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

Alham Fikri Aji[1*†], **Genta Indra Winata**[2*], **Fajri Koto**[3*], **Samuel Cahyawijaya**[4*], **Ade Romadhony**[5,7*], **Rahmad Mahendra**[6,7*], **Kemal Kurniawan**[3,7], **David Moeljadi**[8], **Radityo Eko Prasojo**[9], **Timothy Baldwin**[3,10], **Jey Han Lau**[3], **Sebastian Ruder**[11]

[1]Amazon [2]Bloomberg [3]The University of Melbourne [4]HKUST
[5]Telkom University [6]Universitas Indonesia [7]INACL
[8]Kanda University of International Studies [9]Kata.ai [10]MBZUAI [11]Google Research

afaji@amazon.co.uk, gwinata@bloomberg.net, fajri.phd@gmail.com, rahmad.mahendra@cs.ui.ac.id

Abstract

NLP research is impeded by a lack of resources and awareness of the challenges presented by underrepresented 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 such a 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).

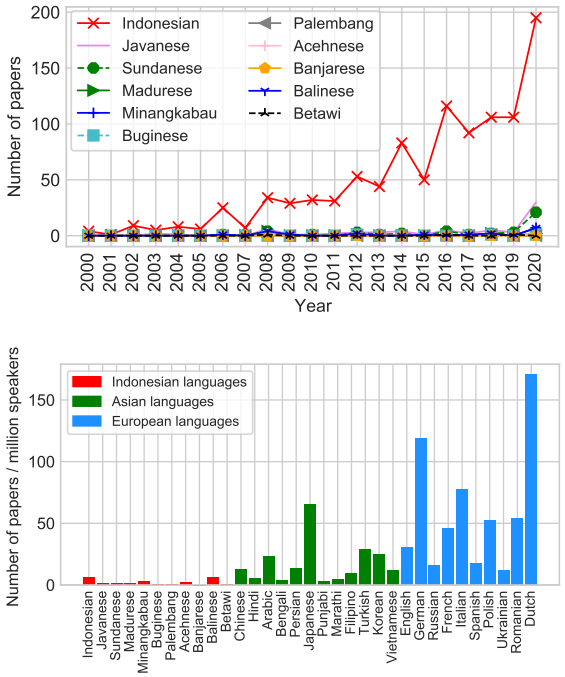

Figure 1: Following Joshi et al. (2020), we compile ACL Anthology to count the distribution of published works that mention languages spoken in Indonesia. **Top:** Distribution of papers in 20 years. **Bottom:** Number of papers per a million speakers. We compare languages spoken in Europe, Asia, and Indonesia.

Among more than 700 Indonesian local languages, many are threatened. 440 languages are listed as endangered and 12 as extinct according to data from Ethnologue (Eberhard et al., 2021) illustrated in Figure 2. Anindyatri and Mufidah (2020) found nearly half of a sample of 98 Indonesian local languages to be endangered while van Esch et al. (2022) observed 71 among 151 Indonesian local languages to have less than 100k speakers.

[*] These authors contributed equally.
[†] Work done prior to joining Amazon.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 7226 - 7249

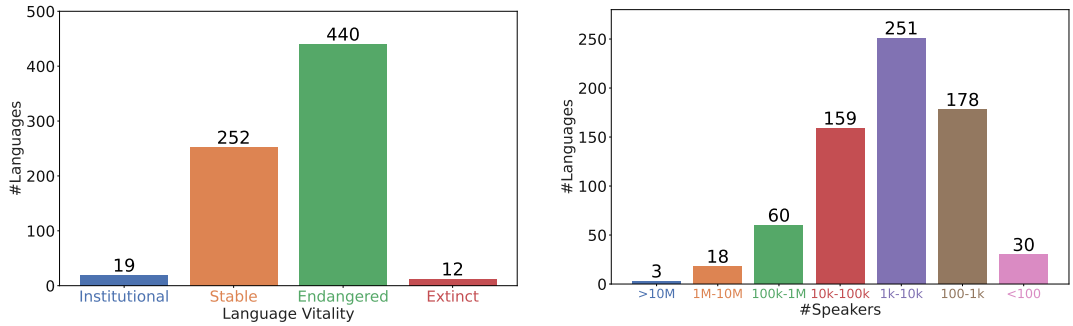

Figure 2: Distribution of 700+ languages spoken in Indonesia according to Ethnologue (Eberhard et al., 2021). **Left:** Language vitality. **Right:** Speaker count.

Language	ISO	# Speakers
Indonesian	ind	198 M
Javanese	jav	84 M
Sundanese / Sunda	sun	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 number of speakers for Indonesian and top-10 most spoken local languages in Indonesia (Eberhard et al., 2021).

Table 1 lists the names of the 10 most spoken local languages in Indonesia (Eberhard et al., 2021). Javanese and Sundanese are at the top with 84M and 34M speakers, respectively, while Madura, Minangkabau, and Buginese each 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 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 a 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. Language 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 and Indonesia's hundreds of 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 the languages in Indonesia but also for other underrepresented languages.

2 Background and Related Work

2.1 History and Taxonomy

Indonesia is one of the richest countries globally 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 3, 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. Lan-

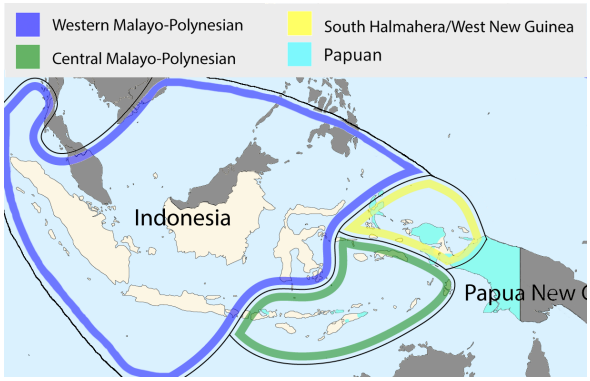

Figure 3: Map of Austronesian and Papuan languages in Indonesia.

guages 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, and further-flung regions such as the Mamberamo River and the Raja Ampat Islands. Meanwhile, the Papuan languages 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 et al., 2011). In the second millennium CE, the Austronesian people moved from Taiwan to the Philippines. 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 Second National Youth Congress participants 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. From independence until the present, Indonesian has functioned as the primary language in education, mass media, and government activities. Many local language speakers are increasingly using Indonesian with their children because they believe it will aid them to attain 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) have been proposed. Their coverage, however, focuses on high-resource languages. 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 (TyDiQA; Clark et al., 2020), common sense reasoning (XCOPA; Ponti et al., 2020), abstractive summarization (Hasan et al., 2021), passage ranking (mMARCO; Bonifacio et al., 2021), cross-lingual visual question answering (xGQA; Pfeiffer et al., 2021), language and vision reasoning (MaRVL; Liu et al., 2021), paraphrasing (ParaCotta; Aji et al., 2021), dialogue systems (XPersona & BiToD; Lin et al., 2021a,b), lexical normalization (MultiLexNorm; van der Goot et al., 2021), and machine translation (FLORES-101; Guzmán et al., 2019) include Indonesian but most others do not, and very few include Indonesian local lan-

guages. 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. Recent studies, however, have raised concerns regarding the quality of such multilingual corpora for underrepresented languages (Caswell et al., 2022).

2.3 Progress in Indonesian NLP

NLP research on Indonesian has occurred across multiple topics, such as POS tagging (Wicaksono and Purwarianti, 2010; Dinakaramani et al., 2014), NER (Budi et al., 2005; Rachman et al., 2017; Gunawan et al., 2018), sentiment analysis (Naradhipa and Purwarianti, 2011; Lunando and Purwarianti, 2013; Wicaksono et al., 2014), hate speech detection (Alfina et al., 2017; Sutejo and Lestari, 2018), topic classification (Winata and Khodra, 2015; Kusumaningrum et al., 2016), question answering (Mahendra et al., 2008; Fikri and Purwarianti, 2012), machine translation (Yulianti et al., 2011; Simon and Purwarianti, 2013; Hermanto et al., 2015), keyphrases extraction (Saputra et al., 2018; Trisna and Nurwidyantoro, 2020), morphological analysis (Pisceldo et al., 2008), and speech recognition (Lestari et al., 2006; Baskoro and Adriani, 2008; Zahra et al., 2009). However, many of these studies either did not release the data or used non-standardized resources with a lack of documentation and open source code, making them extremely difficult to reproduce.

Recently, Wilie et al. (2020), Koto et al. (2020b, 2021), and Cahyawijaya et al. (2021) collected Indonesian NLP resources as benchmark data. Others have also begun to create standardized labeled data for Indonesian NLP, e.g. the works of Kurniawan and Aji (2018), Guntara et al. (2020), Koto et al. (2020a), Khairunnisa et al. (2020), and Mahendra et al. (2021).

On the other hand, there has been very little work on local languages. Several works studied stemming (Sundanese (Suryani et al., 2018); Balinese (Subali and Fatichah, 2019)) and POS Tagging (Madurese; Dewi et al., 2020). Koto and Koto (2020) built a Indonesian Minangkabau parallel corpus and also sentiment analysis resources for Minangkabau. Other works developed machine translation systems between Indonesian and local languages, e.g., Sundanese (Suryani et al., 2015), Buginese (Apriani et al., 2016), Dayak Kanayatn (Hasbiansyah et al., 2016), and Sambas Malay (Ningtyas et al., 2018).

Tanaya and Adriani (2016, 2018) studied Javanese character segmentation in non-Latin script. Safitri et al. (2016) worked on spoken data language identification in Minangkabau, Sundanese, and Javanese, while Azizah et al. (2020) developed end-to-end neural text-to-speech models for Indonesian, Sundanese, and Javanese. Nasution et al. (2017, 2021) proposed an approach for bilingual lexicon induction and evaluated the approach on seven languages, i.e., Indonesian, Malay, Minangkabau, Palembang Malay, Banjar, Javanese, and Sundanese.

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 Unlabeled corpora are crucial for building large language models, such as GPT-2 (Radford et al., 2019) or BERT (Devlin et al., 2019). Available unlabeled 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.

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 respectively while in mC4 (Xue

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

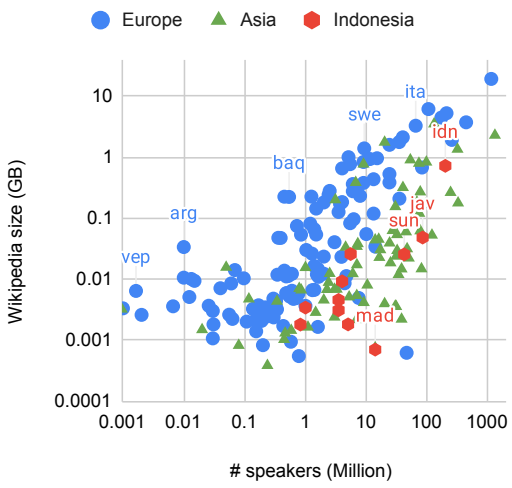

Figure 4: Relationship between the number of speakers and the size of data in Wikipedia for languages spoken in Europe, Asia, and Indonesia.

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 in Figure 4.[2] Much less data is available for the languages spoken in Indonesia, 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 with comparable numbers of speakers have more than 1.5 GB of articles. Similar trends hold for most other Asian languages. Languages in Africa are even more underrepresented in terms of Wikipedia data (see Appendix B).

Beyond the highly spoken local languages, most other Indonesian local languages do not have Wikipedia instances, in contrast to European languages with few speakers. 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 long-tail languages are even more scarce 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 challenging to obtain. These statistics demonstrate that collecting unlabeled 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 NLP systems.

Labeled Data Most work on Indonesian NLP (see §2) has not publicly released the data or models, limiting reproducibility. Although recent Indonesian NLP benchmarks are addressing this issue, they mostly focus on the Indonesian language (see Appendix F). Some widely spoken local languages such as Javanese, Sundanese, or Minangkabau have extremely small labeled 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 labeled 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 reflected in 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) showed that villages across the Jambi province use different dialects of Jambi Malay. Similarly, Kartikasari et al. (2018) showed that Javanese between different cities in central and eastern Java could have more than 50% lexical variation, while Purwaningsih (2017) showed that Javanese in different districts in the Lamongan 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).

[2]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.

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 (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 variations of Javanese dialects and styles across different regions of the Java island. Native speakers are asked to translate the words, given the context.

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 respectively. For Jambi Malay, we sample the result from a prior work (Anderbeck, 2008). For Javanese, we ask native speakers to translate basic words into 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 among regions, we only consider *Krama* from the Eastern speakers' perspective.

Jambi Malay 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).[3] Examples from other languages are shown in Appendix D.

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.

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[4] 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 collect-

[3] *Krama* is used to speak formally (e.g., with older or respected people). However, people prefer to use Indonesian more in a 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.

[4] Our annotators are based in Banyumas, Jogjakarta, and Jember for Western, Central, and Eastern Javanese respectively. Using dialects from different cities might yield different results.

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.

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.

ing textual data in a language. Despite its importance, it is an open research area, particularly for underrepresented languages (Hughes et al., 2006, Caswell et al., 2022).

We compare langid.py (Lui and Baldwin, 2012), FastText (Joulin et al., 2017), and CLD3.[5] 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 underrepresented 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 (Sitaram et al., 2019, Winata et al., 2018, 2019b, Doğruöz et al., 2021). This phenomenon is common in Indonesian conversations (Barik et al., 2019, Johanes et al., 2020, Wibowo et al., 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.

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 orthographies that are mutually intelligible, as they are pronounced the same. Some examples can be seen in Table 6. Such a variety of written forms is common in local languages in Indonesia. This variation leads to a significantly larger vocabulary size, especially for NLP systems that use word-based representations, and presents a challenge to constrain the representations for different spellings of the same word to be similar.

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

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

within small communities may emphasize the spoken form. Some languages are also declining, and speakers may 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 are not well-distributed within the nation. Internet penetration in Indonesia is 73.7% in 2020 but is mainly concentrated on Java. Among the non-Internet users, 39% explain that they do not understand the technology, while 15% state that they do not have a device to access the Internet.[6] In some areas where the 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 in top universities in the country.[7]

The dynamics of population movement in Indonesia also need to be taken into consideration. For example, urban communities 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 underrepresented 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, McMillan-Major et al., 2022), we recommend the following considerations.

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

[7]For instance, we estimate the whole CS Faculty of the country's top university to have fewer than 10 GPUs.

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

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 in other languages. In addition, this metadata can be used to document the formality level of the text, so it may be useful for research on modeling style or style transfer.

4.2 Potential Research Directions

Among the most spoken local languages, a lot of research has been done on mainstream NLP tasks such as hate-speech detection, sentiment analysis, entity recognition, and machine translation. Some research has even been deployed in production by industry. Many of the languages, however, are not widely spoken and under-explored. Focusing on these languages, we suggest future research direction as follows.

Data-Efficient NLP Pretrained language models, which have taken the NLP world by storm, require an abundance 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, Kurniawan et al., 2021), few-shot learning (Winata et al., 2021, 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 Collection 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 for 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 the Appendix) might help facilitate building translation systems with relatively little 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 data can be explored.

Compute-Efficient NLP The costly GPU requirement for current NLP models hinders adoption by local research institutions and industries. Instead of focusing on building yet another massive model, we suggest focusing on developing lightweight and fast neural architectures, for example through distillation (Kim et al., 2019; Sanh et al., 2019; Jiao et al., 2020), model factorization (Winata et al., 2019a) or model pruning (Voita et al., 2019). We also recommend research on more efficient training mechanisms (Aji and Heafield, 2017; Diskin et al., 2021). In addition, non-neural methods are still quite popular in Indonesia. Therefore, further research on the trade-off between the efficiency and quality of the models is also an interesting research direction.

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 explorations 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. This problem is also explored by Tan and Joty (2021) in an adversarial setting.

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 (Chung et al., 2021; Serdyuk et al., 2018), speech recognition (Besacier et al., 2014; Winata et al., 2020a,b), and multimodality (Dai et al., 2020, 2021) in order to improve NLP for such languages.

4.3 Engagement with Communities

As discussed in §3.4, it is difficult to generalize a solution across local languages. We thus encourage the NLP community, such as the Indonesian Association of Computational Linguistics (INACL)[8] to work more closely with native speakers and local communities. Local communities who work on linguistics such as *Polyglot Indonesia*,[9] *Merajut Indonesia*,[10] and *Masyarakat Linguistik Indonesia*[11] would be relevant collaborators to provide solutions and resources that support use cases benefiting the native speakers and communities of underrepresented languages. We advise the involvement of linguists, for example, to aid the language documentation process (Anastasopoulos et al., 2020). We also support open-science movements such as BigScience[12] or ICLR CoSubmitting Summer[13] to help start collaborations and reduce the barrier to entry to NLP research.

5 Conclusion

In this paper, we highlight challenges in Indonesian NLP. Indonesia is one of the most populous countries and the second-most linguistically diverse country of the world, with over 700 local languages, yet Indonesian NLP is underrepresented and underexplored. Based on the observed challenges, we also present recommendations to improve the situation, not only for Indonesian but also for other underrepresented languages.

Acknowledgments

We are grateful to Alexander Gutkin and Ling-Yen Liao for valuable feedback on an earlier draft of this paper. We also thank the anonymous reviewers for their helpful feedback. Lastly, we acknowledge the support of Kata.ai in this work.

[8] https://inacl.id/inacl/
[9] http://polyglotindonesia.org
[10] https://merajutindonesia.id/
[11] https://www.mlindonesia.org/
[12] https://bigscience.huggingface.co/
[13] 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. The Australian National University, Canberra.

Zaenal Abidin, Permata, Imam Ahmad, and Rusliyawati. 2021. Effect of mono corpus quantity on statistical machine translation Indonesian–Lampung dialect of nyo. In *Journal of Physics: Conference Series*, volume 1751. IOP Publishing.

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.

Mirna Adriani, Jelita Asian, Bobby Nazief, Seyed MM Tahaghoghi, and Hugh E Williams. 2007. Stemming Indonesian: A confix-stripping approach. *ACM Transactions on Asian Language Information Processing (TALIP)*, 6(4):1–33.

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.

Alham Fikri Aji, Radityo Eko Prasojo Tirana Noor Fatyanosa, Philip Arthur, Suci Fitriany, Salma Qonitah, Nadhifa Zulfa, Tomi Santoso, and Mahendra Data. 2021. Paracotta: Synthetic multilingual paraphrase corpora from the most diverse translation sample pair. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 666–675.

Alham Fikri Aji and Kenneth Heafield. 2017. Sparse communication for distributed gradient descent. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 440–445.

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. IEEE.

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.

Tri Apriani, Herry Sujaini, and Novi Safriadi. 2016. Pengaruh kuantitas korpus terhadap akurasi mesin penerjemah statistik bahasa bugis wajo ke Bahasa Indonesia. *JUSTIN (Jurnal Sistem Dan Teknologi Informasi)*, 4(1):168–173.

Valentina Kania Prameswara Artari, Rahmad Mahendra, Meganingrum Arista Jiwanggi, Adityo Anggraito, and Indra Budi. 2021. A multi-pass sieve coreference resolution for Indonesian. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 79–85.

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, Hong Kong, China. Association for Computational Linguistics.

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.

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*. CSLI Publications, Stanford, USA.

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.

Luiz Henrique Bonifacio, Israel Campiotti, Vitor Jeronymo, Roberto Lotufo, and Rodrigo Nogueira. 2021. mmarco: A multilingual version of the ms marco passage ranking dataset. *arXiv preprint arXiv:2108.13897*.

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.

Budiono, Hammam Riza, and Chairil Hakim. 2009. Resource report: building parallel text corpora for multi-domain translation system. In *Proceedings of the 7th Workshop on Asian Language Resources (ALR7)*, pages 92–95.

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 Khodra, Ayu Purwarianti, and Pascale Fung. 2021. IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8875–8898, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Isaac Caswell, Julia Kreutzer, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Auguste Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios Gonzales, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Rubungo Andre Niyongabo, Toan Q. Nguyen, Mathias Muller, Andr'e Muller, Shamsuddeen Hassan Muhammad, Nanda Firdausi Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, M. Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine cCabuk Balli, Stella Rose Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi N. Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

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. In Kingsley Bolton, Werner Botha, and Andy Kirkpatrick, editors, *The Handbook of Asian Englishes*, 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 *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5305–5316, Online. Association for Computational Linguistics.

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.

Nindian Puspa Dewi, Joan Santoso, Ubaidi Ubaidi, and Eka Rahayu Setyaningsih. 2020. Combination of genetic algorithm and Brill tagger algorithm for part of speech tagging Bahasa Madura. *Proceeding of the Electrical Engineering Computer Science and Informatics*, 7(2):38–42.

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.

Michael Diskin, Alexey Bukhtiyarov, Max Ryabinin, Lucile Saulnier, Anton Sinitsin, Dmitry Popov, Dmitry Pyrkin, Maxim Kashirin, Alexander Borzunov, Albert Villanova del Moral, et al. 2021. Distributed deep learning in open collaborations. *Advances in Neural Information Processing Systems*, 34.

A. Seza Doğruöz, Sunayana Sitaram, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2021. A survey of code-switching: Linguistic and social perspectives for language technologies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1654–1666, Online. Association for Computational Linguistics.

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).

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

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.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Muhammad Yudistira Hanifmuti and Ika Alfina. 2020. Aksara: An Indonesian morphological analyzer that conforms to the UD v2 annotation guidelines. In *2020 International Conference on Asian Language Processing (IALP)*, pages 86–91. IEEE.

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.

Muhammad Hasbiansyah, Herry Sujaini, and Novi Safriadi. 2016. Tuning for quality untuk uji akurasi mesin penerjemah statistik (MPS) Bahasa Indonesia-bahasa dayak kanayatn. *JUSTIN (Jurnal Sistem dan Teknologi Informasi)*, 4(1):209–213.

Andi Hermanto, Teguh Bharata Adji, and Noor Akhmad Setiawan. 2015. Recurrent neural network language model for English-Indonesian machine translation: Experimental study. In *2015 International conference on science in information technology (ICSITech)*, pages 132–136. IEEE.

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*.

Baden Hughes, Timothy Baldwin, Steven Bird, Jeremy Nicholson, and Andrew MacKinlay. 2006. Reconsidering language identification for written language resources. In *Proceedings of LREC 2006*, pages 485–488, Genoa, Italy.

Muhammad Okky Ibrohim and Indra Budi. 2018. A dataset and preliminaries study for abusive language detection in Indonesian social media. *Procedia Computer Science*, 135:222–229. The 3rd International Conference on Computer Science and Computational Intelligence (ICCSCI 2018) : Empowering Smart Technology in Digital Era for a Better Life.

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.

Rini Jannati, Rahmad Mahendra, Cakra Wishnu Wardhana, and Mirna Adriani. 2018. Stance classification towards political figures on blog writing. In *2018 International Conference on Asian Language Processing (IALP)*, pages 96–101.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174.

Ricky Chandra Johanes, Rahmad Mahendra, and Brahmastro Kresnaraman. 2020. Structuring code-switched product titles in Indonesian e-commerce platform. In *Computational Data and Social Networks*, pages 217–227, Cham. Springer International Publishing.

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 the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

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.

Siti Oryza Khairunnisa, Aizhan Imankulova, and Mamoru Komachi. 2020. Towards a standardized dataset on Indonesian named entity 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: Student Research Workshop*, pages 64–71, Suzhou, China. Association for Computational Linguistics.

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 the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1312–1323, Online. Association for Computational Linguistics.

Young Jin Kim, Marcin Junczys-Dowmunt, Hany Hassan, Alham Fikri Aji, Kenneth Heafield, Roman Grundkiewicz, and Nikolay Bogoychev. 2019. From research to production and back: Ludicrously fast neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 280–288.

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*, pages 10660–10668, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Fajri Koto and Gemala Y Rahmaningtyas. 2017. Inset lexicon: Evaluation of a word list for Indonesian sentiment analysis in microblogs. In *2017 International Conference on Asian Language Processing (IALP)*, pages 391–394. IEEE.

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, Lea Frermann, Philip Schulz, and Trevor Cohn. 2021. PPT: Parsimonious parser transfer for unsupervised cross-lingual adaptation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2907–2918, Online. Association for Computational Linguistics.

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.

Retno Kusumaningrum, M. Ihsan Aji Wiedjayanto, Satriyo Adhy, and Suryono. 2016. Classification of Indonesian news articles based on latent Dirichlet allocation. In *2016 International Conference on Data and Software Engineering (ICoDSE)*, pages 1–5.

Septina Dian Larasati. 2012. IDENTIC corpus: Morphologically enriched Indonesian-English parallel corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 902–906, Istanbul, Turkey. European Language Resources Association (ELRA).

Septina Dian Larasati, Vladislav Kuboň, and Daniel Zeman. 2011. Indonesian morphology tool (morphind): Towards an Indonesian corpus. In *International Workshop on Systems and Frameworks for Computational Morphology*, pages 119–129. Springer.

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. 2021a. 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.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, Peng Xu, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale Fung. 2021b. BiToD: A bilingual multi-domain dataset for task-oriented dialogue modeling. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, pages 342–351.

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*, pages 10511–10527, Online and Punta Cana, Dominican Republic. 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.

Angelina McMillan-Major, Zaid Alyafeai, Stella Biderman, Kimbo Chen, Francesco De Toni, Gérard Dupont, Hady Elsahar, Chris Emezue, Alham Fikri Aji, Suzana Ilić, et al. 2022. Documenting geographically and contextually diverse data sources: The bigscience catalogue of language data and resources. *arXiv preprint arXiv:2201.10066*.

David Moeljadi. 2017. Building JATI: A Treebank for Indonesian. In *Proceedings of The 4th Atma Jaya Conference on Corpus Studies (ConCorps 4)*, pages 1–9, Jakarta.

David Moeljadi, Francis Bond, and Sanghoun Song. 2015. Building an HPSG-based Indonesian resource grammar (INDRA). In *Proceedings of the Grammar Engineering Across Frameworks (GEAF) 2015 Workshop*, pages 9–16, Beijing, China. Association for Computational Linguistics.

David Moeljadi, Aditya Kurniawan, and Debaditya Goswami. 2019. Building Cendana: a treebank for informal Indonesian. In *Proceedings of the 33rd Pacific Asia Conference on Language, Information and Computation*, pages 156–164. Waseda Institute for the Study of Language and Information.

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.

Arbi Haza Nasution, Yohei Murakami, and Toru Ishida. 2017. A generalized constraint approach to bilingual dictionary induction for low-resource language families. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 17(2):1–29.

Arbi Haza Nasution, Yohei Murakami, and Toru Ishida. 2021. Plan optimization to bilingual dictionary induction for low-resource language families. *ACM Transactions Asian Low-Resource Language Information Processing*, 20(2).

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.

Della Widya Ningtyas, Herry Sujaini, and Novi Safriadi. 2018. Penggunaan pivot language pada mesin penerjemah statistik bahasa inggris ke bahasa melayu sambas. *JEPIN (Jurnal Edukasi dan Penelitian Informatika)*, 4(2):173–178.

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

Hiroki Nomoto, Hannah Choi, David Moeljadi, and Francis Bond. 2018. MALINDO Morph: Morphological dictionary and analyser for Malay/Indonesian. In *Proceedings of the LREC 2018 Workshop "The 13th Workshop on Asian Language Resources"*, pages 36–43.

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.

Tiago Pimentel, Maria Ryskina, Sabrina J. Mielke, Shijie Wu, Eleanor Chodroff, Brian Leonard, Garrett Nicolai, Yustinus Ghanggo Ate, Salam Khalifa, Nizar Habash, Charbel El-Khaissi, Omer Goldman, Michael Gasser, William Lane, Matt Coler, Arturo Oncevay, Jaime Rafael Montoya Samame, Gema Celeste Silva Villegas, Adam Ek, Jean-Philippe Bernardy, Andrey Shcherbakov, Aziyana Bayyr-ool, Karina Sheifer, Sofya Ganieva, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Andrew Krizhanovsky, Natalia Krizhanovsky, Clara Vania, Sardana Ivanova, Aelita Salchak, Christopher Straughn, Zoey Liu, Jonathan North Washington, Duygu Ataman, Witold Kieraś, Marcin Woliński, Totok Suhardijanto, Niklas Stoehr, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Richard J. Hatcher, Emily Prud'hommeaux, Ritesh Kumar, Mans Hulden, Botond Barta, Dorina Lakatos, Gábor Szolnok, Judit Ács, Mohit Raj, David Yarowsky, Ryan Cotterell,

Ben Ambridge, and Ekaterina Vylomova. 2021. Sigmorphon 2021 shared task on morphological reinflection: Generalization across languages. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–259, Online. Association for Computational Linguistics.

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.

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, Alvin Andhika, Alfan Farizki Wicaksono, Irfan Afif, and Filman Ferdian. 2016. Inanlp: Indonesia natural language processing toolkit, case study: Complaint tweet classification. In *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*, pages 1–5. IEEE.

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.

Desmond Darma Putra, Abdul Arfan, and Ruli Manurung. 2008. Building an Indonesian wordnet. In *Proceedings of the 2nd International MALINDO Workshop*.

Oddy Virgantara Putra, Fathin Muhammad Wasmanson, Triana Harmini, and Shoffin Nahwa Utama. 2020. Sundanese twitter dataset for emotion classification. In *2020 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM)*, pages 391–395. IEEE.

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.

Valdi Rachman, Septiviana Savitri, Fithriannisa Augustianti, and Rahmad Mahendra. 2017. Named entity recognition on Indonesian Twitter posts using long short-term memory networks. In *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 228–232.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Anna Rogers, Timothy Baldwin, and Kobi Leins. 2021. 'just what do you think you're doing, dave?' a checklist for responsible data use in NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4821–4833, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

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

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Ilham Fathy Saputra, Rahmad Mahendra, and Alfan Farizki Wicaksono. 2018. Keyphrases extraction from user-generated contents in healthcare domain using long short-term memory networks. In *Proceedings of the BioNLP 2018 workshop*, pages 28–34, Melbourne, Australia. Association for Computational Linguistics.

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. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

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)*.

HS Simon and Ayu Purwarianti. 2013. Experiments on Indonesian-Japanese statistical machine translation. In *2013 IEEE international conference on computational intelligence and cybernetics (CYBERNETICSCOM)*, pages 80–84. IEEE.

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.

Sunayana Sitaram, Khyathi Raghavi Chandu, Sai Krishna Rallabandi, and Alan W Black. 2019. A survey of code-switched speech and language processing. *arXiv preprint arXiv:1904.00784*.

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.

Made Agus Putra Subali and Chastine Fatichah. 2019. Kombinasi metode rule-based dan n-gram stemming untuk mengenali stemmer bahasa bali. *Jurnal Teknologi Informasi dan Ilmu Komputer*, 6(2):219–228.

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 Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 17(4):1–28.

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.

Samson Tan and Shafiq Joty. 2021. Code-mixing on sesame street: Dawn of the adversarial polyglots. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3596–3616.

Dipta Tanaya and Mirna Adriani. 2016. Dictionary-based word segmentation for Javanese. *Procedia Computer Science*, 81:208–213.

Dipta Tanaya and Mirna Adriani. 2018. Word segmentation for Javanese character using dictionary, SVM, and CRF. In *2018 International Conference on Asian Language Processing (IALP)*, pages 240–243.

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*.

I Nyoman Prayana Trisna and Arif Nurwidyantoro. 2020. Single document keywords extraction in Bahasa Indonesia using phrase chunking. *Telkomnika*, 18(4):1917–1925.

Rob van der Goot, Alan Ramponi, Arkaitz Zubiaga, Barbara Plank, Benjamin Muller, Iñaki San Vicente Roncal, Nikola Ljubešić, Özlem Çetinoğlu, Rahmad Mahendra, Talha Çolakoğlu, Timothy Baldwin, Tommaso Caselli, and Wladimir Sidorenko. 2021. MultiLexNorm: A shared task on multilingual lexical normalization. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (WNUT 2021)*, pages 493–509, Online. Association for Computational Linguistics.

Daan van Esch, Tamar Lucassen, Sebastian Ruder, Isaac Caswell, and Clara Rivera. 2022. Writing system and speaker metadata for 2,800+ language varieties.

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

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808.

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

Devid Haryalesmana Wahid and SN Azhari. 2016. Peringkasan sentimen esktraktif di twitter menggunakan hybrid tf-idf dan cosine similarity. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, 10(2):207–218.

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*.

Alfan Farizki Wicaksono, Clara Vania, Bayu Distiawan, and Mirna Adriani. 2014. Automatically building a corpus for sentiment analysis on Indonesian tweets. In *Proceedings of the 28th Pacific Asia conference on language, information and computing*, pages 185–194.

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. 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, Suzhou, China. Association for Computational Linguistics.

Andika William and Yunita Sari. 2020. CLICK-ID: A novel dataset for Indonesian clickbait headlines. *Data in Brief*, 32:106231.

Genta Indra Winata. 2021. *Multilingual transfer learning for code-switched language and speech neural modeling*. Hong Kong University of Science and Technology (Hong Kong).

Genta Indra Winata, Samuel Cahyawijaya, Zhaojiang Lin, Zihan Liu, and Pascale Fung. 2020a. 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, Peng Xu, and Pascale Fung. 2020b. Learning fast adaptation on cross-accented speech recognition. *Proc. Interspeech 2020*, pages 1276–1280.

Genta Indra Winata and Masayu Leylia Khodra. 2015. Handling imbalanced dataset in multi-label text categorization using bagging and adaptive boosting. In *2015 International Conference on Electrical Engineering and Informatics (ICEEI)*, pages 500–505. IEEE.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. 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, Jamin Shin, Elham J Barezi, and Pascale Fung. 2019a. On the effectiveness of low-rank matrix factorization for lstm model compression. In *Proceedings of the 33rd Pacific Asia Conference on Language, Information and Computation*, pages 253–262. Waseda Institute for the Study of Language and Information.

Genta Indra Winata, Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018. Code-switching language modeling using syntax-aware multi-task learning. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 62–67.

Genta Indra Winata, Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2019b. 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.

Wilson Wongso, David Samuel Setiawan, and Derwin Suhartono. 2021. Causal and masked language modeling of Javanese language using transformer-based architectures. In *2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 1–7. IEEE.

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 5, we contrast the count of publications related to Indonesian languages compared to European languages. Although there are many more Indonesian speakers than most European languages, the amount of published research relating to Indonesian is still comparatively lower than for European languages.

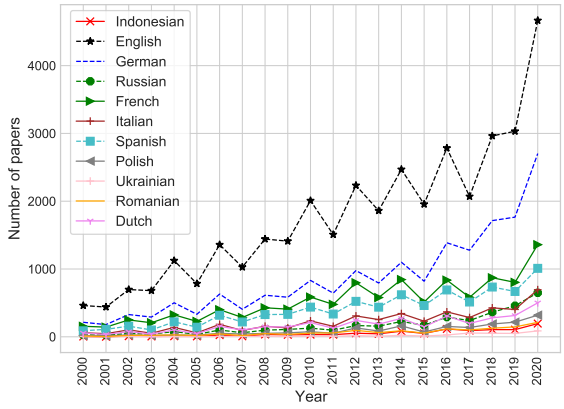

Figure 5: Count of published papers per year on Indonesian and European languages from 2000 to 2020.

B Wikipedia Availability

In Figure 6, we compare Wikipedia size (in GB file size) compared to the number of speakers for various languages. We show that some African and indigenous American languages are even more under-resourced.

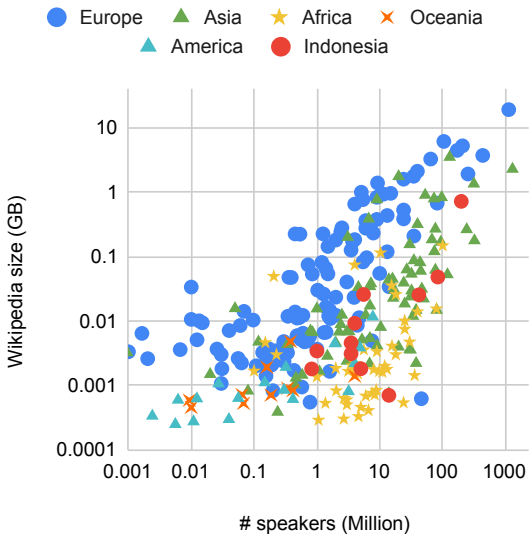

Figure 6: Wikipedia data size (in GB) compared to the number of speakers among Indonesia and different regions across the world.

C Wikipedia Vocabulary Overlap

In Table 7, we present vocabulary statistics for Indonesian languages based on Wikipedia. Due to the noisy nature of Wikipedia, we use "*Kamus Besar Bahasa Indonesia*" (KBBI) third edition,[14] the official dictionary for the Indonesian language to filter out the top 1% and top-100 most frequent words. As expected, the top 1% words are less reliable, with only 59.3% of 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.

Lang	# Vocab		
	All (k)	Top 1% ∩ KBBI (%)	Top 100 ∩ KBBI (%)
ind	2023	59.3	**96**
jav	435	46.8	43
sun	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.[15]

D Dialect Differences

In this section, we present more examples of lexical variations in local Indonesian languages. Maharani and Candra (2018) and Sarwadi et al. (2019) show lexical variations in Balinese and Sasak, respectively, where they asked locals to translate general/common words. Then, they compared the vocabulary across different locations (in this case, villages) to each other. Some of the examples can be seen in Tables 8 and 9. Unfortunately, they did not provide quantitative results. Pamolango (2012) conducted a similar experiment in the Banggai district of South Sulawesi across 31 observation points for the Saluan language. While Pamolango (2012)

[14] https://github.com/geovedi/indonesian-wordlist
[15] KBBI is the official Indonesian dictionary.

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 8: 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 9: Lexical variation of Sasak across different villages in North Lombok district (Sarwadi et al., 2019)

did not provide full examples, they reported up to 23.5% lexical variation among 200 basic vocabulary items.

E Local Language Classification

As shown in Table 10, some Javanese texts are misidentified as Indonesian, English, and Malay. This is because Javanese and Indonesian (which is similar to Malay) share some words. We believe English misclassification is due to the data size bias.

F Indonesian NLP Resources

In Table 11, we present the list of existing monolingual and parallel corpora in Indonesian and local languages. Table 12 shows the list of publicly available NLP datasets for Indonesian. Table 13 shows the list of NLP tools and resources for Indonesian. Table 14 shows the list of publicly available NLP datasets for local languages spoken in Indonesia.

Although the volume of data in local languages is much smaller than that for Indonesian, these resource collections are arguably beneficial for constructing resources in other local languages. This

Dialect/ Style	Method	Class jav	idn	eng	mys
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 10: Language identification misclassification rate.

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).

	Paper	Language(s)	Domain/Source	Size	Link
MONO	Wilie et al. (2020)	ind	multi domain	3.6B	Indo4B
	Conneau et al. (2020)	ind	Web	22.7B	CC–100 Indonesian
		jav	Web	24M	CC–100 Javanese
PARALLEL	Budiono et al. (2009)	ind ↔ eng	News	24k	PANL BPPT
	Larasati (2012)	ind ↔ eng	multi domain	45k	Identic
	Guntara et al. (2020)	ind ↔ eng	Wikipedia	93k	General En-Id
		ind ↔ eng	News	42k	News En-Id
		ind ↔ eng	Religion	590k	Religious En-Id
	Cahyawijaya et al. (2021)	ind ↔ eng	Religion	31k	
		jav ↔ eng	Religion	16k	Bible En-Jav
		sun ↔ eng	Religion	16k	Bible En-Sun
	Koto and Koto (2020)	ind ↔ min	Wikipedia	16k	MinangNLP MT
	Abidin et al. (2021)	ind ↔ abl	Book	3k	Parallel: Indonesian - Lampung Nyo

Table 11: List of monolingual and parallel corpora involving Indonesian and local languages. The size of the monolingual corpora is the number of words, while the size of the parallel corpora is the number of sentences.

Work by	Task	Domain/Source	Size	Dataset link
Pimentel et al. (2021)	Morphology Analysis	Dictionary, Wikipedia	27k	unimorph id
Dinakaramani et al. (2014)	POS Tagging	News	10k	POS bahasa.cs.ui.ac.id
Hoesen and Purwarianti (2018)	POS Tagging	News	8k	IndoNLU/POSP
Hoesen and Purwarianti (2018)	Named Entity Recognition	News	8k	IndoNLU/NERP
NERGrit	Named Entity Recognition	News	23k	nergrit-corpus
Fachri (2014)	Named Entity Recognition	News	2k	Indonesian NER
Alfina et al. (2016)	Named Entity Recognition	Wikipedia	48k	Singgalang modified-dee
Mahendra et al. (2018)	Word Sense Disambiguation	multi domain	2k	Indonesian WSD
Moeljadi (2017)	Constituency Parsing	Dictionary	1.2k	JATI
Moeljadi et al. (2019)	Constituency Parsing	Chat	0.7k	Cendana
Arwidarasti et al. (2019)	Constituency Parsing	News	1k	kethu
McDonald et al. (2013)	Dependency Parsing	News, Blog	5k	UD_Indonesian-GSD
Zeman et al. (2018)	Dependency Parsing	News, Wikipedia	1k	UD_Indonesian-PUD
Artari et al. (2021)	Coreference Resolution	Wikipedia	0.2k	IndoCoref
Purwarianti et al. (2007)	Question Answering	News	3k	IndoNLU/FacQA
Clark et al. (2020)	Question Answering	Wikipedia	18k	tydiqa/id
Kurniawan and Louvan (2018)	Summarization	News	20k	IndoSum
Koto et al. (2020a)	Summarization	News	215k	Liputan6
Mahfuzh et al. (2019)	Keyphrases Extraction	Twitter	1k	IndoNLU/KEPS
Setya and Mahendra (2018)	Natural Language Inference	Wikipedia	0.5k	IndoNLU/WreTe
Mahendra et al. (2021)	Natural Language Inference	Wikipedia, News	18k	IndoNLI
Purwarianti and Crisdayanti (2019)	Sentiment Analysis	Review	13k	IndoNLU/SmSA
Ilmania et al. (2018)	Sentiment Analysis	Review	1k	IndoNLU/CASA
Azhar et al. (2019)	Sentiment Analysis	Review	3k	IndoNLU/HoASA
Koto et al. (2020b)	Sentiment Analysis	Twitter, Review	5k	IndoLEM/sentiment
Saputri et al. (2018)	Emotion Classification	Twitter	4k	Indonesian emotion
Jannati et al. (2018)	Stance Detection	Blog	0.3k	Indonesian stance
Alfina et al. (2017)	Hate Speech Detection	Twitter	0.5k	id hatespeech
Ibrohim and Budi (2018)	Hate Speech Detection	Twitter	2k	id abusive
Ibrohim and Budi (2019)	Hate Speech Detection	Twitter	13k	id multilabel HS
William and Sari (2020)	Clickbait Detection	News	15k	Indonesian clickbait
Wibowo et al. (2020)	Style Transfer	Twitter	2k	STIF-Indonesia

Table 12: List of publicly available NLP datasets for the Indonesian language. The size of the dataset is defined by the number of sentences in most cases, except for morphology analysis (number of words), coreference resolution, summarization, and stance detection (number of articles).

Resource Name	Reference	Description
Sastrawi	–	Indonesian stemming implementation based on Adriani et al. (2007)'s algorithm
Pujangga	–	The interface for InaNLP (Purwarianti et al., 2016), an Indonesian NLP toolkit
INDRA	Moeljadi et al. (2015)	Indonesian resource grammar
MorphInd	Larasati et al. (2011)	Morphology tool
MALINDO Morph	Nomoto et al. (2018)	Morphological dictionary and analyzer
Aksara	Hanifmuti and Alfina (2020)	Indonesian morphological analyzer based on the UD v2 annotation guidelines
WordNet IWN	Putra et al. (2008)	WordNet (UI), consisting of 1,203 synsets and 1,659 words in Indonesian
WordNet Bahasa	Bond et al. (2014)	WordNet (NTU), consisting of 49,668 synsets and 64,131 unique words in Malay and Indonesian
Inset	Koto and Rahmaningtyas (2017)	Indonesian sentiment lexicon of 3,609 positive and 6,609 negative words
masDevid opinion word	Wahid and Azhari (2016)	List of Indonesian opinion words (1,182 positive and 2,402 negative), translated from Liu et al. (2005)
Kamus Alay	Salsabila et al. (2018)	Indonesian formal to informal lexicon
IndoCollex	Wibowo et al. (2021)	Indonesian formal to informal lexicon, categorized by the transformation type

Table 13: List of Indonesian NLP tools and resources.

Work by	Task	Language(s)	Domain/Source	Size	Dataset link
Koto and Koto (2020)	Sentiment analysis	min	Twitter, review	5000	Minang NLP sentiment
Putra et al. (2020)	Emotion classification	sun	Twitter	2518	Sundanese emotion
Putri et al. (2021)	Hate speech detection	jav	Twitter	3478	Javanese HS
Putri et al. (2021)	Hate speech detection	sun	Twitter	2209	Sundanese HS
Javanese NLP CSUI	Dependency parsing	jav	Wikipedia	125	UD Javanese CSUI
Wongso et al. (2021)	Sentiment analysis	jav	Review	100k	Javanese translated IMDB

Table 14: List of publicly available NLP datasets for local languages spoken in Indonesia. The size of the dataset is defined by the number of sentences.