Fast Bootstrapping of Grapheme to Phoneme System for Under-resourced Languages - Application to the Iban Language

Sarah Samson Juan, Laurent Besacier
Laboratoire d'Informatique de Grenoble / Grenoble University
Grenoble, France
`sarah.samson-juan@imag.fr`, `laurent.besacier@imag.fr`

Abstract

This paper deals with the fast bootstrapping of Grapheme-to-Phoneme (G2P) conversion system, which is a key module for both automatic speech recognition (ASR), and text-to-speech synthesis (TTS). The idea is to exploit language contact between a local dominant language (Malay) and a very under-resourced language (Iban - spoken in Sarawak and in several parts of the Borneo Island) for which no resource nor knowledge is really available. More precisely, a pre-existing Malay G2P is used to produce phoneme sequences of Iban words. The phonemes are then manually post-edited (corrected) by an Iban native. This resource, which has been produced in a semi-supervised fashion, is later used to train the first G2P system for Iban language. As a by-product of this methodology, the analysis of the "pronunciation distance" between Malay and Iban enlighten the phonological and orthographic relations between these two languages. The experiments conducted show that a rather efficient Iban G2P system can be obtained after only two hours of post-edition (correction) of the output of Malay G2P applied to Iban words.

1 Introduction

Multilingualism is at the heart of current issues relating to cultural, economic and social exchanges in a globalized world. Thus, people are more likely to evolve in multilingual environments, as evidenced by recent trends of world and society: increasing importance of international organizations (multilateral organizations like European Union, multinational corporations, etc.), increase of cultural exchanges and travel, extensive use of social networks to communicate with people around the world, renewed interest in regional languages and dialects which now coexist with the national languages. Such "language diversity" must be taken into account. Moreover, it is known that among the most widely spoken languages in the world, many are those for which the technologies for written and spoken natural language processing are poorly developed (under-resourced languages). There is a commercial interest in enabling the ~300 most widely spoken languages in the digital domain: if digital technologies work for this group of languages that represents 95 % of humanity. As mobile phones are nearly ubiquitous (87 % penetration worldwide) and global internet access is approaching 1/3 of the human population, enabling "the long tail of languages" in the digital domain increasingly matters. The other ~6,500 languages are not of commercial interest, but there are other reasons to enable them if possible: to provide access to information, to provide a critical new domain of use for endangered languages, for better linguistic knowledge of them, for response in a crisis (surge languages), etc.

In this paper, our work concentrates on Grapheme-to-Phoneme (G2P) conversion system, which is an important component for both automatic speech recognition (ASR) and text-to-speech synthesis (TTS). We proposed a quick and plausible solution in developing a G2P system for an under-resourced language using existing G2P system with a language of the same family. More precisely, a pre-existing G2P of a local dominant language, Malay, is utilized to generate phoneme sequences for Iban, a language from the same family. The outputs are further post-edited manually by an Iban native to get the right sequences. The post-edited transcript is later used to train the first Iban G2P system. An alternative approach is also studied where we experiment on Phoneme-to-Phoneme (P2P) system which translates from

Malay pronunciation (output of a Malay G2P) to Iban pronunciation.

This paper is organized as follows: section 2 describes the languages involved in our work and their relationship is detailed in section 3. Section 4 presents our methodology for fast bootstrapping of Iban G2P system with the help of pre-existing Malay G2P and details of the P2P as well as our experimental results of this study. Last but not least, section 5 concludes this work and provides some perspectives.

2 Malay and Iban Languages

Malay and Iban languages are both spoken in Malaysia. The latter is mostly spoken only by the Iban community while Malay language has many speakers as it is the country's national language. In the following section, we briefly describe about the two languages and the resources that are currently available for our research.

2.1 Malay language and its G2P system

Globally, Malay is not only spoken in Malaysia but also in several other neighboring countries such as Brunei, Indonesia, Singapore and southern Thailand. Although it is spoken across several countries, each country has its own standard of pronunciation and spelling. For our study, we only focus on the Malay language spoken in Malaysia. This language is written using Latin alphabet and considered as an agglutinative language. An agglunative language has words that are formed by adding affixes onto a root word, word composition or reduplication (Rainavo-Malançon, 2004). Furthermore, it is not a tonal language like Mandarin or Vietnamese languages and basically, language users can distinguish general pronunciations directly from the grapheme sequences.

The Malay data that we applied in this study was collected by Universiti Sains Malaysia in Malaysia, which was used to design a Malay speech recognition system. Tan et al. (2009) collected Malay texts from 1998-2008 articles concerning economy, entertainment, sports and general news. The authors built a pronunciation dictionary and language model for the Malay ASR using these texts. From our observation, a total of 76.05K pronunciations was available for 63.9K distinct Malay words (36 different phonemes are used to transcribe Malay words).

To build our Malay G2P system, we utilized a data driven G2P toolkit called Phonetisaurus (Novak, 2012) using the Malay data as our training set. The open source toolkit is able to perform Expectation Maximization (EM)-based alignments between grapheme and phoneme sequences; it also uses a "target" N-gram language model (made up of phone sequences). Models are converted to weighted finite states tranducers (WFST) by Phonetisaurus for decoding.

Our Malay G2P model was constructed from 68K Malay pronunciations taken from the 76K total set (8.05K remain for testing). In our case, the target N-gram model uses $N = 7$ where the model was generated using original Kneser-Ney smoothing utilizing the SRILM toolkit (Stolcke, 2002). This model was chosen among several N-gram models where N values ranged from 3 to 7 and the 7-gram model gave us the lowest phoneme and word error rates compared to the rest.

We tested with the remaining 8.05K data to measure the accuracy of the Malay G2P system and the results are 6.20% phoneme error rate (PER) and 24.98% word error rate (WER). From this point, this system became the starting point before the development of the second language phonetizer, the Iban G2P system. First, we briefly explain about the target language and the preliminary text data available in the next part of this paper.

2.2 Iban language and its initial text resource

2.2.1 Brief background

The Iban language is mainly spoken among the Iban community, one of Sarawak's indigenous group in Malaysia. Sarawak is the biggest state in Malaysia and has more than 690,000 Ibans living across the region (Statistics-Department, 2010). Like Malay, the Iban language belongs to the Malayo-Polynesian branch of the Austronesian language family and Iban falls in the Ibanic language group (Lewis et al., 2013). Speakers can also be found in several parts of the Borneo Island such as in Kalimantan, Indonesia; however, we limit our focus to Iban system from Sarawak. Since the early 90s, schools in the region teaches Iban in the primary and secondary level as a nonobligatory subject. The teaching effort has also recently spread to the university level, where several universities open basic Iban courses for undergraduate students. Despite the fact there are many Iban speakers, resources for

human language technologies (HLTs) are still very limited. Thus, we view this language as a very under-resourced language for technology applications.

2.2.2 Text resources

We began an Iban text collection campaign in November 2012: a total of 7000 articles was extracted from a local newspaper through its website. Articles dated from 2009 (the year when Iban articles started to be published online) to 2012, were mostly on general news, entertainment and sports. These articles were compiled and normalized. The steps in the text normalization process included removing HTML tags, changing numbers to words, converting commonly used abbreviations such as Apr. for April, Dr. for *Doktor* , Kpt. for *Kapten* and Prof. for *Profesor*, spliting paragraphs to sentences and removing punctuation marks except for '-'. For the latter step, we treated words "tied" with '-' as a single item. In Iban as well as in Malay, these words are categorized as reduplication or *jaku pengawa betandu penuh* in Iban. It is common that words are duplicated to form plurals and new words. Plurals are, for example, *bup-bup* or *rumah-rumah* or new words, such as, *bebai-bai, diuji-uji, ngijap-ngijapka* and *beberap-berap* where these words are categorized as partial reduplication. The final normalization step was to convert all capital words to lower case. After the text normalization process completed, we obtained approximately 2.08 M words. Based on this newly acquired corpus and the Malay corpus, we carried out a study on the relationship between Iban and Malay languages.

3 Malay-Iban relationship

3.1 Phonology

According to a reference manual written by the Centre of Curriculum and Sarawak Education Department (Education-Department, 2007), the Iban system is said to be influenced by the Malay system in terms of phonology, morphology and syntax. Omar (1981) provided the first description of the language in 1981 where among her work included the classifications of phonemes for Iban. There are 19 consonants (including semivowels), 6 vowels and 11 vowel clusters. Meanwhile for Malay, Tan et al. (2009) referred to Maris (1979)'s classification of Malay sounds. Based on Maris' work, there are 27 consonants, 6 vowels and 3

diphthongs. From the descriptions made by Omar and Maris, we carried out a comparative study between Malay and Iban phonemes.

Iban vs. Malay	Phonemes
Common	Consonants: p, b, m, w,t,d,n,tʃ,dʒ, s,l,r, ɲ, j, k, g, ŋ,h,ʔ Vowels : a,e,ə,i,o,u V. Clusters: ai, au
Difference	(only appear in Malay) Consonants: f,v,θ, z, x, ɣ, ð, ʃ (only appear in Iban) V. Clusters : ui,ia, ea,ua,oa,iu,iə, uə,oə

Table 1: *Iban and Malay common and different phonemes*

In Table 1, we present the common and different Malay and Iban consonants, vowels and vowel clusters. It is observed that Iban is a subset of Malay consonants and vowels. However, the same could not be said for Iban vowel clusters. These vowel clusters appear as variations in Iban pronunciations and we found that only two of the Iban vowel clusters, /ai/ and /au/ matched with Malay diphthongs, /aj/ and /aw/, respectively. Our next example of Iban words (refer to Table 2) also show that the missing vowel clusters are not expressed in the spellings (grapheme sequences). Hence, it is clear that a Malay G2P will produce incorrect phoneme sequences for Iban words due to the missing phonemes.

Vowel clusters	Phoneme and grapheme sequences
/ai/	/kumbai/ ∼ kumbai
/ui/	/ukui/ ∼ ukui
/ia/	/kiaʔ/ ∼ kiak
/ea/	/rumeah/ ∼ rumah
/ua/	/kuap/ ∼ kuap
/oa/	/menoa/ ∼ menua
/iu/	/niup/ ∼ niup
/au/	/tawn/ ∼ taun
/iə/	/biliəʔ/ ∼ bilik
/uə/	/puən/ ∼ pun
/oə/	/boəʔ/ ∼ buk

Table 2: *Iban vowel clusters with the related grapheme and phoneme sequences*

3.2 Orthography

Orthography relates to the standard writing system for a particular language. Both Malay and Iban are written using latin alphabets and their orthography system is closely related. Ng et al. (2009) investigated their orthographic similarity i.e; finding cognates and non-cognates between Malay and Iban. They applied several methods such as the Levenshtein distance (Levenshtein, 1966) in order to estimate the distances on 200 word pairs. Chosen words were translations of a Swadesh list (Swadesh, 1952), which is a common reference created for linguists to study relationships between languages. Ng et al. (2009) discovered that Iban has the highest cognate percentage of 61% with Malay compared to other Sarawak languages like Kelabit, Melanau and Bidayuh. From this observation, we investigated, at a larger scale, the pronunciation similarity between Iban and Malay words that have the same surface form.

There are 36,358 distinctive words in our Iban texts. From this lexicon, we identified words (surface forms) that match in both Malay and Iban lexicons. Table 3 shows that more than thirteen thousand words are shared between Malay, Iban and also surprisingly English. After removing words included in the CMU English pronunciation dictionary, the number of Malay-Iban common surface forms is 8,472 words.

Corpus	Vocab. size	Identical words	
		with English	w/o English
Malay	76,050	13,774	8,472
Iban	36,358		

Table 3: *Number of identical (same surface form) words found in our Iban and Malay lexicons*

Conclusively, 23% (8,472) of our Iban vocabulary is shared with Malay, 19% (6,707) with English, while the remaining 58% (21,179) purely belong to the Iban language. In other words, 42% of this lexicon is found shared not only with one, but, with two languages, English and Malay. This gave us an idea of language contact and code switching issues related to Iban language.

3.3 Measuring Malay-Iban pronunciation distance

To measure pronunciation distances, our concern was on Malay-Iban common surface forms only.

The purpose was to study what was the minimum cost to transform a Malay phoneme sequences to an Iban one. We chose Levenshtein distance as the estimation method following the study by Heeringa and de Wet (2008). To carry out this investigation, we selected 100 most frequent common surface forms and prepared the pronunciation transcripts. Malay G2P was employed to generate an initial pronunciation transcript for Iban and the transcript was later post-edited by a native speaker.

The Levenshtein distance was computed for all 100 pronunciation pairs. As a result, we obtained 17% of errors (phoneme Insertions, Substitutions, Deletions) between Malay and Iban pronunciations but we found out that 47% of the Malay pronunciations were kept unchanged for Iban! This result confirms that the use of a Malay G2P is probably a good starting point to bootstrap an Iban G2P system. To put this result in perspective, we quote the work of Heeringa and de Wet (2008) who measured the average distance between Afrikaans and Dutch pronunciations and found that it was significantly smaller than between Afrikaans and Frisian ; as well as between Afrikaans and German.

No.	Words	Iban	Malay
1	ke	/kə/	/kə/
2	nya	/ɲaʔ/	/ɲə/ or / ɲa/
3	iya	/ija/	/ija/
4	ba	/baʔ/	/ba/
5	dua	/duwa/	/duwə/ or /duwa/
6	sida	/sidaʔ/	/sida/
7	puluh	/puluəh/	/puloh/
8	raban	/raban/	/raban/
9	lalu	/lalu/	/lalu/
10	orang	/urang/	/orang/

Table 4: *Example of ten words with phoneme sequences for Iban and Malay*

Table 4 shows an example of words and their corresponding Malay / Iban phoneme sequences. We analyzed several phonemes that were frequently substituted and inserted in the Malay-to-Iban transformation process. The phoneme /o/ is frequently substituted with /uə/ , for example, the word *puluh* in Table 4 is transcribed as /puloh/ in Malay while in Iban, it is /puluəh/. This substitution occurred due to the vowel cluster /uə/ in Iban utterance. Also, we found out that phoneme /e/ was frequently substituted by /iə/ in sequences

such as /pəsiser/ in Malay to /pəsisiər/ in Iban for the word *pesisir*. The glottal stop /ʔ/ was inserted at almost all words ending with a vowel. As an example, *kepala* transcribed as /kəpala/ in Malay needs a glottal stop at the final vowel to transform to /kəpalaʔ/ and another example, *nya*, changes from /ɲa/ to /ɲaʔ/.

To summarize, this preliminary study on Malay-Iban pronunciation distance suggested that the Malay G2P system can be used as a basis for transcribing Iban words. The suggestion was also supported by the closeness of these two languages based on phonological and orthographical aspects particularly for Malay-Iban cognates. However, we need to investigate the Malay G2P performance on non-related / non-common words specifically, the "pure" Iban words in the lexicon. The strategy is detailed and experimented in the following section.

4 Obtaining Iban G2P training data via post-editing the Malay G2P output

4.1 Methodology

Our proposed methodology involves the following process:

- Choose two different development sets of common Malay-Iban and pure Iban words, from the Iban vocabulary.

- Apply Malay G2P model on each set to obtain initial phoneme sequences.

- Post edit Malay G2P outputs to produce correct Iban pronunciations and measure time taken to complete this phase.

- Train Iban pronunciations from previous step as a first Iban G2P system (using Phonetisaurus)

- Run and evaluate both Malay and Iban G2P systems on new Iban test set

We chose two sets of 500 most frequent Iban words (500 common Malay-Iban and 500 pure Iban words) and then applied our Malay G2P system to convert words into phoneme sequences. Then, a native speaker manually edited the output to obtain correct sequences for Iban. Moreover, we identified phonemes that were frequently substituted or inserted during the process and measured the phoneme and word (sequence) error

rates using a scoring toolkit by the National Institute of Standards and Technology (NIST, 2010). Thereafter, we combined the post-edited transcripts into one list and trained our first Iban G2P system.

Upon achieving this, we selected another two sets of different words from the Iban vocabulary to evaluate the Malay G2P again and the new Iban G2P system. The two sets contain second 500 most frequent Iban words (500 common Malay-Iban and 500 pure Iban words).

4.2 G2P output evaluation

Phonetizer	Corpus	PER (%)	WER (%)	Post-edit (mins)
Malay G2P	500_{IM}	6.52	27.2	30
	500_I	15.8	56.0	42
Iban G2P	500_{IM}	13.6	44.2	45
	500_I	8.2	31.8	32
Iban P2P	500_{IM}	16.6	53.5	-
	500_I	7.3	31.9	-

Note: IM for common Malay-Iban words and I for pure Iban words

Table 5: *Malay G2P and Iban G2P systems (+ Iban P2P) peformance for an Iban phonetisation task*

Our first 1000 sequences generated by the Malay G2P scored at 11.88% PER and 48.9% WER. The scoring was based on the post-edited transcript that was completed by the native within 1 hour and 34 minutes. Now, Table 5 presents the evaluation results of our Iban G2P output and of Malay G2P, for comparison which was done on a second 1000 words data set. The test sets, 500_{IM} and 500_I, were taken from the second most frequent items in the Iban lexicon. The values presented are phoneme error rate (PER), word error rate (WER) and post-editing effort in minutes. Based on these results, we discovered that the Malay G2P system performed best for Malay-Iban matching words and the result is consistent with Malay G2P performance on Malay test set. On the other hand, the Iban G2P system gave better results for pure Iban words. Despite of the small amount of data used to train the Iban G2P (1000 sentences), it was able to perform less than 10% PER. The time spent to correct Iban G2P output was also less for pure Iban words compared to the post-editing effort for Malay G2P output. However, the Iban G2P system seems to be not suitable

to phonetize common Malay-Iban words (PER increases from 6.52 % to 13.6 %).

We examined in detail on each G2P outputs to find wrongly substituted and deleted phonemes. In the case of pure Iban words test results, Malay G2P substituted phonemes /uə/, /iə/, /ea/, with /o/, /e/ and /a/, respectively, while the glottal stop /ʔ/ and phoneme /r/ were missed out. Meanwhile, Iban G2P substituted phonemes /ə/ for /e/, /iə/ for /i/, /uə/ for /u/ and /uə/ for /o/. We also found that the glottal stop was inaccurately inserted. As for the Malay-Iban common words results, similar phonemes were wrongly predicted by both G2P systems. However, Malay G2P conversion was more accurate compared to Iban G2P because many original phoneme sequences were retained for Iban due to Malay word adoption (e.g; words such as *parlimen* (parliament), *menteri* (minister) and *muzik* (music)). For pure Iban words, Iban G2P gave better sequences because it included vowel clusters (combined Malay phonemes) that were missing in the initial Malay G2P.

4.3 Converting pure Iban words using P2P system

Apart from the phonetization tasks by G2P models, we also developed P2P system and conducted phonetization tasks using P2P phonetizer. Recalling our Malay G2P outputs which was later post-edited to get data for Iban G2P system, we took these outputs and the corrected pronunciations as the training corpus for an Iban P2P system.

For experimentation, the 1000 pairs of phoneme sequences were randomized and then divided into 10 portions. Later, we built ten systems with different training data sizes (add one portion to training set after each model developed) and evaluated the systems on pure Iban words.

All phone error rates acquired from applying the G2P and P2P systems were plotted as shown in Figure 1 (see non-dotted line). The results are between 6.4% to 7.6% and found to be rather stable for the P2P model. Also on the same graph, we plotted PERs that were obtained by applying 10 Iban G2P systems of different training data sizes where each system had the same Iban phoneme sequences as in each P2P system. Based on our results, the G2P systems gave worse results compared to the P2P systems' results. Unlike the slightly unstable P2P results, phone accuracy improved gradually after adding more data in the

G2P. Finally, using 1000 words for training, Iban P2P and G2P systems' PER results are quite close and both systems have equal WER (31.9% - see also last line of table 5).

4.4 Phonetization of our whole Iban lexicon and final evaluation

Given results obtained, we decided to build the pronunciation lexicon for Iban using both G2P modules (Malay and Iban). The strategy was as follows: the Malay G2P phonetizes all Malay-Iban common words while the Iban G2P phonetizes all pure Iban words. In total, we phonetized 29,651 words (Iban-English not included) automatically. This lexicon can be used for further ASR or TTS system development. In addition to this phonetization task, we also developed a second lexicon using one of the P2P modules from previous experiments as explained in section 4.3. We chose the P2P that contains 1000 phoneme pairs (Malay-Iban phonemes) to phonetize only pure Iban words and we kept Malay G2P to phonetize Malay-Iban common words.

As a final evaluation, we post-edited 2000 random outputs taken from both lexicon. In this random set, there are 1426 pure Iban words and 574 Malay-Iban common words. The random outputs scored at 8.1% PER and 29.4% WER based on the G2P strategy, whereas the results from the P2P strategy are 10.2% PER and 38.1% WER.

Compared to our previous experimental results as shown in Table 5, our strategy to phonetize the Iban lexicon using G2Ps actually gave favourable outputs. Although the error rates are not the lowest, they are also not the worst. The PER falls in the range between 6.52% (Malay G2P score) and 8.2% (Iban G2P score) and WER in between 27.2% and 31.8%. Unfortunately, the P2P strategy returned lower accuracy values compared to the G2P strategy results. Table 6 presents a summary of the Malay and Iban phonetizers and their performances. In summary, the Iban G2P and P2P performances on Iban lexicon are a little worse than those of Malay G2P on Malay test set (6.2% PER; see Section 2.1).

5 Conclusions and future work

In this paper, we described the language contact between a local dominant language Malay and an under-resourced language from the same family, Iban. This study involved the comparison of
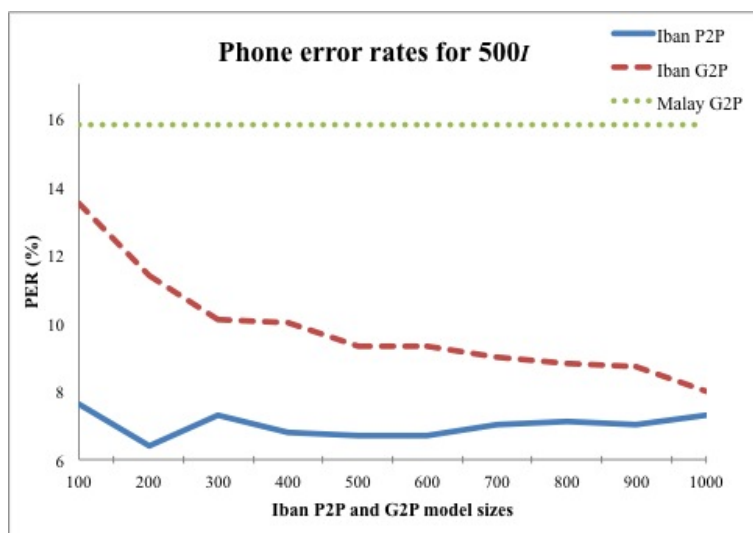
Figure 1: Phonetization results for 500 pure Iban words (*I*) obtained by employing different sizes of G2P and P2P models. Malay G2P (round dots) result is also plotted on the graph as our baseline study.

Phonetizer	#words	PER	WER
Malay G2P	8050 (Malay)	6.2	24.98
Iban G2P	2000 (Iban)	8.1	29.4
Iban P2P		10.2	38.1

Table 6: *Performance of Malay and Iban phonetizers. Measurement based on percentage (%) .*

Malay and Iban phonemes and pronunciation distance measurement on Malay-Iban common surface forms using the Levenshtein distance method. Due to our findings on the Malay-Iban connection, we built our first Iban G2P using post-edited Malay G2P output which, was done in less than 2 hours of manual post-editing by a native speaker.

Our preliminary results on two testing sets revealed that the two phonetizers, Malay and Iban G2Ps, are necessary to handle different word groups (Malay-Iban or pure Iban words). Thus, both Malay and Iban G2Ps were used in our attempt to produce the first Iban pronunciation lexicon. Besides that, we also employed Malay G2P and Iban P2P as our second strategy to obtain the lexicon.

To compare sample outputs with the post-edited list, we selected 2000 random phoneme sequences from the G2P and P2P outputs. We discovered that the G2P (results : 8.1% PER and 29.4% WER) is more accurate than the P2P. However, both Iban G2P and P2P performed lower than the Malay G2P.

As a continuation work on Iban phonetizers, our next research focus will be on tying Malay G2P and Iban P2P as a potential way to reduce the "knowledge gap" between Malay and Iban pronunciations. At the moment, the Malay G2P suits better for phonetizing Malay-Iban common words. While results on 2000 sample outputs from the Iban lexicon are not in favour of Iban P2P, the phonetizer did give higher accuracies compared to G2P phonetizers in our initial testings on pure Iban words. Hence, a suitable tying approach such as weighted finite states transducers, for example, could probably improve our phonetizer's accuracy. Besides that, our future goal is to build an Iban ASR system using Malay acoustic models and our new Iban pronunciation lexicon.

References

Sarawak Education-Department, 2007. *Sistem Jaku Iban di Sekula*. Sarawak, Malaysia, 1st edition.

W. Heeringa and F. de Wet. 2008. The origin of afrikaans pronunciation: a comparison to west germanic languages and dutch dialects. In *Proceedings of Conference of the Pattern Recognition Association of South Africa*, pages 159–164.

V. I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. In *Soviet Physics-Doklady*, volume 10, pages 707–710.

M. P. Lewis, Gary F. Simons, and Charles D. Fennig. 2013. Ethnologue : Languages of the world, sil international. available at : http://www.ethnologue.com.

Y. M. Maris. 1979. *The Malay Sound System*. Siri Teks Fajar Bakti, Kuala Lumpur.

Ee Lee Ng, Alvin Wee Yeo, and Bali Ranaivo-Malançon. 2009. Identification of closely-related indigenous languages:an orthographic approach. In *Proc. of International Conference on Asian Language Processing.*, number 230-235. IEEE.

NIST. 2010. Speech recognition scoring toolkit (sctk). available at : http://www.nist.gov/speech/tools/.

Josef R. Novak. 2012. Phonetisaurus: A wfst-driven phoneticizer. available at : https://code.google.com/p/phonetisaurus.

Asmah Haji Omar. 1981. Phonology. In *The Iban Language of Sarawak*, pages 16–41, Kuala Lumpur, Malaysia. Dewan Bahasa dan Pustaka.

Bali Rainavo-Malançon. 2004. Computational analysis of affixed words in malay language. In *Internal Publication, Universiti Sains Malaysia*.

Malaysian Statistics-Department. 2010. Negeri sarawak:total population by ethnic group, subdistrict and state. Technical report, Statistics Department, Malaysia.

Andreas Stolcke. 2002. Srilm - an extensible language modeling toolkit. In *Proc. of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*, pages 901–904.

Morris Swadesh. 1952. Lexico-statistic dating of prehistoric ethnic contacts. In *Proc. of the American Philosophical Society*, volume 96, pages 452–463.

T. P. Tan, H. Li, E. K. Tang, X. Xiao, and E. S. Chng. 2009. Mass: a malay language lvsr corpus resource. In *Proc. of 2009 Oriental COCOSDA International Conference*, pages 26–30.