Applications of Natural Language Processing in Bilingual Language Teaching: An Indonesian-English Case Study

Zara Maxwell-Smith[1] **Ben Foley**[2] **Simón González Ochoa**[1] **Hanna Suominen**[1,3–4]

1. The Australian National University / Canberra, ACT, Australia
2. University of Queensland / Brisbane, QLD, Australia
3. Data61/Commonwealth Scientific and Industrial Research Organisation / Canberra, ACT, Australia
4. University of Turku / Turku, Finland

Zara.Maxwell-Smith@anu.edu.au, b.foley@uq.edu.au,
Simon.Gonzalez@anu.edu.au, Hanna.Suominen@anu.edu.au

Abstract

Multilingual corpora are difficult to compile and a classroom setting adds pedagogy to the mix of factors which make this data so rich and problematic to classify. In this paper, we set out methodological considerations of using automated speech recognition to build a corpus of teacher speech in an Indonesian language classroom. Our preliminary results (64% word error rate) suggest these tools have the potential to speed data collection in this context. We provide practical examples of our data structure, details of our piloted computer-assisted processes, and fine-grained error analysis. Our study is informed and directed by genuine research questions and discussion in both the education and computational linguistics fields. We highlight some of the benefits and risks of using these emerging technologies to analyze the complex work of language teachers and in education more generally.

1 Introduction

Using quantitative methods to understand language learning and teaching is difficult work as limitations in the recording, transcribing, and analyzing of data continue to constrain the size of datasets. It is not surprising then, that quantitative studies looking at *second language*[1] *acquisition* have been critiqued for their low statistical power (Plonsky, 2013). Usage-based analyses of teacher corpora are an important next stage in understanding language acquisition (Ellis, 2017). Given the magnitude of worldwide investment in *L2 teaching and learning*, drawing on developments in automated methods of compiling this kind of speech data is timely.

Consequently, we sought to address the following main research question in this paper: How can *automated speech recognition* (ASR) be adapted for this use? More specifically, i) How well do these speech-to-text tools perform on this type of data? and ii) How do these tools and datasets relate to the overall purpose of opening a window into the practice of language teachers? Such an endeavor requires careful consideration of how ASR models are built, and what the underlying training data and desired output of such models might be. In this paper we use the term *ASR model* to refer to statistical models used to map speech sequences and sounds to respective text sequences (Jurafsky and Martin, 2009, pp. 38, 286, 287).

Our study is drawn from a project investigating the teaching of *Indonesian*. Data was collected from a tertiary Indonesian language program at an Australian university. A single teacher's speech was recorded throughout one semester of a second-year language program. In investigating Indonesian language teaching, ideally various instances of linguistic features and non-standard Indonesian would be annotated to allow for analyzing various topics, including, for instance, the comprehensibility of teachers' speech, movement between the L2 and assumed *first language* (L1), representations of regional Indonesian languages, and non-standard varieties and loanwords. Yet, the tools tend to restrict the data structures for annotating the audio. As an example from the conclusions of this paper, the classification of data as belonging to the L2 (Indonesian) or the L1 (*English*) quickly emerged as a very significant issue.

The paper is organized as follows: We begin by presenting an overview of related work in transcription and ASR before describing our methodological approach, with subsections on bilingual and classroom teacher data. This is followed by a more detailed description of our materials and methods to train and evaluate ASR models. Finally, we present experimental results of our machine transcription, discuss them, and conclude the study.

[1] a.k.a. target language or L2

Proceedings of the 15th Workshop on Innovative Use of NLP for Building Educational Applications, pages 124–134
July 10, 2020.

2 Background

Transcription is a complex task traditionally seen by linguists from the perspective of linguistic theory and documentation of complex language structures and phenomena. Linguists and their research teams become extremely familiar with their data during the process of transcribing, and their publications usually make reference to data-specific guidelines developed for their transcription teams (BNC-Consortium, 2007). Often these are adapted from generic guidelines or rules for annotating language which aim to record the "most basic transcription information: the words and who they were spoken by, the division of the stream of speech into turns and intonation units, the truncation of intonation units and words, intonation contours, medium and long pauses, laughter, and uncertain hearings or indecipherable words" (Du Bois et al., 1993). Most teams use sophisticated software tools, which provide a method for rich interlinear annotation of speech data by humans.[2] These annotations allow linguists to record more than 'just' the words used in human communication, but obviously cannot represent all characteristics of the audio data.

Acknowledging the *time constraints* and *subjectivity* or *bias* that enter the transcription process as transcription guidelines are developed is important. The purpose of these guidelines — namely, to create uniformity of practice from individual, and teams of transcribers — may not be achievable (Hovy and Lavid, 2010). In fact, experiments looking at the subjectivity of transcription led Lapadat and Lindsay (1998) to conclude that "the choices researchers make about transcription enact the theories they hold and constrain the interpretations they draw from their educational practice". Moreover, a transcription survey carried out by the *Centre of Excellence for the Dynamics of Language, Transcription Acceleration Project* (CoEDL TAP) team documented a significant variety in the way linguists go about transcribing their data. The survey also found that each minute of data takes, on average, 39 minutes for a linguist to transcribe, creating the well-known 'transcription bottleneck' (Durantin, 2017).

Advances in ASR and other *natural language processing* (NLP) bring researchers closer to overcoming this bottleneck, but many open challenges remain (Hirschberg and Manning, 2015). ASR tools can help by providing a first-pass hypothesis of audio for languages with large datasets to train the underlying models (Google, 2019; Nuance, 2019).[3] However, financial and ethical restrictions may prevent a study from using these off-the-shelf systems and cloud computing services.[4] Existing solutions may also have insufficient coverage of the domain-specific language used by speakers or not support a given L2.

Recognizing the potential benefit that integrating ASR tools into a linguist's workflow could have, the CoEDL TAP team has been building *Elpis* (Foley et al., 2018), an accessible interface for researchers to use the powerful but complex *Kaldi ASR toolkit*.[5] According to Gaida et al. (2014) "Compared to the other recognizers, the outstanding performance of Kaldi can be seen as a revolution in open-source [ASR] technology". This project constitutes an early use of the Elpis pipeline to prepare training data, ready for Kaldi to build ASR models, which can then be used to "infer" a hypothesis for un-transcribed audio.

The *interdisciplinary work* involved in this project shines a spotlight on the limitations on the type of data used by ASR systems; human transcribers often face difficult decisions as to what should and what should not be recorded in the training data.[6] Since bias and error may multiply in ASR models and create unreliable and undesirable outcomes, sharing the best practices and having transparent processes for creating training and evaluation data and protocols is of utmost importance (Hovy and Lavid, 2010). ASR trained on carefully compiled data can then be scientifically tested and variations to the training data analysed for their impact (Baur et al., 2018).

3 Methodological Approach

Our methodological considerations addressed the following: which ASR tool to use, how to prepare training data for this tool, and how to best manage the bias of the training data inherent in all transcription processes. Kaldi's orthographic transcription

[2] e.g., ELAN, Transana, and FLEx (last accessed on 4 December 2019)

[3] E.g., Google's Cloud Speech-to-Text and Nuance's ASR Self-Service support 120 languages and 86 languages/dialects, respectively (last accessed on 4 December 2019).

[4] E.g., purchasing a license for all participating teachers and obtaining each participant's informed consent for their speech data to be saved on a cloud owned by a private corporation, possibly using them to develop their ASR and other commercial products, may be infeasible or questionable.

[5] last accessed on 4 December 2019

[6] See our examples below for illustration.

capabilities and Elpis' processing and output of time-aligned ELAN files were a good fit with the broader research goals in lexical analysis, including dispersion analysis.

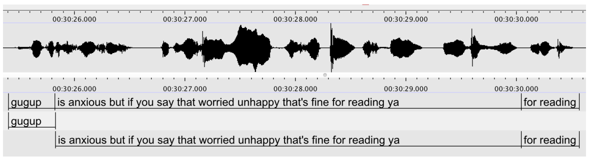
In general, we took a pragmatic approach to managing the loss of data from audio recordings, viewing information such as rising intonation[7] as something that was unnecessary to our lexical focus and which could be added later if the data were used for different research purposes. We were able to minimize some loss using a tier structure in the ELAN training data and this allowed us to maintain syntax relationships and other information used by Kaldi in the data.

Entwined in the issue of data loss, was the management of subjectivity in transcription. Indonesian and English native speakers, linguists, and an Indonesian language teacher worked together to transcribe our training data and we used extended discussions of specific samples to develop our transcription guidelines, including some discussions with our teacher participant. Meanwhile the tier structure allowed consideration of the teacher's behavior from an alternative framework discussed below; that is, translanguaging.

3.1 Transcription Decisions with Bilingual Data

Turell and Moyer (2009) argue that "transcription is already a first step in interpretation and analysis" and add that the complexity of the task inevitably increases when more than one language is at play: as the number of lexical items, morphemes, pragmatic strategies, and countless other linguistic possibilities increase, a transcriber must consider multiple possible 'first step' interpretations of their data.

In our data, the teacher used Australian English and Indonesian, the target language. Target languages are often understood as abstract and definable entities (Pennycook, 2016) used by imagined communities of native speakers (Norton, 2001). This is problematic as it hides the complexity and variation of natural languages, especially for Indonesian which exists in a highly diverse linguistic ecosystem; in Indonesian, complex concepts of social relationships play out in its variation across different speaking situations (Djenar, 2006, 2008; Morgan, 2011; Djenar and Ewing, 2015). Indonesian teachers, consciously or not, participate in and negotiate the politics of ethnic diversity and variation in urban and rural Indonesia (Goebel, 2010, 2014). Furthermore, Indonesian could be considered diglossic, with two varieties of the language in use in everyday situations (Sneddon, 2003).



gugup | is anxious but if you say that worried unhappy that's fine for reading ya | for reading

gugup

is anxious but if you say that worried unhappy that's fine for reading ya | for reading

Figure 1: Community of Practice shared repertoire: 'reading'

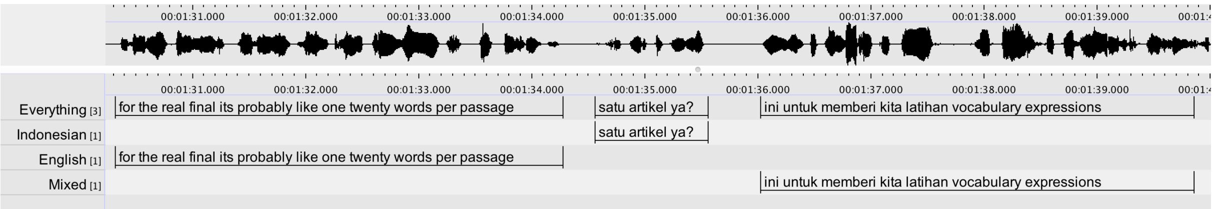
In our case study interview, the teacher explicitly acknowledged the diglossic nature of Indonesian and expressed the desire and intention to include *Colloquial Jakartan Indonesian* (CJI) in lessons as a speaking and listening target for students, while also stating that the written resources given to students focused more on the standardized or high variety of Indonesian. The teacher's intentions were consistent with the training data, which contained numerous CJI lexical items[8] and standard Indonesian. While not encountered in our small training dataset, our transcribers considered multiple English varieties due to the diverse English speaking experience of our teacher participant.

In addition to diglossic Indonesian and one variety of English, the teacher also used language consistent with the *Community of Practice* (CofP) framework, which, according to Wenger (1998, p. 76), involves a) mutual engagement, b) a joint negotiated enterprise, and c) a shared repertoire of negotiable resources accumulated over time. The teacher used language, or a repertoire, developed by the class through their interaction as a CofP. For example, the word 'reading' (Figure 1) was repurposed by the teacher participant to refer to a program-specific activity, assessment, and skillset that does not match with a general understanding/definition of this word in Australian English; it has become shorthand, or jargon, for something like a 'reading task'.

Thus far in this paper, we have relied on a presumption that it is desirable and theoretically sound to categorize teacher's speech into different languages. Such categorization rests on theorizing that languages are discrete entities and that teachers and students 'code-switch' — or alternate —

[7] which linguists often transcribe through special characters

[8] e.g., according to Sneddon (2006), 'gitu', 'dong', 'nggak' are from CJI



Everything [3] for the real final its probably like one twenty words per passage | satu artikel ya? | ini untuk memberi kita latihan vocabulary expressions

Indonesian [1] satu artikel ya?

English [1] for the real final its probably like one twenty words per passage

Mixed [1] ini untuk memberi kita latihan vocabulary expressions

Figure 2: ELAN tier structure

"between two languages or dialects of the same language within the same conversation" (Boztepe, 2003). Recent discussions of an alternative framework — *translanguaging* (Garcia and Wei, 2014) — propose that multilinguals employ only one, expanded repertoire of linguistic features. This repertoire may contain two or more languages which are officially and externally recognized as distinct systems, but according to translanguaging theory, the distinction between the systems is not internally valid.

By using several ELAN tiers to create parallel structures for storing data (Figure 2), we balanced technological requirements without taking a particular stance in relation to translanguaging nor the internal mechanisms of bilinguals. The uppermost 'Everything' tier included all orthographical annotations for the data. The next two tiers contained data, which according to various phonological, syntactical, and morphological factors were separated by our transcription team into Indonesian and English according to a code-switch paradigm. Finally, we also created a tier labeled 'Mixed' to contain annotations, which were difficult to separate.

While some researchers battle with technologies to represent very different orthographies[9], we worked with two languages that are both written in the Roman alphabet. This presented some challenges of its own. Some words, for example, 'status' *(status)* and 'level' *(level)*, were spelled identically in both languages; meanwhile, names could have been represented in a number of different ways. Our decisions to use a certain orthography in training data impacted statistical relationships between words and phonemes in the ASR models. We chose to approximate all names in the Indonesian orthography as these proper nouns are somewhat language independent.[10] For example, our 'Indonesianized' class list included 'Jorj' (George), 'Shantel' (Chantelle), and 'Medi' (Maddy). This decision allowed us to maintain the names within both Indonesian and English sentences, however, it did require manual creation of a phonemic map for that lexical item.[11] Similarly, we used only the Indonesian phonemic map for 'status' and 'level' as our participant's English incorporated Indonesian phonological characteristics (accent), and our intention was to strengthen our Indonesian computational model.

Decisions about orthography and tier allocations were very difficult and we made them only after extensive discussion in our transcription team. In some cases, within word changes between the typical Indonesian phonology and English phonology occurred. For example, in one segment, the teacher produced the first vowel of 'status' as [eɪ] (as in 'bait'), an English phoneme, but finished the word with the Indonesian /u/ (similar to 'book'). Even with a common set of characters used for bilingual data, the decisions taken developing training data had to be clearly documented and their impact considered in the ASR evaluation.

3.2 Toward Interpreting Teacher Speech

The complex bilingual transcription process outlined above was further complicated by the transcriber's interpretation of the educational setting. Given the conceivable criticism of a given teachers' professional practice made possible through the creation of corpora, we carefully considered the impacts of this scrutiny while developing our transcription guidelines and sought to minimize unfair or inaccurate treatment of teacher data. We also wish to proclaim the limitations of corpus data in this setting.

Although a full description and examination of these issues is beyond the scope of this paper, we identified some pertinent methodological implications of our own data structure. First, the task of analyzing the teacher's speech is likely to be over-simplified into binary L1 versus L2 catego-

[9] (e.g., Halai (2007) for Urdu and English)

[10] They were also annotated in all tiers when they occurred alone.

[11] See Section 4 below.

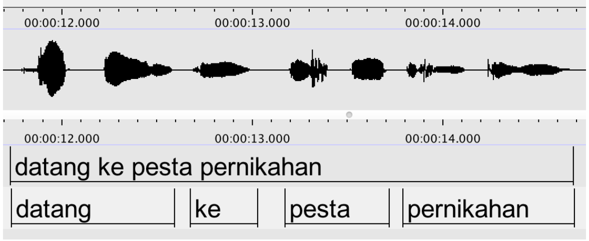


Figure 3: Treatment of pauses in teacher speech

rization of teacher's speech. The aforementioned methodological difficulties of teasing apart speech data and the questionable validity of delimiting languages raised by the translanguaging framework were central to our transcription guidelines. We also note that pauses in the teacher's language and other easily overlooked phenomena might skew the time counted towards a given language (Figure 3). We assessed that L2 was vulnerable to this skewing as the teacher extended pauses between words unfamiliar to the students, thus expanding the time counted as L2 speech. Conversely, cutting the L2 use apart when a teacher paused removed between-word-time from a cumulative L2 count and artificially shortened the time spent in the language.

Second, the goal of modifying sociolinguistic norms which brings people to language classrooms precipitated a level of variance and unpredictability unusual in other speech contexts as students learn and progress in their acquisition. We viewed variation in teacher speech from a pedagogically 'generous' perspective; for example, unusual linguistic forms were interpreted in line with research on language simplification (Saito and Poeteren, 2012; O Dela Rosa and Arguelles, 2016) or identity work with students (Norton and Toohey, 2011). However, a transcriber might note that in the Australian second language teaching setting, teachers often have less than 'native' proficiency in either the L2 or classroom L1. A proficiency-focused transcriber could be particularly sensitive to the teacher's productions of loanwords.[12] Thus, a transcriber's own perception of proficiency and speech errors, as well as their knowledge of, and stance on, pedagogical approaches are implicated in the interpretation of teacher speech.

With so many possible interpretations, asking the teacher to comment on or transcribe their own data might seem useful. However, the intent of a teacher in using specific linguistic features is likely to be highly complex, as well as difficult to ascertain as this work is often the result of internally reasoned, impromptu responses to student feedback (Borko et al., 1990). With these features put together over thousands, possibly millions of teaching decisions each lesson, we were cautious in our asking our teacher participant to recall or explain what they were doing in retrospect. We noted that any disparity in teacher intention and the recorded data, or inability to recall the purpose of specific interactions, language choices and other behaviors may create an air of scrutiny which could skew resulting interpretation (Gangneux and Docherty, 2018).

Ensuring that teachers, their work, and their decisions are not misrepresented or misunderstood was important to us. We emphasized and are urging caution in the use of corpora to assess teacher practice until methodological questions have received prolonged and rigorous attention across a wide-range of datasets, including at the minimum different L1 and L2, teachers, pedagogical styles, and teaching situations.

4 Materials and Methods

The audio data was recorded in a second year tertiary Indonesian language program at an Australian university (Ethics Approval No. 2017/889 of the ***Australian National University*** Human Research Committee for the Speech Recognition; Building Datasets from Indonesian Language Classrooms and Resources protocol). The teacher, who was recorded over the course of one semester, grew up using Indonesian in school and public places, and a regional language at home. A semester of over 32 hours of class was recorded.[13]

The teacher wore a head-mounted microphone and wireless bodypack linked to a ZOOM recorder set to record 44.1 kHz, 36-bit WAV format audio. Because students were not the target of the study, the microphone settings were optimized to exclude their voices. Three lessons of approximately 50 minutes were chosen for transcription as training and test data for the ASR. The lessons were selected to contain a range of content, instructional styles, and activities. The remaining audio recordings were held out from training and testing.[14]

[12] E.g., the Indonesian loanword 'kelas' (class) might be interpreted as English should it not meet a transcriber's personal threshold for an Indonesian production.

[13] Excluded classes in which formal assessments took place, an introductory class, and some lesson segments where technical problems resulted in loss of data.

[14] They were kept in reserve for future experiments and anal-

Model	Training tier[a]	Tokens in training	Languages	n-gram	WER[b]	WER full set[c]	Correct	Long text spans	
								Words	Text spans
Bilingual_1G	*Everything* tier	6194	English Indonesian Mixed	1	77% (56/73)	117% (117/100)	38/100	4 3	2 1
Bilingual_3G	*Everything* tier	6194	English Indonesian Mixed	3	89% (65/73)	133% (133/100)	34/100	3	2
Indonesian_3G	*Indonesian* tier	3377	Indonesian	3	64% (23/36)	134% (134/100)	22/100[d] 22/51[e]	4 3	1 1

[a] See Figure 2 for tier structure
[b] Results when testing only with words found in training data
[c] Results including training words and testing words not found in training data
[d] In the full test set
[e] Indonesian and non-language specific words in the test set

Table 1: *Word Error Rates* (WER) from 3 ASR Models

PRAAT auto-segmentation with settings at the minimum pitch of 70 Hz, silence threshold of -50 Db, and minimum silent interval of 0.25 was used to segment the data. Segments were then manually edited to remove remnant student voices and extreme modality sounds[15] to avoid confusing the Kaldi acoustic training. Care was taken to find the boundaries between speech sounds and discriminate between the languages used, with challenging sections examined in PRAAT by the transcription team. Transcription was completed in ELAN and initially all teacher speech was transcribed on one tier before being expanded onto other tiers (see Figure 2).

To use the Kaldi toolkit, a lexicon with each word's phonemic representation was required. Due to the bilingual dataset in this study, we built a lexicon with consistent *grapheme-to-phoneme* (G2P) mapping across two orthographies. Our lexicon was built by adding missing English words to the *Carnegie Melon University* (CMU) Pronunciation Dictionary. Although the pronunciations of this dictionary are based on American English, it was the best available match with our teacher participant. We then merged this lexicon with an Indonesian lexicon, which was built using Elpis functionalities.[16] The tools used the regular G2P mapping in Indonesian to generate a pronunciation dictionary based on the orthographical representation of each word.

We trained three models (Table 1) on two lessons selected from the semester of teaching. We then used the three models to automatically transcribe a 100-word[17] test subset of data from a third lesson. We used the *word error rate* (WER)[18] as the primary evaluation measure in this analysis.

The two bilingual models, which were trained on all parts of the audio recordings, are referred to as bilingual models for ease of reference. However, it should be noted that there was nothing binary in these models:[19] Bilingual_1G and Bilingual_3G were each a single model, where 1G and 3G refer to n-grams.[20] We chose the unigram and trigram models to assess the importance of word sequences.

5 Preliminary Results from Automated Speech Recognition and Their Analysis

The WER of three models was from 64% to 89% (Table 1). This was large compared with those reported by major commercial ASR transcription services; however, this comparison requires interrogation.

The WER of the large commercial services is typically related monolingual tasks, usually on English data, and outside the classroom context. In a monolingual Spanish classroom environment, an impressively small WER of 10% was reported using a tailored, commercial ASR system with test data of two 50-minute university lectures and one 50-minute seminar with 10–16 year-old students (Iglesias et al., 2016). In contrast, for monolin-

yses.

[15] e.g. laughter, outbreaths, unintelligible whispers

[16] incl. the Indonesianised names

[17] duration of approx. 1 minute

[18] i.e., the number of deleted, inserted, and substituted words, divided by the total number of words

[19] No differentiation was made by the Kaldi toolkit between languages.

[20] i.e., a sequence of n words

gual US English-speaking teachers' speech, a WER from 44% to 100% was reported for five ASR systems, which were free of cost to use and required no additional supervised learning to train the ASR model (Nathaniel et al., 2015).

In our results, we analyzed teacher speech phenomena, such as emphasized articulation. For example, an instance of 'sma', an acronym for a senior high school produced as the Indonesian names of the letters, [es em ah] was hyper-articulated. Our Bilingual_3G and Indonesian_3G models produced reasonable approximations: 'ah sma aha' and 'hasan aha', respectively. Given the variation this sort of phenomena introduced into lexical items, teacher speech characteristics seem likely to have impacted our ASR performance.

ASR performance degrades in multilingual settings, but a range of techniques for reducing WER are available (see Yilmaz et al. (2016); Nakayama et al. (2018); van der Westhuizen and Niesler (2019); Yue et al. (2019)). Many of these studies note their shortage of training data and some report success in using training data from high resource languages to work with low resource languages. For example, Biswas et al. (2018) experimented with a new South African soap opera corpus in which five languages were present and found that the incorporation of monolingual, out-of-domain training data reduced their WER. Working with the same corpus, Biswas et al. (2019) first trained bilingual systems and a unified five-lingual system, and then experimented with adding convolutional neural network layers to these models. Overall they achieved WERs ranging from 43% to 64% for paired languages and from 26% to 78% for single languages.

While performance between different language pairs might not be suitable for comparison due to the interplay of language typologies interacting in distinctive ways, WERs from codeswitch bilingual data were more similar to our WER, especially given our small amount of training data. Yeong and Tan (2014) studied Indonesian, Iban, and Malay codeswitching in written work, however, to the best of our knowledge, our work was the first work on spoken Indonesian–English data.

WER rates were useful in relating our results with the overall progress being made in ASR, but given our goal to expedite human transcription, for us, it was more fruitful to analyze the number and length of correctly recognized text spans in the ASR-based transcription. We theorized that these tools could begin to change workflow or decrease cognitive load for human transcribers by generating a draft transcript for revision.

The two 4-word and one 3-word correct text spans produced by our Bilingual_1G model would probably be the most useful in speeding transcription (Table 1). However, the preliminary results produced by the Indonesian_3G model were comparable to the two bilingual models. This was impressive given that nearly 50% of the test data was in English. Supposing a research interest in only the Indonesian spoken by the teacher, or the use of an English language model for the other data, the Indonesian model could reasonably be assessed as scoring 22 correct words from the 51 Indonesian words in the test data.

Proceeding to a more detailed study of the performance of the models, we undertook an error analysis to elucidate the type of errors occurring. We analyzed them as segments, from multiple perspectives (Figure 4). There was a high incidence of resyllabification[21] in the machine transcription, as words were split, concatenated with the preceding or succeeding word(s), a middle consonant was omitted, and/or an initial consonant was omitted. For example, 'perguruan' in the reference transcript and 'per keren' by ASR accumulated three errors: resyllabification and two counts of substitution.[22] Another example is, 'it' in the reference transcript produced as 'old' in the ASR output. This error was coded for a vowel change and consonant change.[23]

Given the small test set, using this error analysis, we made the tentative note that the Bilingual_3G model seemed slightly less likely to make errors of insertion and deletion, indicating that the errors were perhaps less 'disruptive' than the errors in the other models. Thus, despite the model's worse overall performance, it might improve rapidly with more training data.[24]

6 Discussion

As our principal result, we concluded that Kaldi, in conjunction with the Elpis interface, can expedite the transcription of teacher corpora. The time taken to transcribe speech can be extreme; in our project,

[21] word split

[22] consonant g > k and vowel change monophthong > diphthong

[23] monophthong > diphthong and t > d, respectively

[24] Our analysis of the Indonesian excluded all English words, reducing the test sample significantly.

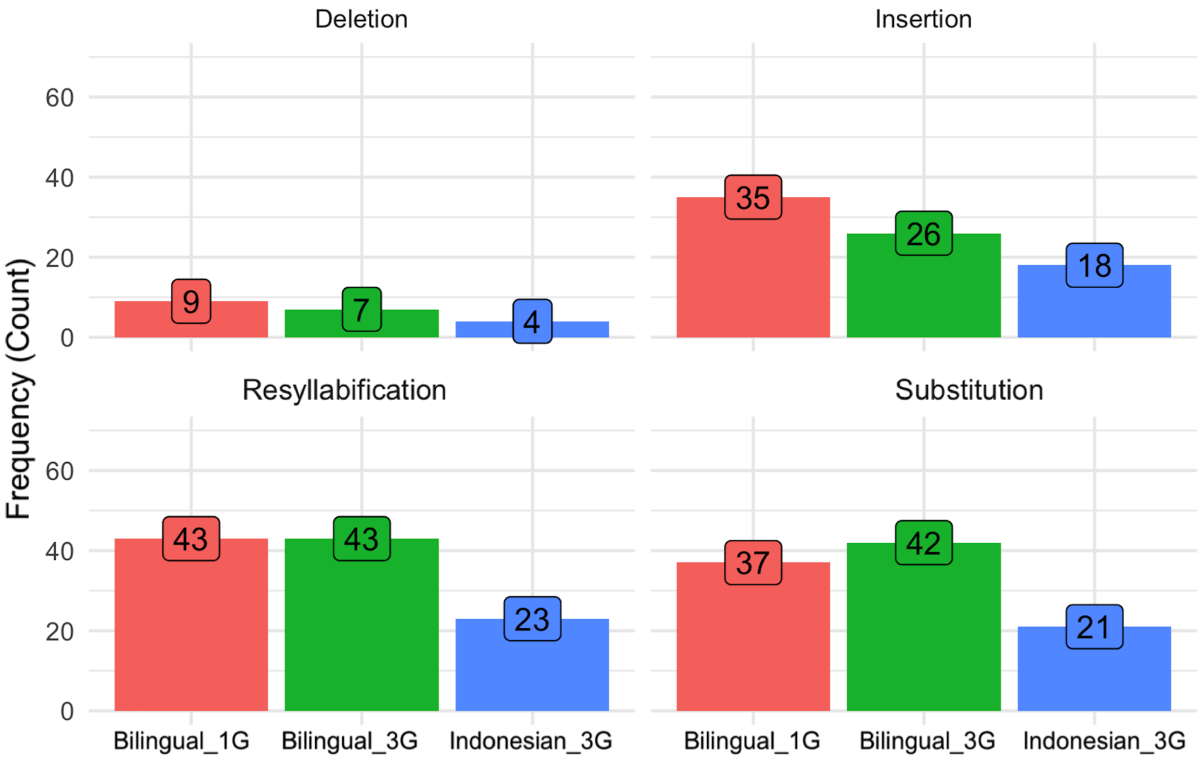


Figure 4: Segment analysis showing the frequency of error types at the phonemic level

transcribers spent months familiarizing themselves with the participants speech and setting up extensive transcription guidelines. Our final 51 minutes of test data took approximately $1,024$ minutes[25] to transcribe.[26] However, the use of Kaldi and Elpis was also time-consuming and required significant training and expertise. The continued development of Elpis may make the tool more viable for ASR-assisted transcription in research.

Our detailed discussion of methodological issues arising during human transcription of training data cannot prescribe a solution for all language teacher corpora; as Helm and Dooly (2017, p. 170) say of their own methodology paper examining the transcription of online language classroom data, their methods necessarily reflect "the research questions and the situated context of the study". However, we do hope to provide a baseline of discussion for those developing training datasets with this kind of complex speech. Similarly to Helm and Dooly (2017, p. 181), we hope to "highlight how we can try to be reflexive and critical in our research practices, increasing the transparency and accountability of our work and opening it up for discussion with others". This is especially pertinent as we develop machine learning-based technologies, which often lack transparency and trustworthiness (Pynadath et al., 2018).

Beyond the goals of this study, our findings contribute to expanding bodies of research into the use of ASR with small datasets (Gonzalez et al., 2018), in educational and classroom settings, as well as ASR of multilingual data. Our results gave some indication that while developing an initial (small) training dataset, using a simpler unigram model with less lexical information is better. Of course, ASR could be enhanced with a larger training dataset and supplementary text corpora from teaching resources.

Data loss was inevitable when we converted enacted classroom interactional phenomena into the linear, rather two-dimensional written format of orthographic transcription. This loss of complexity causes us to raise a cautionary flag; datasets produced through these methods can be used to support teacher reflection on their practice, but should never be taken as the entirety of a teacher's work and metrics derived from them should be viewed with a careful understanding of how much they reduce the complexity of the phenomena they record. Losing the context of data is not an "obscure problem apparent to a few philosophers focused on cy-

[25] i.e., 17 hours

[26] i.e., the turnaround of about 1:20

bernetics" (Bornakke and Due, 2018, p. 1). In this paper, we highlight the importance of decisions about 'what data to lose' when transcribing, making tactical decisions that are justified by research questions (Gangneux and Docherty, 2018), and how transcription bias can be multiplied in unknown ways by computational processes.

Investments are necessary to convert the tools we used to a useable workflow for practicing teachers. Elpis is likely to make significant headway in this area, but the complex nature of transcribing bilingual teaching data requires specialized skills. Training teachers to do this work seems a useful area of technological investment in languages education. It could incorporate established uses of teacher corpora for teacher training and professional development with new goals of elucidating the language input teachers provide in the classroom. Teachers who transcribe a small training dataset of their own speech may gain deep insight into their own language use. Using ASR to accelerate transcription could lead to teachers having the capacity to build larger datasets, analyze their own teaching, and thereby progress their practice. Given the workload issues often associated with teaching, asking teachers to transcribe their own lessons may be unrealistic in the initial development of this tool but could be more appropriate in teacher training settings where it could be included as part of their studies. Engaging with concerns in education about the use of teaching technologies as performance management tools (Page, 2017; Tolofari, 2005), this tool in teachers' hands could advance action research and protect teachers from it being used as a supervision/performance management tool.

7 Conclusion

Having trained and applied ASR in the form of Kaldi and Elpis to a dataset of carefully prepared Indonesian language teaching data, it is clear that the applicability of these technologies is limited with such a small set of training data. Yet, further investigation and development toward the goal of expedited transcription is warranted because of the virtuous cycle of ASR-assisted human-workflow.

The limitations and risks of these technologies must be considered if we hope to use them to gain real insight into the practice of language teachers. However, it is crucial that education is not excluded from technological advances. Empirical information about teacher practice for teachers, curriculum writers, educational researchers, and policy makers could be used to inform and advance the education sector the same way as these computational advancements are already routinely used in industry and other sectors.

Acknowledgments

We gratefully acknowledge the contribution of anonymous reviewers, colleagues, and study participant to this paper.

References

Claudia Baur, Andrew Caines, Cathy Chua, Johanna Gerlach, Mengjie Qian, Manny Rayner, Martin Russell, Helmer Strik, and Xizi Wei. 2018. Overview of the 2018 spoken call shared task. In *Interspeech 2018*, pages 2354–2358.

Astik Biswas, Ewald van der Westhuizen, Thomas Niesler, and Febe de Wet. 2018. Improving asr for code-switched speech in under-resourced languages using out-of-domain data. In *SLTU 2018*, pages 122–126.

Astik Biswas, Emre Yilmaz, Febe de Wet, Ewald van der Westhuizen, and Thomas Niesler. 2019. Semi-supervised acoustic model training for five-lingual code-switched asr. In *Interspeech 2019*.

BNC-Consortium. 2007. *Reference Guide for the British National Corpus (XML Edition)*.

Hilda Borko, Carol Livingston, and Richard J. Shavelson. 1990. Teachers' thinking about instruction. *Remedial and Special Education*, 11(6):40–49.

Tobias Bornakke and Brian L. Due. 2018. Big–thick blending: A method for mixing analytical insights from big and thick data sources. *Big Data & Society*, 5(1).

Erman Boztepe. 2003. Issues in code-switching: Competing theories and models. *Issues in CS: Competing Theories and Models*, 3(2).

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

Dwi Noverini Djenar. 2008. Which self? pronominal choice, modernity, and self-categorizations. *International Journal of the Sociology of Language*, 189:31–54.

Dwi Noverini Djenar and Michael C. Ewing. 2015. Language varieties and youthful involvement in indonesian fiction. *Language and Literature*, 24(2):108–128.

John W. Du Bois, Stephan Schuetze-Coburn, Susanna Cumming, and Danae Paolino. 1993. Outline of discourse transcription. In Jane A. Edwards and Martin D. Lampert, editors, *Talking data: Transcription and coding in discourse research*, pages 45–89. NJ: Erlbaum, Hillsdale.

Gautier Durantin. 2017. Early results from survey exploring transcription processes.

Nick C. Ellis. 2017. Cognition, corpora, and computing: Triangulating research in usage-based language learning. *Language learning*, 67(1):40–65.

Ben Foley, Josh Arnold, Gautier Durantin, T. Mark Ellison, Daan van Esch, Scott Heath, Nicholas Lambourne, Zara Maxwell-Smith, Ola Olsson, Aninda Saha, Nay San, Hywel Stoakes, and Janet Wiles. 2018. Building speech recognition systems for language documentation: The coedl endangered language pipeline and inference system (elpis). In *The 6th International Workshop on Spoken Language Technologies for Under-Resourced Languages*.

Christian Gaida, Patrick Lange, Rico Petrick, Patrick Proba, Ahmad Malatawy, and David Suendermann-Oeft. 2014. Comparing open-source speech recognition toolkits. In *NLPCS 2014 : 11th International Workshop on Natural Language Processing and Cognitive Science*.

Justine Gangneux and Stevie Docherty. 2018. At close quarters: Combatting facebook design, features and temporalities in social research. *Big Data & Society*, 5(2).

Ofelia Garcia and Li Wei. 2014. *Translanguaging: Language, Bilingualism and Education*. Palgrave Macmillan, New York.

Zane Goebel. 2010. Identity and social conduct in a transient multilingual setting. *Language in Society*, 39(02):203–240.

Zane Goebel. 2014. Doing leadership through signswitching in the indonesian bureaucracy. *Journal of Linguistic Anthropology*, 24(2):193–215.

Simòn Gonzalez, Catherine E. Travis, James Grama, Danielle Barth, and Sunkulp Ananthanarayan. 2018. Recursive forced alignment: A test on a minority language. In *17th Speech Science and Technology Conference*.

Google. 2019. Google cloud speech-to-text api language support.

Nelofer Halai. 2007. Making use of bilingual interview data: Some experiences from the field. *The Qualitative Report*, 12(3):344–355.

Francesca Helm and Melinda Dooly. 2017. Challenges in transcribing multimodal data: A case study. *Language Learning & Technology*, 21(1):166–185.

Julia Hirschberg and Christopher D. Manning. 2015. Advances in natural language processing. *Science*, 349(6245):261–266.

Eduard Hovy and Julia Lavid. 2010. Towards a 'science' of corpus annotation: A new methodological challenge for corpus. *International Journal of Translation*, 22(1):13–36.

Ana Iglesias, Javier Jimenèz, Pablo Revuelta, and Lourdes Moreno. 2016. Avoiding communication barriers in the classroom: the apeinta project. *Interactive Learning Environments*, 24(4):829–843.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing (2nd Edition)*. Prentice-Hall, Inc, Upper Saddle River, NJ, USA.

Judith C. Lapadat and Anne C. Lindsay. 1998. Examining transcription: A theory-laden methodology. Report, Social Sciences and Humanities Research Council of Canada, Ottawa (Ontario).

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

Sahoko Nakayama, Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2018. Speech chain for semi-supervised learning of japanese-english code-switching asr and tts. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 182–189.

Blanchard Nathaniel, Michael Brady, Andrew M. Olney, Marci Glaus, Xiaoyi Sun, Martin Nystrand, Borhan Samei, Sean Kelly, and Sidney D'Mello. 2015. A study of automatic speech recognition in noisy classroom environments for automated dialog analysis. In *Artificial Intelligence in Education*, pages 23–33, Cham. Springer International Publishing.

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

Bonny Norton and Kelleen Toohey. 2011. Identity, language learning, and social change. *Language Teaching*, 44(4):412–446.

Nuance. 2019. Nuance recognizer language availability.

John Paul O Dela Rosa and Diana C. Arguelles. 2016. Do modification and interaction work? – a critical review of literature on the role of foreigner talk in second language acquisition. *i-Manager's Journal on English Language Teaching*, 6(3):46–60.

Damien Page. 2017. Conceptualising the surveillance of teachers. *British Journal of Sociology of Education*, 38(7):991–1006.

Alastair Pennycook. 2016. Posthumanist applied linguistics. *Applied Linguistics*, 39(4):445–461.

Luke Plonsky. 2013. Study quality in sla: An assessment of designs, analyses, and reporting practices in quantitative l2 research. *Studies in Second Language Acquisition*, 35(4):655–687.

David V. Pynadath, Michael J. Barnes, Ning. Wang, and Jessie Y.C. Chen. 2018. Transparency communication for machine learning in human-automation interaction. In J. Zhou and F. Chen, editors, *Human and Machine Learning*, pages 75–90. Springer, Cham.

Kazuya Saito and Kim van Poeteren. 2012. Pronunciation-specific adjustment strategies for intelligibility in l2 teacher talk: results and implications of a questionnaire study. *Language Awareness*, 21(4):369–385.

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

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

Sowaribi Tolofari. 2005. New public management and education. *Policy Futures in Education*, 3(1):75–89.

Maria Teresa Turell and Melissa G. Moyer. 2009. Transcription. In Li Wei and Melissa Moyer, editors, *The Blackwell Guide to Research Methods in Bilingualism and Multilingualism*, book section 11. Blackwell Publishing Ltd, Carlton, Australia.

Etienne Wenger. 1998. *Communities of Practice: Learning, Meaning, and Identity*. Cambridge University Press, Cambridge.

Ewald van der Westhuizen and Thomas R Niesler. 2019. Synthesised bigrams using word embeddings for code-switched asr of four south african language pairs. *Computer Speech & Language*, 54:151–175.

Yin-Lai Yeong and Tien-Ping Tan. 2014. Language identification of code switching sentences and multilingual sentences of under-resourced languages by using multi structural word information. In *Interspeech 2014*, pages 3052–3055.

Emre Yilmaz, Henk van den Heuvel, and David Van Leeuwen. 2016. Investigating bilingual deep neural networks for automatic speech recognition of code-switching frisian speech. In *5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU 2016*, volume 81, pages 159–166.

Xianghu Yue, Grandee Lee, Emre Yilmaz, Fang Deng, and Haizhou Li. 2019. End-to-end code-switching asr for low-resourced language pairs. In *IEEE ASRU Workshop 2019*.

Appendices

A Code Versions

This study used custom scripts for data preparation, along with the Elpis and Kaldi software projects to train and apply ASR models.

Elpis is a wrapper for the Kaldi speech recognition toolkit. At the time of the study, Kaldi was at version 5.5.

The version of Elpis used was v0.3 of the kaldi_helpers code. The current version of Elpis can be accessed at github.com/CoEDL/elpis. DOI: 10.5281/zenodo.3833887

We have released the data preparation scripts under Apache License Version 2.0. These scripts handled the generation of pronunciation lexicons required by Kaldi, and the preparation of audio files for inferencing. The scripts can be downloaded by following the respective DOIs. Version 0.1 of kaldi-helpers-pron-lexicon was used to identify and make pronunciations for English and Indonesian words. DOI: 10.5281/zenodo.3835586. A script, released as kaldi-helpers-segment-infer v0.1, was written to segment long audio into shorter segments as required by Elpis v0.3. DOI: 10.5281/zenodo.3834016