	590k	100%	0%
MT in MinangNLP (Koto and Koto, 2020)	Machine Translation	11K	0%	100%
EN↔ID MT (Cahyawijaya et al., 2021)	Machine Translation	31K	100%	0%
SU↔ID MT (Cahyawijaya et al., 2021)	Machine Translation	16K	0%	100%
JV↔ID MT (Cahyawijaya et al., 2021)	Machine Translation	16K	0%	100%
IndoSum (Kurniawan and Louvan, 2018)	Summarization	20k	100%	0%
Liputan6 (Koto et al., 2020a)	Summarization	215k	100%	0%
Kethu (Arwidarasti et al., 2019)	Constituency Parsing	1k	100%	0%
UD-Id GSD (McDonald et al., 2013)	Dependency Parsing	5k	100%	0%
UD-Id PUD (Zeman et al., 2018)	Dependency Parsing	1k	100%	0%
(Mahendra et al., 2018)	Word Sense Disambiguation	2k	100%	0%
IndoCoref (Artari et al., 2021)	Coreference Resolution	0.2k	100%	0%
NTP and Tweet Ordering (Koto et al., 2020b)	Discourse	7k	100%	0%
	Pretraining Corpora			
Indo4B (Wilie et al., 2020)	-	3.6B words	100%	0%
Indo4B-Plus (Cahyawijaya et al., 2021)	-	4.0B words	89.64%	10.36%

Table 8: Statistics of publicly available datasets, most datasets are covered on the existing Indonesian languages NLP benchmarks (Wilie et al., 2020; Koto et al., 2020b; Cahyawijaya et al., 2021).

E Dialect Differences

In this section, we present more examples of lexical variation of other local languages. Maharani and Candra (2018) and Sarwadi et al. (2019) show lexical variation of Balinese and Sasak, respectively, where they ask locals to translate general/common words. Then, they compare the vocabulary across different locations (in this case, villages) to each other. Some of the examples can be seen in Table 9 and 10. Unfortunately, they did not provide quantitative results. Pamolango (2012) conducted a similar experiment in the Banggai district in South Sulawesi across 31 observation points for the Saluan language. While Pamolango (2012) did not provide full examples, they reported up to 23.5% lexical variation among 200 basic vocabulary items.

F Local Language Classification

As shown in Table 11, some of the Javanese texts are misidentified as Indonesian, English, and Malaysian. Javanese and Indonesian (which is similar to Malasyian) share some words. We believe English mis-classification is due to the data size bias.

English	Kedonganan	Jimbaran	Unggasan
I/me	Tyang	Tyang	Aku
You	Béné	Béné	Éngko
Umbrella	Pajéng	Pajéng	Pajong
Hat	Capil	Topong	Cecapil, Tetopong
How	Engken	Engken	Kengen
Where	Dijé	Dijé	Di joho
All	Konyangan	Onyé	Konyangan, onyang
Swallow (vb)	Gélék, ngélék	Gélék, ngélék	Ngélokang
Scratch (vb)	Gagas	Gagas	Gauk
Cough (vb)	Kokoan	Dékah	Kohkohan
Dawn	Plimunan	Plimunan	Sémongan
Afternoon	Sanjé	Sanjé	Sanjano

Table 9: Lexical variation of Balinese across different villages in South Kuta district, Bali (Maharani and Candra, 2018)

English	Pemenang Timur	Jenggala	Genggelang	Kayangan	Akar-Akar
Here	Ite	ite	ite	ite	tinI
There	Ito	ito	ito	ito	tinO
You	diʔ	sita	diʔ	sita	diʔ
Husband	kurənan	sawa	sawa	sawa	sawa
No	deʔ	deʔ	deʔ	deʔ	soraʔ
Paddle	bose	bose	dayung	dayung	bose
Spear	tɜr	cinəkan	tɜr	tombak	tombak
Black	birəŋ	birəŋ	birəŋ	birəŋ	pisak
Red	bənəŋ	bənəŋ	bənəŋ	bənəŋ	abaŋ
White	putɜʔ	putɜʔ	putɜʔ	putɜʔ	pətak
Worm	gumbər	loŋa	gumbər	gumbər	gumbər

Table 10: Lexical variation of Sasak across different villages in North Lombok district (Sarwadi et al., 2019)

Dialect/ Style	Method	classified as jv	id	en	ms
Western-*Ngoko*	Langid	0.241	0.103	0.172	0.069
	FastText	0.069	0.276	0.276	0.069
	CLD3	0.759	0.000	0.000	0.034
Central-*Ngoko*	Langid	0.345	0.138	0.069	0.069
	FastText	0.379	0.310	0.069	0.069
	CLD3	0.828	0.000	0.000	0.034
Eastern-*Ngoko*	Langid	0.276	0.103	0.069	0.138
	FastText	0.103	0.310	0.103	0.034
	CLD3	0.552	0.103	0.000	0.000
Eastern-*Krama*	Langid	0.345	0.241	0.034	0.172
	FastText	0.379	0.310	0.069	0.034
	CLD3	0.897	0.000	0.000	0.000

Table 11: Language identification mis-classification rate.