HYBRID DISTANCE-STATISTICAL-BASED PHRASE ALIGNMENT FOR ANALYZING PARALLEL TEXTS IN STANDARD MALAY AND MALAY DIALECTS

Yen-Min Jasmina Khaw[1*], Tien-Ping Tan[2], and Ranaivo-Malançon Bali[3]

[1]Faculty of Information Communication and Technology
Universiti Tunku Abdul Rahman
Kampar, 31900, Malaysia

[2]School of Computer Sciences
Universiti Sains Malaysia
Penang, Malaysia

[3]Universiti Malaysia Sarawak
Kota Samarahan, Malaysia

Emails: khawym@utar.edu.my[1], tienping@usm.my[2], malanconbali@gmail.com[3]

Abstract

Parallel texts corpora are essential resources in linguistics and natural language processing, especially in translation and multilingual information retrieval. The publicly available parallel text corpora are limited to certain genres, types and domains. Furthermore, the parallel dialect text is scarce, even though they are important in the analysis and study of a dialect. Collecting parallel dialect text is challenging because dialects typically appear in the form of speech and very limited dialectic texts exist. Moreover, there is no standard orthography in most dialects. The contributions of this paper are threefold. First, the paper describes a methodology in acquiring a parallel text corpus of Standard Malay and Malay dialects, particularly Kelantan Malay and Sarawak Malay. Second, we propose a hybrid of distance-based and statistical-based alignment algorithm to align words and phrases the parallel text. The results show that the precision and recall values of the proposed alignment algorithm are more than 95% and better than the state-of-the-art GIZA++. Third, the alignment obtained were compared to find out the lexical similarities and differences between Standard Malay and the two studied Malay dialects, contributing valuable insights into the linguistic variations within the Malay language family.

Keywords: Malay dialects; Parallel text; Word alignment

1.0 INTRODUCTION

"Dialect" according to the Oxford dictionary is "a particular form of a language which is peculiar to a specific region or social group.". Dialectology compares and describes various dialects, or sub-languages, of a common language, which are used in different areas of a region [1]. Dialectometry, a sub-component of dialectology, is "the measurement of dialect differences, i.e. linguistic differences whose distribution is determined primarily by geography" [2]. The difference between a language and a dialect is sometimes difficult to state. Mutual intelligibility among speakers is a method used by linguists to distinguish between two different dialects of the same language and two different languages [3]. The speakers of different dialects of a language are able to understand each other because the dialects differ in systematic ways [4]. The systematic ways rely on phonological differences, lexical differences, and syntactic differences [4].

Many studies in dialect look at the phonological and phonetic differences between dialects. Heeringa [5] has proposed to measure the pronunciation differences of Dutch dialects using Levenshtein distance. A more focused work in studying the Dutch dialect variation is the proposition of a model based on articulography that measures the position of tongue and lips during speech [6]. Dialects can also vary in the writing. For instance, Wieling et al. [7] investigate the differences in lexical between Tuscan dialects that is spoken in the area of central Italy and standard Italian using

generalised additive mixed-effects regression model. On the other hand, Grieve [8] highlighted the regional variation in written American English, while Szmrecsanyi [9] explored the differences in grammars of British English dialects through a corpus of recorded speech using a dialectometrical analysis of their morpho-syntactic variation. All these systematic ways can group dialects into discrete classes or organise in a continuum or space without boundaries [10].

Malay is a good case study for dialectometry as it presents many dialects. In Malaysia, the formal Malay language used in the country is known as Standard Malay (SM). SM is from Johor, Riau dialect. The Johor, Riau Malay dialect is one of the main varieties used in Malaysia due to the influence and importance of Johor empire in the 19th century [11]. The Malay dialects in Malaysia can be grouped based on their geographical distribution [12]. Geographically, Malaysia is composed of two non-contiguous parts: Peninsular Malaysia (or West Malaysia) on the Malay Peninsula and East Malaysia on the island of Borneo. Peninsular Malay dialects have been classified differently in the literatures. For example, Onn [13] has proposed four basic groups: southern group, north-eastern group (Kelantan, Terengganu, and Pattani dialects), north-costal group, and Negeri Sembilan group. This work follows by the seven main groups proposed by Asmah [11] as shown in Fig. 1. In East Malaysia, there are Sabah and Sarawak dialects [12] as shown in Fig. 2. This paper investigates two dialects: Kelantan Malay dialect (KD) from Peninsular Malaysia, and Sarawak Malay dialect (SD) from East Malaysia. KD is one the most interesting Malay dialects because it is very different compared to other Malay dialects [14]. SD that is spoken in the largest state in Malaysia can be divided into sub-dialects [15, 16]. The one studied in this work is the SD spoken in Kuching, the capital city of Sarawak.

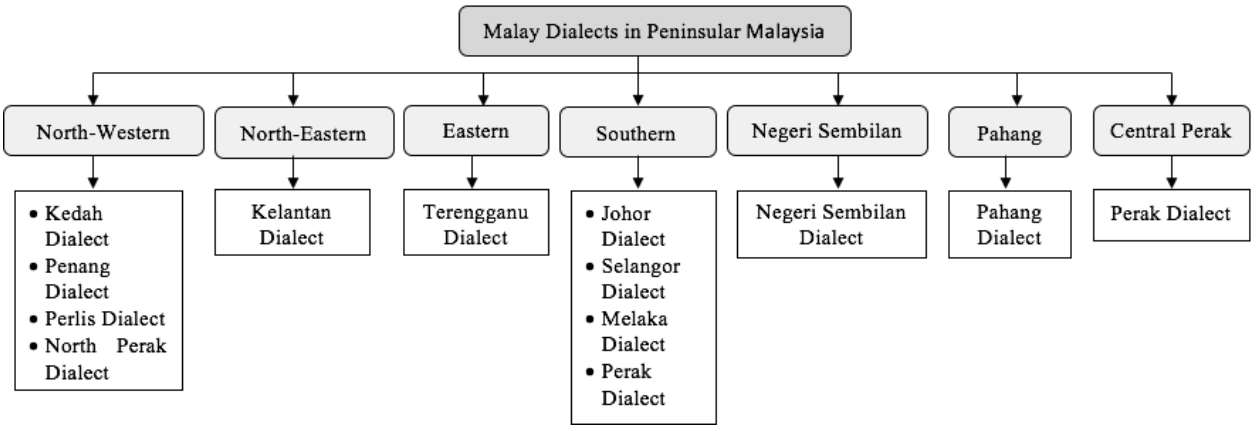

Fig. 1: Malay dialects in Peninsular Malaysia [11]

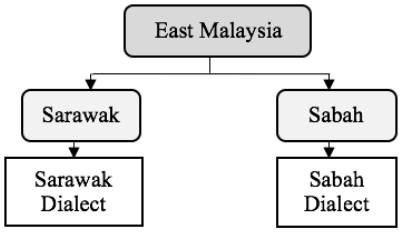

Fig. 2: Malay dialects in East Malaysia [12]

In Malaysia, most of the works in dialectometry focus on the phonology aspect. Asmah [17] proposed the classification of Malay dialects into two groups based on the pronunciation of the grapheme 'a' in word-final position as either [ə] or [a] (e.g. the word "I" or "me" in SM, *saya* can be pronounced as either [saja] or [sajə]). The schwa dialect group in which the letter 'a' is read as [ə] consists of the Malay dialects from the central, eastern, and southern parts of Peninsular Malaysia. The second group, called the a-variety group, consists of the dialects from the northern states of Peninsular Malaysia and dialects from the East Malaysia. In the a-variety group, the letter 'a' is pronounced as [a] with varying degree of openness. According to Asmah [17], the existence of the two main streams of varieties particularly in Peninsular Malaysia is due to the influence of two large Malay sultanates that are the Johor empire and the Kingdom of Kedah. Teoh [18] has analysed and compared the phonology of some of the dialects in schwa variety

and a-variety. He found that even the dialects within the same group can be very distinct. For example, SM and KD are both in the schwa group. Nevertheless, SM has six vowels, /i/, /u/, /e/, /ə/, /o/, and /a/ [19], while KD has fifteen vowels, /a/, /e/, /i/, /o/, /u/, /ə/, /ɛ/, /ɔ/, /ã/, /ẽ/, /ĩ/, /õ/, /ũ/, /ɛ̃/, and /ɔ̃/ [20]. Pronunciation of a word with nasalized vowel and oral vowel will give different meaning as in Table 1.

Table 1: Pronunciation of KD words with oral vowel and nasalised vowel [20]

Malay	***Nasalised Vowel***	***English***	***Malay***	***Oral Vowel***	***English***
kunci	/k u ʧ i/	key	kucing	/k u ʧ ĩ/	cat
pancur	/p a ʧ o/	jet of water	pancung	/p a ʧ õ/	behead
panggil	/p a ŋ e/	call	pangan	/p a ŋ ẽ/	food

Moreover, there are three diphthongs of /au/, /ai/ and /oi/ in SM, but KD does not have any diphthong [20]. SD has the same vowels and diphthongs as SM, but SD belongs to the a-variety group [17].

While most of the studies on Malay dialects focus on the phonology aspect, less has been put on the studies of the other aspects, for instance the writing of the dialect speakers. The only exception is SM, where the orthography [21], morphology [22], and grammars are well documented. Malay dialects do not have a standardized orthography. Native dialect speakers will write in native dialect using a combination of spelling based on SM and dialect. In this paper, we look at dialectometry from the perspective of writing, particulary in lexical differences. The study of the lexical differences is interesting because native speakers communicate also through writing, besides speech, often in social media such as Twitters and Facebook. A parallel text corpus would allow us to analyse the spelling and grammars between dialects, through the alignment of word or phrase in the parallel text. A parallel text corpus is a text in one language with its corresponding translations in another language. Parallel text corpora are essential resources in linguistics and natural language processing, especially in translation [23] and multilingual information retrieval [24]. However, the publicly available parallel corpora are limited to certain genres or areas, and there is a lack of such resources for the general domain [25]. The parallel text of dialects is even more scarce, albeit the resource is valuable in dialectometry and the development of dialect processing applications.

2.0 METHODS FOR BUILDING PARALLEL CORPUS

Today, many parallel corpora have been created for various purposes. These language resources are freely or commercially available. It happens often that the existing parallel corpora do not fit the requested purpose of certain users. In other context, the acquirer simply cannot afford to pay for the language resource. Therefore, the parallel corpus needs to be created from scratch. The parallel corpus is important in solving many tasks. If the parallel corpus is going to be used by a translator, then a translation dictionary can be created from the parallel corpus. A search tool is normally used by translators to show the queried word and translation examples using a concordance table. On the other hand, models are normally created from the parallel corpus if it is used for natural language processing task such as machine translation, speech translation, and multilingual information retrieval. In all these cases, word/phrase alignment is one of the most important processing that must be performed on the parallel text.

2.1 Parallel corpus acquisition

The evident source of texts is the Web as illustrated by the work of Resnik and Smith [26], who viewed the Web as a parallel corpus, besides the fact that many Web documents are free for download. The Web as a parallel corpus means that one webpage written in a source language has its fully or partially translated version in other language stored in another webpage. There are dedicated tools for harvesting parallel Web documents, such as STRAND Resnik and Smith [27], ILSP-FC [28] and Bitextor [29] that are used for downloading, preprocess and extracting candidate parallel sentences. If the location of the parallel translations is known, then the task is simply to retrieve the documents. The task of locating parallel texts becomes challenging when it has to be done automatically. A search tool needs to locate webpages that might have parallel translations. To overcome the problem, different strategies have been proposed in the literature. In STRAND, the location of webpages is based on the structural relation between a parent webpage and its sibling webpage. The parent webpage contains hypertext links that connect the parent page to its sibling pages, corresponding to the different versions of the parent document in other languages. Thus, the task of a search engine is to locate webpages that share in their anchor texts the name of two languages. Other heuristics can also be explored

to find possible parallel translations. For instance, the publication date of a document, especially news documents, can be used to reduce the scope of the search [30]. Zhu et al. [31] proposed to use the digital fingerprint of a website to find the translated website. The digital fingerprint of a document is a sequence of numbers (in digits) that contain in a document. However, this approach will not work well if there are no sufficient unique numbers in a text. Esplà-Gomis et al. [32] provided a non-exhaustive list of strategies that can be used to locate automatically parallel texts: similarities in the URLs corresponding to webpages, parallelisms in the structure of HTML files, content similarity such as maximum bag-of-words overlapping [32, 33] or shared n-grams [34], file size comparison, language markers in the HTML structure, mutual hyperlinks between webpages [26], and images co-occurrence. In addition, it is possible to extract parallel phrases or sub-sentential text from nearly identical documents that are independently created [30], but this process is complex, requires a very large set of documents for searching, and prone to errors. In some other cases, parallel text may also locate in the same document. For example, in Malaysian journal articles and thesis, it is customary to have the abstract of the article to be written in Malay and English. This observation was explored to extract Malay-English parallel sentences [35]. Once the candidate pairs of texts are identified, the sentences may have to be aligned because many times the text is not translated in the same order or completely. There are algorithms for aligning sentences that apply the sentence length [36], dictionary [37] and BLEU score [38].

When the required data is not available in the Web, researchers need to either locate the data in different supports or construct a corpus from scratch. This is usually the case for collecting speech corpora, low density languages including dialects, and non-digitised documents such as novels [39], technical manuals from translation memories [40], written records on folk culture and folktales [41], and bilingual dictionary [35]. One interesting example is the Basic Travel Expression Corpus (BTEC) [42]. The corpus contains more than 200 thousand common phrases and sentences in Japanese-English extracted from travelling phrase books. The initial project was later extended to cover other language pairs such as Chinese-English, Arabic-English, Italian-English and Indonesian-English. Another Japanese-English bilingual travel corpus is the SLDB (Spoken Language DataBase) corpus. The parallel corpus contains conversation speech between a tourist and a front desk clerk [43]. The speech was transcribed and translated by an interpreter from Japanese to English or English to Japanese. An example in the domain of English-Malay machine translation was the construction of parallel text from examples in the bilingual English-Malay dictionary, where more than a hundred thousand parallel phrases and sentences were collected [35]. In other specialized domain in English-Malay translation, Rahman et al. [44] collected fifteen thousand parallel sentences in the domain of agriculture and health. Government data can be also a source of parallel texts. For example, European Parliament Proceeding generated a very large parallel text in eleven European languages [23]. The corpus consists of 110 language pairs.

There are a few works in the construction of dialect parallel corpus. Almeman, Lee and Almiman [45] reported a parallel Arabic dialects speech corpora. In this work, the speech in Modern Standard Arabic (MSA), Gulf, Egypt and Levantine dialect were recorded. The text for the MSA was first prepared. The text which consists of more than a thousand sentences was then translated to the other 3 dialects. This is followed by recording of the read speech. In total 32 hours of speech was recorded. Another work is the parallel speech corpus for Japanese dialects [46]. 100 balanced sentences were read by 25 dialect speakers from 5 areas: Tokyo, Tohoku, San-yo, Kansai and Kyushu. Since Japanese characters were used for all the dialects are the same, the speech was only transcribed to Japanese pronunciation and phoneme transcription, without requiring any translation. The speech was evaluated on automatic speech recognition task, where the acoustic model was trained from Corpus of Spontaneous Japanese. The results show that the accuracies for dialects that are not from Tokyo are lowered. On the other hand, Dipper and Schultz-Balluff [47] reported a parallel German dialect corpus of ancient text, the Anselm corpus. The corpus contains a collection of 50 medieval text *Interrogatio Sancti Anselmi de Passione Domini*, written in different dialects from Early New High German and Middle Low German from the 14th to 16th century. The texts are in three different versions: verse versions, short prose versions and long prose versions.

The work to collect Malay dialect parallel corpus described in this paper is slightly different, compared to the works in other parallel dialect corpus. The Malay dialect speakers do not use SM words to write just like in the case of Japanese speakers that use Japanese words [46] or Arabic speakers using Arabic from a particular country [45]. The writing used by the Malay dialect speakers even from the same area may differs because they are no standard dialect words. Standardized Malay dialects do exist, but only for formal Malay used in countries such as Malaysia, Indonesia, Brunei, and Singapore. To obtain Malay dialect translation, instead of using the approach in the Arabic dialect speech corpus, where all translations for a dialect is carried out, we used the idea similar to BTEC corpus. In the BTEC corpus collection English is the pivot. Most language pairs are translated from English, because the travelling phrases used in different languages are often similar, and English is one of the most widely spoken languages. Besides, it is also

easier to extend to other language pairs in the future. In our case, the Malay dialects are translated to SM only. The practical reason is because the native speakers and transcribers only know their native dialect and SM. In Malaysia, there are 13 states and if we were to assume each state has one dialect (even though many states have more than one dialect for example Perak [48], and we were to record and translate every dialect to every other dialect, it will be challenging. The approach will also allow us to extend the corpus to other Malay dialects easier in the future. Besides that, only the SM has a standard orthography compared to other Malay dialects that we studied. Many studies already available on SM [16-19]. A dialect being a variant of the formal/official language used shares many of the same characteristics such as grammar and morphology. Therefore, when collecting dialect resource, we attempt to reuse existing resources that are in abundance in the formal language.

2.2 Data alignment

Alignment is an important process in linguistics and natural language processing. The alignment step attempts to identify correspondence between two or more things. The word alignment in machine translation involves identifying corresponding words between two sentences that are translations of each other. On the other hand, in speech processing such as acoustic phonetics analysis and speech synthesis, the interest is in finding the alignment between the segment of speech signal and phone. The alignment can be done manually by human, or automatically using algorithms. Manual alignment by human experts is however expensive and takes a long time. Alignment algorithms can be divided to distance-based, statistical-based, neural networks, and heuristics. The word alignment in machine translation is a challenging problem as words in the source and target sentence may get reordered, drop or inserted.

The distance alignment algorithms are normally used for string matching. The matching of two strings can be viewed as a sequence alignment. For example, in spell checking, a character in the source word may align to the character in a target word if they are of the same character. If all the characters can be aligned in sequence, this means the word match in the vocabulary. On the other hand, the word error rate (WER) used in automatic speech recognition is calculated by finding the minimum number of words in the hypothesis and reference that are unable to align. The most used distance alignment algorithm is the minimum edit distance or Levenshtein distance algorithm. From the perspective of alignment, the algorithm finds the maximum number of sequential alignments that can be formed. Levenshtein distance has been used in many computational linguistics and natural language processing tasks. In the dialect study for instance, Heeringa [5] applied the algorithm to measure pronunciation differences between dialects.

The statistical approach is one of the most used approach in word alignment. There are many variations of the alignment algorithms, notably the IBM alignment model 1 to 4. The IBM models use the expectation maximization (EM) approach to find the alignment and translation probabilities. The intuition of the EM algorithm is that the words that are often observed together are the translation of each other. The EM algorithm consists of iterative steps: expectation (E) step and maximization (M) step. At the initialization step, every word in the target sentence is aligned to all words in the source sentence. All alignments are equally likely. The E step then estimates the probability of the alignments*, $p(a|t, s)$,* where a is the alignment between the target word t and the source word s. Followed by the M step to gather the count, $c(t|s)$. A lexical table is created at the end, which contains the probability of the alignment between words. Besides EM algorithm, Chen et al. [49] proposed to use multiobjective evolution algorithm to find the alignment. The IBM models use word as the elementary unit of translation. The word-based alignment algorithm is good with modelling parallel sentences that are literally translated, but will have problem when modelling the word relation in the parallel sentences if they are translated by phrase, because there will be a lot of mismatched word alignments. Machine translation that based on phrase unit was proposed by Koehn, Och and Marcu [50] to solve this problem. A phrase translation table is created during alignment through three steps: word alignment, extraction of phrase pairs and scoring of phrase pairs. The phrase-based statistical machine translation was further enhanced with factored translation model that enable additional information such as part-of-speech be used during alignment and decoding [51]. In aligning text of closely related languages, the alignment algorithm has to be modified to take advantage of the similar orthography in the words. Nakov and Tiedemann [52] proposed a word alignment algorithm for closely related languages, where Macedonian and Bulgarian were examined. The statistical approach combines the strength of character-level and word-level alignment. The character-level alignment is exploited since the language pair has overlap vocabulary and strong lexical similarity. In the character level alignment, each word in a sentence is split to character bigram sequences, and alignment is carried out using phrase-based GIZA++. A typical word-level alignment was also produced, and the phrase table created (after converted to bigram sequences) was combined with the phrase table from the character-level alignment.

Analysing the source-to-target word translation is one area of interest in translation, and another area is the changes in the order of the words. In the alignment of parallel text, the reordering model is a model that governs the changes of the word order. The reordering model of a statistical machine translation can be analysed to study the grammar of a target language. Other approaches to study the grammar of a language will requires a part-of-speech tagger or language parser. If parallel text for a language pair exists, it is possible to train a part-of-speech tagger for a target language by leveraging existing resource in a resource rich source language [53]. The main idea is to first generate the part-of-speech for the source language text of the parallel text. The part-of-speech tagger for the target language text can then be built by using the source language tags generated. Zeman and Resnik [54] adapted a Swedish language parser from Danish, a closely related language of Swedish. The approach combines reranking and self-training domain adaptation algorithm for the purpose.

Recently, many studies showed that neural networks produce very good results in solving many problems such as image classification, automatic speech recognition, sentiment analysis and others. In machine translation, a type of neural network known as the recurrent neural networks (RNN) are used. Recurrent neural networks are similar to feedforward neural networks, except that the recurrent neuron has an additional connection pointing backward to allow the knowledge in sequential data to be captured. The recurrent neurons arranged in an encoder-decoder architecture [55, 56] with attention mechanism [57] have been shown to be very good in sequence-to-sequence modeling, for example part-of-speech tagging [58], machine translation [23, 34, 35, 40], and speech translation [59]. The word/phrase alignment in encoder-decoder networks can be visualized through the attention matrix [60]. For closely related languages, character-level and subword models were also explored in sequence-to-sequence modeling. The character-level model combines convolutional neural networks and RNN and they show impressive results in text classification [61] and language modelling [62]. Finally, the heuristics alignment algorithms use specific associative measures instead of statistical measures to find the alignment. The heuristics alignment algorithms are often used in bioinformatics domain to find alignments quickly in a large database. Examples of heuristics alignment algorithms are K-vec alignment algorithm and word alignment using dice coefficient.

The distance-based alignment algorithm, particularly Levenshtein distance algorithm is efficient in matching string, and it can be used to match words with similar spelling. Thus, it can align words in dialect parallel text. Nevertheless, the statistical information that tells the co-occurrence of two words is also important. This information can be used together to decide on the word alignment. On the other hand, while neural models may have outperformed statistical models in many machine translation tasks recently, but when the amount of the data is small especially in the dialect parallel text case, the alignment accuracy may not be as good as the other approaches.

3.0 BUILDING MALAY DIALECT PARALLEL TEXT CORPUS

Malay dialect text is very limited. Some dialect texts can be found on Internet blogs and forums. These resources are however difficult to be reused as the text are out of context for us, and there might be ambiguity in understanding the contents. Thus, creating parallel text from these texts is challenging. In this paper, we propose to build a Malay dialect parallel text corpus by transcribing and translating from a dialect dialogue speech corpus. The methodology used here is similar to the collection of Japanese-English bilingual travelling speech corpus (SLDB) [43]. The process goes through three main steps: recording dialect dialogues by native speakers, transcribing the dialogues, and then translating the dialect transcription manually to SM. The recording and transcribing of audio take a lot of resources. Nevertheless, by capturing a dialect in speech form, it would allow us to study and understand the dialect in the original form, and without ambiguity, as Malay dialects (except SM) are without a formal writing system. Additionally, the speech corpus will be interesting to be used to study the acoustic phonetic aspects of the dialect or for speech translation task in the future.

3.1 Recording dialect dialogues

The dialogue recordings were conducted in noise free rooms at Universiti Sains Malaysia (USM), Penang and Universiti Malaysia Sarawak (UNIMAS), Sarawak. Two Malay dialect speakers were asked to discuss different topics of interest to them in separate room through a telephone. The two speakers were separated so that the conversation speech from each person can be clearly recorded without overlapping. A microphone headset was also mounted to each speaker and it was connected to a computer. The conversation speech was captured by the headset and recorded using the CoolEdit software. The speech was recorded at a sampling rates of 16 kHz/16 bits per sample. About five hours of KD conversations were recorded. One male and nine female native Kelantan speakers, between 21 to 24

years old took part. Thirty conversations with different topics were recorded. Two participants carried out a conversation, which took 10 minutes on a specific topic. On the other hand, for SD about one hour and twenty minutes of conversations were recorded. Two native SD speakers, one male and one female, participated in the dialogues in eight different topics. They are both 31 years old. Due to time and other resource constraints, we recorded a smaller amount of conversations and speakers in SD. Refer to Table 2.

Table 2**:** Summary of recorded speech conversation

Criteria	**Recorded Speech Conversation**	
	KD	**SD**
Age	21-24	31
Female	9	1
Male	1	1
Duration (10 minutes per topic)	5 hours	1 hour and 20 minutes
Total topics	30	8
Transcribed topics	12	8
Location	Universiti Sains Malaysia (USM)	Universiti Malaysia Sarawak (UNIMAS)

3.2 Transcribing and translating dialect dialogues

After the recording, the dialogues were transcribed in the target dialect (i.e. KD) by native dialect speakers. Native dialect speakers will listen to the recording and then write them in words in his/her dialect and then translated to SM. One of the samples of parallel sentences in KD and SD is shown below:

Kelantan dialect: *Teh adik tawa hebey keh tok letok gulo.*
Standard Malay: *Teh adik rasa tawar kerana terlupa letak gula.*

Sarawak dialect: *Zul ada sik kitak nangga dalam Astro.*
Standard Malay: *Zul ada tak kamu menonton dalam Astro.*

Each dialogue consists of around two hundred to four hundred sentences. Only twelve of the total thirty dialogues in KD were transcribed and all eight dialogues in SD were transcribed as listed in Table 3. There were two transcribers for each dialect. In total, the manual transcription produces 2755 of KD/SM parallel sentences and 3115 of SD/SM parallel (Table 4).

Table 3: Dialogue topics in KD and SD

No.	Dialogue Topics	
	KD	**SD**
1.	Accident (Hospital)	Malaysian Artist
2.	Sport (Buying Coupon)	Areophane
3.	Play Truant	Student Discipline
4.	Food in Kelantan	“Hari Raya”
5.	Cultural Arts in Kelantan	Birth
6.	Flood	Disease
7.	Social Problem	Research
8.	Marriage issue	TV Series
9.	UKM	
10.	Playground	
11.	“Mat Rempit”	
12.	Ball	

Table 4: Number of sentences and tokens in the parallel texts

	KD / SM	SD / SM
Number of sentences	2755	3115
Number of tokens	17191/ 16638	13436/12983

3.3 Aligning transcribed dialect words and phrases

The alignment of words and phrases is executed after acquiring the parallel sentences. The alignment results are useful to study and analyse the Malay dialect orthography, specifically on the similarity and difference in vocabularies, and grammar. We propose a hybrid distance-statistical-based phrase alignment algorithm that uses Levenshtein distance and statistical approach. The alignment algorithm was improved from Khaw and Tan [63] to include phrase matching. The alignment algorithm consists of four steps as shown in the pseudocode in Fig. 3

```
ALIGN(DIALECT TEXT, STANDARD TEXT)
        Input: Parallel text (Dialect <-> Standard)
        Outputs: (1) Vocabulary; (2) Alignment rules
        Step 1: Align similar words with Levenshtein distance
        Step 2: Align non-similar words using pigeonhole principle
        Step 3: Refine aligned word pairs using maximum likelihood estimation
        Step 4: Align word-to-phrase and phrase-to-word based on conditional probability estimates
```

Fig. 3: Hybrid distance-statistical based dialect phrase alignment algorithm.

3.3.1 Step 1: Aligning similar words with Levenshtein distance

The first step of the alignment algorithm is to align similar words in the parallel sentences. Similar words are words in a target language (e.g. SM), that are perceptually and semantically close to words in a source language (e.g. Malay dialect). Our hypothesis is that source and target word that are similar in spelling are also semantically similar. For example, the word '*masa*' (English: time) and '*tak*' (English: no) in SM are written as '*maso*' and '*tok*' in KD. Parallel sentences are first tokenized into words before the distance between the target language word and every word in the source language sentence is calculated using Levenshtein distance. Levenshtein distance is a measurement of minimum number of edits between two strings. The distance calculated is the score between a source and target word pair. The lower the score, the more similar the word pair. A score of zero means both strings are totally matched.

The step is illustrated with an example in Fig. 4. The parallel sentences used in the example are '*saya bawa nasi.*' and '*kawe bawak nasi.*' (English: I brought rice). Each sentence is tokenized using the space character. Fig. 5 illustrates the distance matrices calculated with the example.

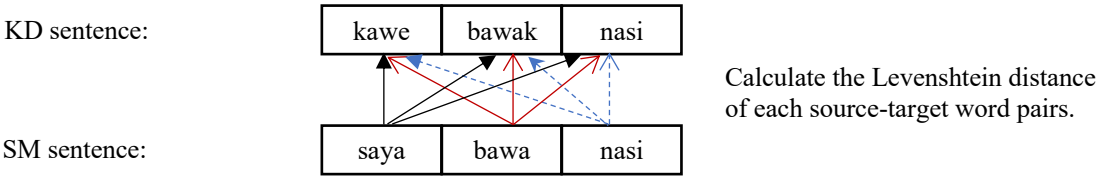

Fig. 4: Levenshtein distance comparison for a word in SM to all KD words

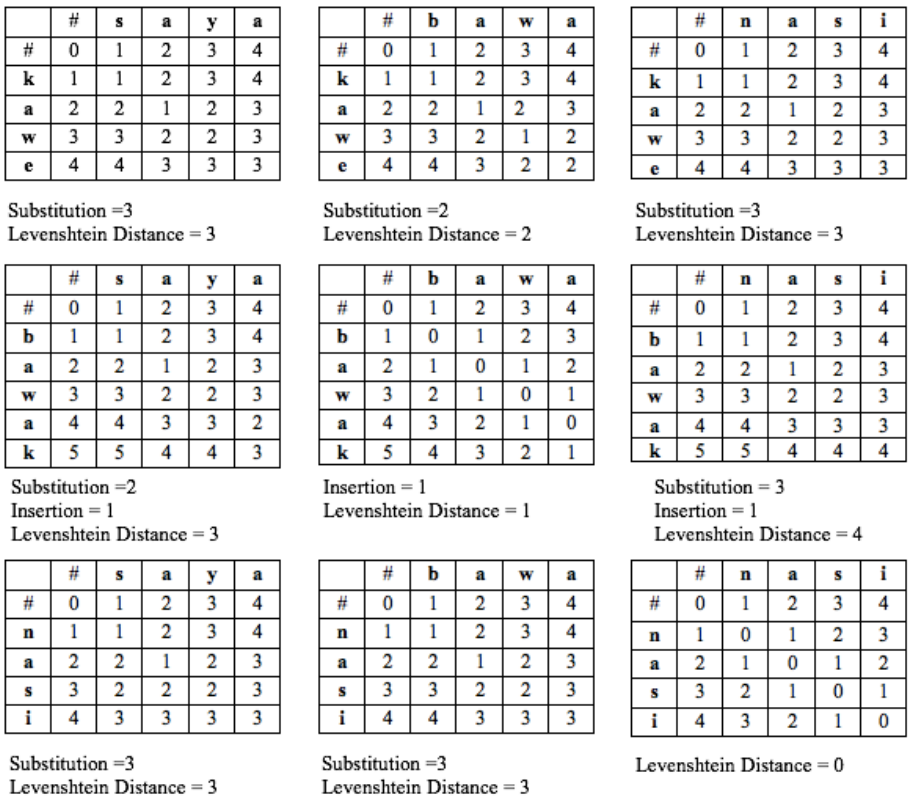

Fig. 5**:** Distance matrix for SM and KD words: Cost of substitution, insertion, and deletion are 1.

The Levenshtein ratio is then calculated for each source and target word pair using equation (1) below. The word pair that has the lowest Levenshtein ratio is aligned together, if the value is less than a predefined threshold. Refer to equation (2), $a(w_s, w_t)$ is the alignment of the similar source language word, w_s and target language word, w_t. In the example given in Fig 5., the SM word '*bawa*' and '*nasi*' will be aligned to the KD word '*bawak*' and '*nasi*' respectively, but the word '*saya*' is not aligned to any dialect words because the Levenshtein ratio of the closest pair is more than the predefined threshold. Alignment threshold is set at 0.4 based on the development data. Some examples of similar words in KD and SD compared to SM are shown in Table 5.

$$ratio_{Lavenshtein} = \frac{distance_{Levenshtein}(w_s, w_t)}{length(w_s)} \quad (1)$$

$$a(w_s, w_t)' = \ argmin\ ratio_{Lavenshtein}(w_s, w_t) \qquad if\ ratio_{Lavenshtein}(w_s, w_t) < threshold \quad (2)$$

Table 5**:** Examples of similar words in KD and SD compared to SM

Similar Words			
KD	**SM**	**SD**	**SM**
mano	*mana*	*pake*	*pakai*
abe	*abang*	*pulo*	*pulau*
naka	*nakal*	*mberi*	*memberi*
pula	*pulau*	*ngisi*	*mengisi*
anok	*anak*	*nyesal*	*menyesal*

3.3.2 Step 2: Aligning non-similar words using pigeonhole principle

At this point, there might be some words in the target language (SM) that are not aligned to any word in the source language (Malay dialect). The source language word that is not aligned to any target language word will be aligned to the remaining target language word without any aligment using pigeonhole principle. In general, the pigeonhole principle states that if there are *n* pigeons and *m* holes, where *n* is more than *m*, then there will be at least one hole that contains more than one pigeon. Therefore, in our earlier example, since the number of source language words and target language words in the parallel sentence are the same, then the word '*saya*' will be aligned to '*kawe*'. Table 6 shows some examples of unique words alignments of KD/SM and SD/SM.

Table 6: Examples of non-similar word alignment of KD/SM and SD/SM

Unique Words			
KD	**SM**	**SD**	**SM**
bokali	*mungkin*	*molah*	*buat*
oyak	*kata*	*madah*	*beritahu*
cakno	*peduli*	*sik*	*belum*
hok	*yang*	*kamek*	*saya*
puwe	*perempuan*	*sidak*	*mereka*
so	*satu*	*ya*	*itu*
dok	*sedang*	*siney*	*mana*
loni	*sekarang*	*sitok*	*sini*
sokmo	*selalu*	*nak*	*yang*
katok	*pukul*	*mun*	*kalau*

3.3.3 Step 3: Refining alignment based on most frequent word pairs

The previous steps may produce erroneous word alignments or a source language word that aligns to many target language words. In this step, the algorithm will update the word alignments using the statistics obtained from the preliminary alignments produced in previous steps. The best alignment for a source language word is the target language word that gives the highest probability. See equation (3).

$$a(w_s, w_t)' = \text{argmax}\, \text{P}(w_t|\, w_s) \tag{3}$$

$$a(w_s, w_t)' = \text{argmax}\, \frac{C(w_s, w_t)}{C(w_s)} \tag{4}$$

In equation (3), w_s is the source word and w_t is the target word. $P(w_t|w_s)$ is the conditional probability distribution of w_t given w_s. $C(w_s, w_t)$ is the count of w_s and w_t, and $C(w_s)$ is the count of w_s. For example, the KD words '*kawe*', '*sera*', and '*sayu*' are aligned to the word '*saya*' in SM (English: I, me) with the total count of 10, 1 and 3 respectively. Thus, the alignment of '*kawe*' and '*saya*' is kept.

3.3.4 Step 4: Aligning word-to-phrase and phrase-to-word based on conditional probability estimation

A word can be translated using more than a word forming a phrase (one-to-many translation), or a phrase can be translated to a single word (many-to-one translation). We assume that an unaligned word, w_i in the source or target language might be a component of a phrase. Thus, the unaligned word w_i can be combined with its neighboring word w_{i-1} or w_{i+1} to form a phrase. In this study, the length of a phrase is limited to a window of two, that is a bigram. A phrase is then identified by finding the most probable word w_{i-1} or word w_{i+1}, which is computed by the formula in equation (5) where W' is the most probable phrase.

$$\text{W'} = \text{argmax}\, (\text{P}(w|w_{-1}), \text{P}(w_{+1}|w)) \tag{5}$$

A phrase formation threshold can be used to determine whether a phrase should be formed. The threshold was set at 2. If the (bigram) probability of a sequence is low, we assume it is not a valid sequence. A development set data can

be used to estimate the threshold. There are 55 of phrases of length two in KD have been identified while 19 of phrases of length two are found in SD.

After identifying the phrases, the alignment of the phrase can be determined. For a phrase with the size of two, the alignment for the phrase will simply be the alignment determined by w_{i-1} or word w_{i+1}. For example, in the KD sentence '*demo lagu mana ni.*', where the word '*lagu*' is unaligned, the count for the sequence '*demo lagu*' is 3, while the count for the sequence '*lagu mana*' is 10. So, the phrase '*lagu mana*' will be created. Table 7 shows the examples of phrase alignment of KD/SM and SD/SM.

Table 7: Examples of phrase alignment of KD/SM and SD/SM

Phrases Alignment			
KD	**SM**	**SD**	**SM**
manih leting	*sangat manis*	*duak igek*	*dua*
tawa hebe	*sangat tawar*	*macam ney*	*bagaimana*
sesok do'oh	*sangat miskin*	*ndak brani*	*takutnya*
air batu	*air sejuk*	*giney giney*	*walau bagaimanapun*
kering tekok	*dahaga*	*jaik hati*	*sedih*
tepung boko	*kuih talam*	*kinek kinek*	*sekarang*
tak dok	*tiada*	*cam tok*	*begini*
keto sewo	*teksi*	*ujong ujong*	*akhirnya*
jokong jokong	*batu bata*	*musin musin*	*memutarkan*
sak ni	*tadi*	*tek dah*	*telahpun*

4.0 EVALUATION AND ANALYSIS OF THE DIALECT ALIGNMENT ALGORITHM

Experiments were performed to evaluate the proposed word alignment algorithm by comparing it to the state-of-the-art GIZA++ word alignment algorithm. The proposed algorithm works faster when the parallel text is small because a smaller parallel text generally has a smaller vocabulary, and therefore, the task of computing Levenshtein distance will be less. The calculation of the Levenshtein distance is time-consuming as it has the time complexity of $O(|VS|*|VT|*m*n)$, where $|VS|$ is the size of the source vocabulary, $|VT|$ is the size of the target vocabulary, m is the average size of the source word and n is the average size of the target word. After computing the Levenshtein distance, many alignments were found, and the following steps will be less computation intensive, whereas GIZA++ does many iterations (average 4-5), in each iteration, it does $O(|VS|*|VT|)$. When the size of training corpus is large, the size of vocabulary found in the corpus will be large. Therefore, the word alignment stage will be the most time-consuming for large training corpus. However, GIZA++ requires a training (parallel) text with sufficient content to produce good alignments of words for bilingual sentences [64]. Alignment was evaluated through precision and recall as follows:

$$Precision = \frac{Number\ of\ Correct\ Alignment}{Total\ Number\ of\ Reference\ Alignment} \quad (6)$$

$$Recall = \frac{Number\ of\ Correct\ Alignment}{Total\ Number\ of\ Proposed\ Alignment} \quad (7)$$

There were 2755 sentences of KD and 3115 sentences of SD from the transcribed dialogue speech corpus. Two thousand sentences from each Malay dialect were selected for training, and 30% of the sentences were randomly chosen from the parallel text in KD and SD for evaluation. The precision and recall for KD and SD are shown in Table 8.

Table 8: Precision and recall of the alignment evaluation

	GIZA++ (baseline)		Proposed approach	
Malay dialect	**Kelantan**	**Sarawak**	**Kelantan**	**Sarawak**
Precision	0.9341	0.9282	0.9542	0.9503
Recall	0.9304	0.9204	0.9502	0.9432

In general, the higher the precision and recall the better the alignment algorithm. The experiment was evaluated on dialect sentences with formal and informal SM. The average precision and recall of the alignment between Malay dialect and formal SM obtained from our proposed approach were 0.9542 and 0.9502 for KD, and 0.9503 and 0.9432 for SD. The overall results show that the proposed algorithm is better than the baseline GIZA++. The higher precision and recall are due to the usage of Levenshtein distance for matching similar words in the parallel sentences. The word similarity matching used allows us to align sequences that do not appear frequently. Besides that, another advantage of the proposed algorithm is that it produces one-to-one, one-to-many, many-to-one or many-to-many alignment, whereas GIZA++ produces one-to-one or one-to-many alignments, but it does not posit many-to-one or many-to-many relationships [65]. Table 9 shows some KD word and phrase alignments obtained from the parallel text.

Table 9: Top three examples of alignment results in KD

Alignment type	**SM**	**KD**
One-to-one	*awak* *balik* *semakin*	*demo* *kelik* *koho*
One-to-many	*dahaga* *tadi* *teksi*	*kering tekok* *sak ni* *keto sewa*
Many-to-one	*di sana* *macam mana* *tidak mahu*	*ssana* *gano* *tokse*
Many-to-many	*sangat tawar* *sangat miskin* *sangat manis*	*tawa hebe* *sesok do'oh* *manih leting*

The alignment algorithm also clusters variants of the same word together. These variants (Table 10) occur due to the different transcriptions provided by the native speakers who transcribed the dialogues.

Table 10: Examples of variants in KD and SD

Clustering of word variants			
SM	**KD**	**SM**	**SD**
rumah	*a. ghumoh*	*memberi*	*a. mberik*
	b. rumoh		*b. memberik*
boleh	*a. boleh*	*mengisi*	*a. ngisik*
	b. buleh		*b. ngisi*
kereta	*a. kheta*	*hujung*	*a. ujung*
	b. kreta		*b. ujong*
	c. kereta		*c. hujong*

Table 11 shows the size of KD vocabulary and SD vocabulary found in the parallel text. The vocabulary is divided to 3 groups based on their similarity to the SM words: similar words, non-similar words and same words. The size of the KD and SD vocabulary are 3237 and 2676 respecitively. The number of non-similar (unique) words in KD and SD are about 12%. This indicates that about 10% of the dialect words can not be found in SM. Interestingly, KD has about 64% of similar words, which mean that the pronunciation of the KD words differs a lot compared to SM. The number of similar words in SD is lower, which is at 43%. On the other hand, SD has more same words compared to KD. This shows that the percentage for a SM word appears in SD and KD stands at 44% and 24% respectively.

Table 11: The size of KD and SD vocabulary

Malay Dialect	Total Vocabulary	# Similar Words		# Same Words		# Non-Similar Words	
		Total	Percentage	Total	Percentage	Total	Percentage
KD	3237	2062	63.70%	792	24.47%	383	11.83%
SD	2676	1162	43.42%	1171	43.76%	343	12.82%

5.0 MALAY DIALECT LEXICAL ANALYSIS

This section examines the lexical similarities and differences between SM and Malay dialect through the analysis of similar words found in word alignment. Many of the findings are supported by the studies in Malay phonology and phonetics indirectly in the literature. Speech and writing are very closely connected. Phoneme is the smallest unit of sound that distinguish a word in a language. Grapheme is the letters that represent a phoneme. The analysis of the writing by the native speakers is important for natural language processing purpose such as dialect identification, sentiment analysis, speech synthesis etc.

5.1 KD lexical analysis

After analysing the spelling of similar words in KD-SM, we found 13 unique group of letters used in KD but not in SM which we hypothesized are KD graphemes, in addition to the 32 graphemes [66] in SM (and minus the two diphthongs). These unique group of letters are ‘pp’, ‘bb’, ‘tt’, ‘dd’, ‘kk’, ‘gg’, ‘ss’, ‘cc’, ‘jj’, ‘ll’, ‘mm’, ‘nn’, and ‘ww’, which were identified manually from the analysis of similar words (e.g. sini in SM vs ssini in KD). In addition, we generalize 16 differences in writing between SM and KD. The first 15 in Table 12 describe the lexical differences, while the other two involves the word order. Table 12 below lists the differences in details and examples.

Table 12: Differences in writing between SM and KD words

No.	Differences	Description	SM	KD	Meaning
1.	**Final ‘s' Substitution**	The letter ‘s’ at the end of the SM base word is substituted by a letter ‘h’ if it precedes with a letter ‘a’.	*pedas* *atas*	*pedah* *atah*	spicy above
2.	**Final ‘l’ and ‘r’ Deletion**	The letter ‘l’ or ‘r’ at the end of a SM base word is deleted if it precedes by an ‘a’.	*mahal* *lapar*	*maha* *lapa*	expensive hungry
3.	**‘a’ followed by ‘ng’, ‘n’ or ‘m’ Substitution**	The letter ‘a’ followed by a letter/group of letter ‘ng’, ‘n’ or ‘m’ in the last syllable of a SM base word is substituted by a letter ‘e’.	*malang* *cawan* macam	*male* *cawe* *mace*	unfortunate cup same as
4.	**‘a’ followed by ‘h’ or ‘k’ Substitution**	The letter ‘a’ followed by a letter ‘h’ or ‘k’ in the last syllable of a SM base word is substituted by an ‘o’ in KD.	*anak* *salah*	*anok* *saloh*	child wrong
5.	**Final ’a’ Substitution**	The letter ‘a’ at the end of a SM word is substituted by an ‘o’.	*masa*	*maso*	time
6.	**‘m’, ‘n’, and ‘ng’ Deletion**	The letter ‘m’, ‘n’ and ‘ng’ in a SM base word that appears at the coda of the syllable is deleted if the syllable is not the last syllable.	*kampung* *pintu* *bungkus*	*kapung* *pitu* *bukuh*	village door package
7.	**Final ‘ai’ and ‘au’ Substitution**	The group of letter ‘ai’ and ‘au’ at the end of a SM base word is substituted by a letter ‘a’.	*pulau* *kedai*	*pula* *keda*	island shop
8.	**‘r’ in Prefix ‘ber’ and ‘ter’ Deletion**	The letter ‘r’ in the prefix ‘ber-’ and ‘ter-’ of a SM word is deleted if the base word starts with a consonant except ‘h’.	*berlatih* *tertelan* *berikat*	*belatih* *tetele* *berikat*	to train swallowed belted

		If base word starts with a 'h', the letter 'h' is dropped.	*berhulur terangkat terhanyut*	*berulo terakat teranyut*	is giving upraised adrift
9.	**'e' of Prefix 'se-' Deletion**	The letter 'e' in the prefix 'se-' of a SM word is deleted if the base word starts with a vowel. If base word starts with a letter 'h', 'h' is dropped.	*sehijau seindah*	*sija sindoh*	as green as beautiful
10.	**Suffix '-kan' Substitution**	A SM word with suffix '-kan' is substituted by a prefix 'pe-' for base word that starts with a consonant except 'h' or the prefix 'per-'. If the base word starts with 'h', the 'h' is dropped.	*tidurkan ingatkan hangatkan*	*petido peringat perangat*	to snooze to remind to heat up
11.	**Suffix '-an' Substitution**	Suffix '-an' in SM is written as '-e' in KD.	lebihan harapan	lebihe harape	surplus hope
12.	**Particle '-lah' Deletion**	Particle '-lah' in SM is written as '-la' in KD.	*sinilah*	*sinila*	over here
13.	**Particle '-kah' Substitution**	Particle '-kah' in SM is written as '-ko' in KD.	yakah	yoko	is it?
14.	**Double Consonants**	a) The preposition is deleted and the first consonant of the next word is duplicated	*ke sini di dalam pada baju*	*ssini ddalam bbaju*	there inside clothes
		b) The first element of the reduplication word is aborted and at the same time the initial consonants in the second element of the first syllable is doubled.	*jalan-jalan*	*jjalan*	stroll
		c) When words made up of three syllables, the first syllable is dropped. The dropped syllable will be replaced by raising the length of the first consonant in the second syllable of the word. The dropped syllable could be a prefix or phonological features of a word that supports such syllable, which does not support any meaning.	*membakar sebenar menjual terkejut*	*bbaka bbena jjual kkejut*	to burn real to sell shocked
15.	**Swapping Perfect Marker Position**	In SM, the perfective marker *sudah* occurs before an intransitive verb. In KD, the same perfective marker written as *doh* occurs after an intransitive verb.	*Dia sudah makan.*	*Dia makan doh.*	He has already eaten.
16.	**Swapping Intensifier Position**	In SM, the intensifiers '*sangat*', '*sungguh*', and '*benar*' occur before an adjective In KD, the same intensifiers occur after the adjective.	*Dia sangat letih.*	*Dia letih sangat.*	He is very tired.

Most of the findings observed in the dialect writing are supported indirectly by the Malay phonological studies, due to the relationship between spelling and pronunciation in a language that can be captured with letter-to-sound rules. There are some new observations not found in the literatures. This show that the dialect parallel text is an equally effective medium if not better in capturing dialect phenomena.

5.1.1 Deletion of consonants in word-final position

Adelaar [67] and Abdul Aziz [68] described the final /s/ substitution by /h/. Their finding is similar to our observation, where the final SM letter 's' is substituted with letter 'h' in KD (refer to Table 12, item 1). This correspondence has also been found in most of the Malay dialects spoken in Thailand [69, 70] as well as in Ulur Muar Malay and Kedah Malay, but not in Johor Malay [71]. The loss of word-final /s/ is reported to exist also in Eastern Romance languages and Taiwanese language [72]. Brown et al. [73] recorded 19 languages that have the correspondence /s/ > /h/. Besides the final /s/ substitution, Abdul Aziz [68] also reported that the consonant deletion in the final phoneme /l/ and /r/. This observation is similar to our finding, where SM words with the final letter 'l' and 'r' are deleted in KD. Refer to Table 12, item 2. In term of pronunciation, the vowel before the deletion will be lengthened, e.g. 'kapal' as [ka:pa:]. This is however not expressed in the writing.

Adelaar [67] and Abdul Aziz [68] also describe the glottalisation of final plosive. The substitution of final stop by a glottal stop is however not reflected in the dialect writing. For example, 'dakap' to /dakaʔ/ and 'ikat' to /ikaʔ/. In this case, the native speakers may prefer maintaining the same spelling because the letter 'k' is associated with the sound /ʔ/ in SM. In Tioman Malay, several grammatical particles end with the glottal stop /ʔ/ [74]. The glottalisation of consonants in word-final position can be found in other Malay dialects: Bangkok Malay [70].

Interestingly, when the letter 'a' is followed by a letter 'm', 'n' or 'ng' in the last syllable of the SM base word, the letter 'a' and the subsequent letter is replaced by the letter 'e' in KD. Refer to Table 12, item 3. This observation is similar to the finding by Abdul Aziz [68], where he pointed that /a/ and subsequent nasal consonant is replaced by the nasalized vowel /ɛ̃/. Refer to Table 12. The speakers in this case use the letter 'e' to represent the sound /ɛ̃/. On the other hand, according to [71, 75], the sound change depends on the vowel preceding the nasal. If the vowel is not /a/, then the nasal is changed into /n/. However, if the vowel is /a/, the nasal becomes /ɛ/. In our analysis, we do not find any changes of "nasal letters" to 'n'. Ajid found that some Kelantan dialect speakers do not nasalise the final vowel [14]. The final nasal deletion and the regressive rule of nasalisation are "the two most salient phonological features" that differentiate Kelantan dialect from the other Malay dialects [14]. The regressive rule of nasalisation exists only in Kelantan dialect and not in other Malay dialects [14]. Refer Table 13.

Table 13: Regressive rule of nasalisation in KD [14]

Standard Malay word	***jalan***
Underlying form	/jalan/
Vowel raising	/jalɛn/
Regressive nasalisation	/jalɛn/
Word final nasal deletion	/jalɛ/
Final form	[jalɛ]

5.1.2 Simplification of nasal + voiceless stop clusters

A nasal consonant is deleted when it is followed by a voiceless stop consonant like /t/ [75]. The description is similar to the finding we get in item 6 of Table 12. The SM words *kampung*, *pintu* and *bungkus* are prounounced as [kampoŋ], [pintu] and [buŋkus] respectively. These words appeared as *kapung*, *pitu* and *bukuh* respectively in the KD parallel text.

5.1.3 R-drop of a prefix

As noticed by Yunus [76], in SM, the final 'r' of the prefixes *ber-, per-*, and *ter-* may either be pronounced or not at all when the base begins with a consonant. However, when the base starts with a vowel, the 'r' is pronounced in a normal way. In KD, the 'r' in the prefixes *ber-* and *ter-* is systematically dropped when the base starts with a consonant except 'h'. Refer to Table 12, item 8. The r-drop process for the prefix *ter-* can be found also in Johor Malay dialect [71].

5.1.4 Gemination of the initial consonant

The geminate consonants are identical consonants that may be realised in pronunciation as a single long consonant [77]. Our result shows that the germination in KD writing is realised by doubling the consonant. This increases the number of hypothesized graphemes available in KD. From the native KD writing, we found 13 additional unique group of letters ('pp', 'bb', 'tt', 'dd', 'kk', 'gg', 'ss', 'cc', 'jj', 'll', 'mm', 'nn', and 'ww') not found in SM [66]. In KD, like in Pattani Malay, the geminate consonants occur only in word initial position. In the examples given in Table 10, the gemination results from a morphological process in a prepositional phrase: the preposition is deleted and the initial consonant of the noun is geminated. Hamzah et al. [78] give examples of gemination in phoneme transcription and meaning, based on this information we infer the word is a reduplication word (pagi-pagi > *ppagi* 'early morning') and an affixed word (*tertidur* > *ttido* 'sleep by change'). From our finding, we observed double consonants are used to substitute reduplication words and words made up of three syllables. Refer to item 15 in Table 10. However, this substitution does not happen all the time, and it is still acceptable of using reduplication words and words made up of three syllables in KD. "Word-initial geminates are typologically rarer than word-medial geminates." [79]. Like in KD, the Maltese lexical geminates are conditioned morphologically to derive passive and reflexive forms [79].

5.1.5 Vowel lowering in word-final position: **/u/ > [o]**

Based on the explanation of Adelaar [67] that "the high vowel /u/ spelt as 'u' is lowered to the vowel [o] spelt as 'o' in word-final position", the /u/ at the final-syllable position before a final stop will be realised either as [u] or [o], for example /masuk/ as [masoʔ], but if /u/ is before a final-syllable position before /n/ and /ŋ/, it will be realised as [õ]. In our text analysis, we do not encounter any example that show the changes in the grapheme 'u' in the context mentioned. One likely reason is because the grapheme 'u' in these contexts in SM are also pronounced as [o], for example the phonetic transcription for the SM word *masuk* is [masoʔ]. Thus, the native speakers may not attempt to differentiate the words in the writing.

5.1.6 Vowel raising in word-final position: **/a/ > [ɔ]**

Our study shows that when the letter 'a' is followed by the letter 'k' or 'h' in the last syllable of the SM base word, the vowel will be written as 'o' in KD. Refer to item 4 in Table 12 This observation is similar to the analysis reported by [71, 67, 19] that describe the vowel /a/ is pronounced [ɔ] when it is followed by /ʔ/, /h/, or in word-final position. Some dialects located in the sea-side of Sarawak also show similar sound change [80].

5.1.7 Monophthongisation in word-final position

Monophthongisation occurs when a diphthong becomes a monophthong. In KD, the vowel sequence /ai/ and /au/ are reduced to /a/ [71]. In writing, we found that the native speakers also replace the letters 'ai' and 'au' with the letter 'a'. Refer to item 7 in Table 12 Other Malay dialects that change /ai/ to /a/ are Pattani Malay and Terengganu Malay [16]. In Pahang dialect, the diphthong /au/ becomes /a/ (e.g. *kala* for *kalau* 'if' in SM), but the diphthong /ai/ becomes a nasalised /e/ (e.g. *sunge* for *sungai* 'river' in SM) [19] like in SD. There is another diphthong in Malay, which is /oi/. This is a very rare sound in Malay and only appears in limited number of words. Unfortunately, we do not manage to capture words with the grapheme 'oi' in the parallel text.

5.1.8 KD Word Order

Besides the lexical differences, we also analyse the word order in KD. Most of the words and phrases have one to one mapping in sequence between SM and KD. Some exceptional cases exist such as the position of the perfect marker *sudah* and the adverb *sangat*. As reported by [81], the sentence structure of Kelantan dialect shows three salient features: the construction of the passive, the position of the perfect marker /doh/, and the position of the adverb *sangat*. In our analysis of the KD word order, we found only the two lasts features.

The prescriptive grammar of SM state that *sudah*, or its contracted form *dah*, is placed before a verb (e.g. *dia sudah makan* 'he has already eaten'). However, in KD, the perfect marker is represented by its contracted form with the vowel change, and it is placed after the verb (e.g. *dia makan doh* [81]). The prescriptive grammar of SM state that *sangat* can be placed before or after an adjective (e.g. *dia sangat letih* or *dia letih sangat* 'he is very tired'). However, in KD, the adverb is always placed after the adjective [81] (e.g. *dia letih sangat*). This construction is attested in all the examples given in the Wikipedia webpage of KD (*Bahasa Melayu Kelantan*). Nonetheless, none of the examples

use the word *sangat*. For instance, the sequences in SM, *sangat manis* 'very sweet' and *sangat masin* 'very salty, have as equivalents in KD *manis lleting* and *masing ppeghak*. The first expression occurs in our corpus with different spelling (*manih letting*), which illustrates how the spelling of the dialect is not standardised yet. In informal SM, it is not common for native speakers to say *dia makan sudah* or *dia letih sangat* although the meaning can be understood.

5.2 SD spelling analysis

In our analysis of SD, we do not find any new hypothesized graphemes besides the graphemes in SM. We generalize 10 differences between SM and SD in Table 14 below. From the 10 differences, there are 8 substitutions, 1 insertion and 1 deletion of graphemes in Standard Malay words. From the 8 substitutions, 3 are performed on the final letters of a word, 5 are performed on the prefix of a word. It does not show any changes in word order.

Table 14: Differences in orthography between Standard Malay (SM) and Sarawak Malay words

No.	Differences	Description	Standard Malay	Sarawak Malay	Meaning
1.	**Final 'ai' Substitution**	The letters 'ai' at the end of the base of a SM word is substituted by an 'e' in Sarawak dialect.	*pakai*	*pake*	to wear
2.	**Final 'au' Substitution**	The letters 'au' at the end of the base of a SM word is substituted by an 'o' in Sarawak dialect.	*pulau*	*pulo*	island
3.	**Deletion of Initial 'h'**	The initial letter 'h' in the base of a SM word is deleted in Sarawak dialect.	*hias*	*ias*	to decorate
4.	**Appending of 'k'**	The letter 'k' is appended to the final vowel of the base of a SM word in Sarawak dialect.	*lupa* *lagi*	*lupak* *lagik*	forget more
5.	**Final 'ng' and 'm' Substitution**	The letters 'ng' and 'm' at the end the base of a SM word is substituted by a letter 'n' if it precedes the letter 'i' in Sarawak Malay.	*kering* *musim*	*kerin* *musin*	dry season
6.	**Prefix 'men-' Substitution**	The prefix 'men-' in SM word is written as 'en-' in Sarawak Malay.	*menjama*	*enjamah*	to taste
7.	**Prefix 'mem-' Substitution**	The prefix 'mem-' in SM word is written as 'm-' in Sarawak Malay.	*memberi*	*mberi*	to give
8.	**Prefix 'meng-' Substitution**	Prefix 'meng-' in SM word is written as 'ng-' in Sarawak dialect.	*mengisi*	*ngisik*	to fill
9.	**Prefix 'men(s)-' Substitution**	The prefix 'men-' in SM is deleted if the prefix is followed by a base word that starts with a 's', the letter 's' is substituted by the letters 'ny'.	*menyesal (*base: *sesal)*	*nyesa*	to regret
10.	**Prefix 'men(t)-' Substitution**	The prefix 'men-' in SM is deleted if the prefix is followed by a base word that starts with a 't', the letter 't' is substituted by the letter 'n'.	*menawar (*base: *tawar)*	*nawar*	to offer

The following subsections discussed previous works particularly in SD phonology that align with our findings from the dialect parallel text.

5.2.1 Glottalisation in word-final position

Some SD words tend to add the glottal stop [ʔ] in word-final position [82, 80], and the sound is rendered by the letter 'k' by the native speakers based on our finding. This phenomenon can be found in other Malay dialects like Tioman Malay dialect [82], and Sabah Malay dialect [83]. For Sabah Malay dialect, Wong [83] proposed the following rule, /∅/ → [ʔ] / V_ #, indicating that the insertion of the glottal occurs if the word ends with a vowel. Our SD examples (in Table 14, item 4) seem following that rule. Examples where the letter 'k' represent the sound /ʔ/ are also seen in other Malay dialects, for example in KD, which is due to the influence of SM, where the letter 'k' at the coda of a syllable is read as /ʔ/.

5.2.2 Deletion of initial /h/ in word-initial position

The phoneme /h/ in SM is realised as voiceless glottal fricative [h] in all environments [83]. However, that phoneme cannot appear in word-initial position in SD except in loan words [82, 19]. Sabah Malay dialect [83] and Ulu Kapuas Malay dialect [84] show similar behaviour. Like Sarawak, Sabah and Ulu Kapuas are also situated on the island of Borneo. The deletion of the phoneme /h/ is also reflected in the spelling as shown in the example in Table 14, item 3.

5.2.3 Nasal substitution in word-final position

In SD, all nasals preceded by the vowel /i/ in word-final position are assimilated to the dental nasal [n] [82]. The SD writing in our dialect parallel text also show similar observation. Refer to Table 14, item 5. As explained earlier, KD constraints the sound changed on the vowel /a/. Madzhi [80] studied extensively the phonology of SD as spoken in Kuching does not mention at all that sound change except through three examples found in his book: /bisin/ for *bising* for 'noisy', /jəɤin/ for *jering* 'soaking', /kəɤin/ for *kering* 'dry' [80]. The Ulu Kapuas Malay dialect seems to have similar behaviour: [ucin] for *kucing* 'cat', [palin] for *paling* 'most' [84].

5.2.4 Monophthongisation in word-final position

Like in KD, the SM diphthongs go through a sound change in SD. The two vowel sequences /ai/ and /au/ become monophthongs, /e/ and /ɔ/ respectively [82, 80, 19]. Our finding in item 1 of Table 14, shares a similar observation. In other Malay dialect for instance Terengganu dialect, the diphthong /au/ becomes a nasalised /o/ (e.g. *piso* for *pisau* 'knife' in SM), but the diphthong /ai/ becomes /a/ (e.g. *bala* for *balai* 'hall' in SM) [19] like in KD.

5.2.5 Nasal alternation of prefix N-

If in SM, the prefix is meN-, then in Sarawak dialect it consists of a single archiphoneme N-. The term "single archiphoneme" is borrowed from Goddard [85], while he was reporting the nasal alternation of the single archiphoneme N- in Javanese. The morphological process of adding the prefix meN- or N- to a base is the same in SM and SD, and it depends on the initial sound of the base. From our finding in SD writing, the rules are as follows: (a) if the base is 'b', then the N- is realised as 'm'; (b) if the base starts with 't', then N- appears as 'n'; (c) if the base starts with 'j', then N- is spelt as 'en-'; (d) if the base starts with 's', N- appears as /ɲ/, which is spelt as 'ny'; (e) if the base starts with a vowel, N- appears as /ŋ/, which is written with the grapheme 'ng'. Refer to Table 14, items 6 to 10. Similar rules can be found with the Javanese N- and the Tagalog maN- [85].

6.0 CONCLUSIONS AND FUTURE WORK

In this paper, we describe our work in collecting a parallel text corpus of SM and Malay dialects. A dialogue speech corpus in Malay dialects was first recorded, and it was then transcribed and translated to SM. We propose a phrase-based alignment algorithm that uses Levenshtein distance and statistical technique for aligning words in dialects. The results show that the alignment algorithm works better than the statistical phrase-based alignment, GIZA++. The alignment algorithm in this study serves two purposes, clustering variants of a word, and analyzing similar words in dialects. From our analysis, we found that most of the Malay dialect words are similar in writing to the SM words, with around 10% of unique words found. There are systematical lexical differences in Malay dialect and SM. Most of the differences happens in the end of a word. Even though it is possible for native dialect speakers to use SM words to represent Malay dialect, they do not do that. The usage of similar but different words in the writing show that native dialect speakers' intension to use a different writing scheme than SM, probably to indicate a different social group they attached to. In term of grammars, Malay dialects show a similar syntactic structure compared to SM, except in a few cases in KD. The parallel dialect text is a very good record that describe the lexical similarities and differences between SM and Malay dialects.

For future works, the writing of other interesting Malay dialects such as Terengganu Malay, Perak Malay, and Kedah Malay can be acquired to give a more comprehensive analysis. The speech corpus acquired can also be used for

acoustic phonetic study or speech processing research such as automatic speech recognition and speech synthesis. The dialogue speech corpus and parallel text corpus will be released at Github (https://github.com/).

REFERENCES

[1] I. A. Bolshakov and A. Gelbukh, Computational Linguistics: Models, resources, applications, Mexico: Fondo De Cultura Economica, 2004.

[2] J. Nerbonne and W. Kretzschmar, "Introducing computational techniques in dialectometry," *Computers and the Humanities,* vol. 37, no. 3, pp. 245-255, 2003.

[3] G. Yule, The study of language, Cambridge University Press, 2010.

[4] V. Fromkin, R. Rodman and N. Hyams, An introduction to language, Wadsworth: Cengage Learning, 2014.

[5] W. Heeringa, Measuring dialect pronunciation differences using levenshtein distance, Netherlands: Ph.D. thesis, Rijksuniversiteit Groningen., 2004.

[6] M. Wieling, F. Tomaschek, F. Arnold, D. Tiede, M. Bröker, Thiele, S. N. Wood and R. H. Baayen, "Investigating dialectal differences using articulography," *Journal of Phonetics,* vol. 59, pp. 122-143, 2016.

[7] M. Wieling, S. Montemagni, J. Nerbonne and R. H. Baayen, "Lexical differences between Tuscan dialects and standard Italian: Accounting for geographic and sociodemographic variation using generalized additive mixed modeling," *Language,* vol. 90, no. 3, pp. 669-692, 2014.

[8] J. Grieve, Regional variation in written American English, Aston University: Cambridge University Press, 2016.

[9] B. Szmrecsanyi, "Grammatical Variation in British English Dialects: A Study in Corpus-based Dialectometry," in *Cambridge University Press*, Cambridge, 2013.

[10] W. Heeringa and J. Nerbonne, "Dialect areas and dialect continua," *Language Variation and Change,* vol. 13, no. 03, pp. 375-400, 2001.

[11] H. O. Asmah, Aspek bahasa dan kajiannya: kumpulan siri ceramah peristilahan, Kuala Lumpur: Dewan Bahasa dan Pustaka, 1991.

[12] J. T. Colins, "Malay Dialect Research in Malaysia: the Issue of Perspective, Bijdragen tot de Taal-, Land- en Volkenkunde," pp. 235-264, 1989.

[13] F. M. Onn, Aspects of Malay phonology and morphology: a generative approach, Bangi, Selangor: Universiti Kebangsaan Malaysia, 1980.

[14] C. K. Ajid, "Word final nasal in Malay dialects," in *Proceedings of the Seventh International Conference on Austronesian Linguistics*, Leiden, 1994.

[15] M. Johari, Fonologi Dialek Melayu Kuching (Sarawak), Kuala Lumpur: Dewan Bahasa Dan Pustaka, 1988.

[16] A. J. Moain, “Bahasa Melayu dan dialek daerahnya,” *Sari,* vol. 11, pp. 63-86, 1993.

[17] H. O. Asmah, The Phonological Diversity of the Malay Dialects, Kuala Lumpur: Bahagian Pembinaan dan Pengembangan Bahasa, Dewan Bahasa dan Pustaka, 1977.

[18] B. S. Teoh, The sound system of Malay revisited, Kuala Lumpur: Dewan Bahasa dan Pustaka, 1994.

[19] H. O. Asmah, Susur Galur Bahasa Melayu, Kuala Lumpur: Dewan Bahasa dan Pustaka, 2008.

[20] H. M. Abdul, Sintaksis Dialek Kelantan, Kuala Lumpur: Dewan Bahasa dan Pustaka, 2006.

[21] M. Z. Abd Rozan and Y. Mikami, “Orthographic reforms of Standard Malay online: Towards better pronunciation and construction of a cross-language environment,” *Journal of Universal Language,* vol. 8, pp. 129-159, 2007.

[22] B. Ranaivo-Malançon, “Computational analysis of affixed words in Malay language,” in *International Symposium on Malay/Indonesian Linguistics, in ISMIL*, Penang, 2004.

[23] P. Koehn, “Europarl: A Parallel Corpus for Statistical Machine Translation,” in *in: MT Summit*, 2005.

[24] K. Parton, K. R. McKeown, J. Allan and E. Henestroza, “Simultaneous multilingual search for translingual information retrieval,” in *In Proceedings of the 17th ACM conference on information and knowledge management*, California, United States, 2008.

[25] D. Stefanescu and R. Ion, “Parallel-Wiki: A Collection of Parallel Sentences Extracted from Wikipedia,” *In Proceedings of the 14th Internation- al Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2013), Samos, Greece,* 2013.

[26] P. Resnik and N. A. Smith, “The web as a parallel corpus,” *Computational Linguistics,* vol. 29, no. 3, pp. 349-380, 2003.

[27] P. Resnik and N. A. Smith, “The web as a parallel corpus,” *Computational Linguistics,* vol. 29, no. 3, pp. 349-380, 2003.

[28] V. Papavassiliou, P. Prokopidis and G. Thurmair, “A modular open-source focused crawler for mining monolingual and bilingual corpora from the web,” in *Proceedings of the 6th Workshop on Building and Using Comparable Corpora*, Sofia, Bulgaria, 2013.

[29] M. Esplà-Gomis, “Bitextor, a free/open-source software to harvest translation memories from multilingual websites,” in *Proceedings of MT Summit XII*, Ottawa, Canada, 2009.

[30] D. Munteanu and D. Marcu, “Extracting Parallel Sub-sentential Fragments from Non-parallel Corpora,” in *ACL 2006*, Sydney, 2006.

[31] Z. Zhu, L. Miao, C. Lei and S. Zheng, “Automatic Construction of Chinese-Mongolian Parallel Corpora from the Web Based on the New Heuristic Information,” in *International Conference on Asian Language Processing*, Penang, 2011.

[32] M. Esplà-Gomis, F. Klubička, N. Ljubešić, S. Ortiz-Rojas, V. Papavassiliou and P. Prokopidis, “Comparing two acquisition systems for automatically building an English Croatian parallel corpus from multilingual websites,” in *Proceedings of the 9th international conference on language resources and evaluation*,

Reykjavik, Iceland, 2014.

[33] E. Morin, A. Hazem, E. Loginova-Clouet and F. Boudin, "LINA: Identifying comparable documents from Wikipedia," in *Proceedings of the 8th Workshop on Building and Using Comparable Corpora*, Beijing, China, 2015.

[34] J. Uszkoreit, J. M. Ponte, A. C. Popat and M. Dubiner, "Large Scale Parallel Document Mining for Machine Translation," in *Coling 2010*, Beijing, 2010.

[35] Y.-L. Yeong, T.-P. Tan, C. K. Lim, K. H. Gan and M. Siti Khaotijah, "Malay-English Parallel Text Acquisition and Comparison of Translation Modeling in Statistical Machine Translation and Neural Machine Translation with Low Resources," in *In Proceedings of ONA*, Phnom Penh, 2017.

[36] W. A. Gale and W. C. Kenneth, "A program for aligning sentences in bilingual corpora," *Comp. Ling.,* vol. 19, no. 1, p. 75–102, 1993.

[37] X. Ma, "Champollion: a robust parallel text sentence aligner," in *LREC 2006*, Genova, Italy, 2006.

[38] R. Sennrich and M. Volk, "Iterative, MT-based sentence alignment of parallel texts," in *In 18th Nordic Conference of Computational Linguistics*, Riga, 2011.

[39] L. Huang, "Building a Chinese-English parallel corpus of modern and montemporary Chinese novels," in *Style in translation: A corpus-based perspective new frontiers in translation studies*, Springer, 2015, pp. 31-41.

[40] A. Menezes and S. D. Richardson, "A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora," in *Recent advances in example-based machine translation*, vol. 21, M. Carl and A. Way, Eds., The Netherlands, Springer Netherlands, 2003, pp. 421-442.

[41] V. Giouli, N. Glaros, K. Simov and P. Osenova, "A web-enabled and speech-enhanced parallel corpus of Greek-Bulgarian cultural texts," in *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage Social Sciences, Humanities, and Education (LaTeCH – SHELT&R 2009)*, Athens, Greece, 2009.

[42] T. S. E. S. F. Y. H. &. Y. S. Takezawa, "Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world," in *In Proceedings of the International Conference on Language Resources and Evaluation*, 2002.

[43] T. K. G. M. M. &. S. E. Takezawa, "Multilingual spoken language corpus development for communication research," *Computational Linguistics and Chinese Language Processing Journal,* vol. 12, no. 3, pp. 303-324, 2007.

[44] S. A. Rahman, N. Ahmad, H. A. Hashim and A. W. Dahalan, "Real time on-line English-Malay machine translation (MT) system," in *Proceedings of the 3rd real-time technology and application symposium*, 2006.

[45] K. Almeman, M. Lee and A. A. Almiman, "Multi dialect Arabic speech parallel corpora," in *Proceedings of the 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*, Sharjah, United Arab Emirates, 2013.

[46] K. Yoshino, N. Hirayama, S. Mori, K. Itoyama and H. G. Okuno, "Parallel speech corpora of Japanese dialects," in *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC*

2016), Portorož, Slovenia, 2016.

[47] S. &. S.-B. S. Dipper, "The anselm corpus: Methods and perspectives of a parallel aligned corpus," in *In Proceedings of the workshop on computational historical linguistics*, Oslo, 2013.

[48] A. Zaharani, The Phonology & Morphology of the Perak Dialect, Kuala Lumpur: Dewan Bahasa dan Pustaka, 1991.

[49] Y. Chen, X. Shi, C. Zhou and Q. Hong, "A word alignment model based on multiobjective evolutionary algorithms," in *Computer and Mathematics with Applications*, 2009.

[50] P. Koehn, F. J. Och and D. Marcu, "Statistical Phrase-based Translation," in *In Proceedings of the Human Language Technology Conference*, Edmonton, 2003.

[51] P. Koehn and H. Hoang, "Factored translation models," in *In Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2007)*, Prague, 2007.

[52] P. Nakov and J. Tiedemann, "Combining word-level and character-level models for machine translation between closely-related languages," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, 2012.

[53] D. Das and S. Petrov, "Unsupervised part-of-speech tagging with bilingual graph-based projections," in *In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, 2011.

[54] D. Zeman and P. Resnik, "Cross-language parser adaptation between related languages," in *In Proceedings of IJCNLP-08 Workshop on NLP for Less Privileged Languages*, Hyberabad, India, 2008.

[55] K. Cho, M. B. van, C. Gulcehre, F. Bougares, H. Schwenk and Q. V. Le, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *in Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, 2014.

[56] I. Sutskever, O. Vinyals and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems*, 2014.

[57] D. Bahdanau, K. Cho and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *In Proceedings of ACL-IJCNLP*, 2014.

[58] T.-P. Tan, B. Ranaivo-Malançon, L. Besacier, Y.-L. Yeong, K. H. Gan and E. K. Tang, "Evaluating Lstm Networks, Hmm and Wfst in Malay part-of-speech tagging," *Journal of Telecommunication, Electronic and Computer Engineering,* vol. 9, no. 2-9, pp. 79-83, 2017.

[59] A. Bérard, O. Pietquin, L. Besacier and C. Servan, "Listen and translate: A proof of concept for end-to-end speech-to-text translation," in *In Proceedings of Conf. on Neural Information Processing Systems*, Barcelona, 2016.

[60] D. Britz, "Attention and memory in deep learning and nlp," 3 January 2016. [Online]. Available: http://www.wildml.com/2016/01/attention-and-memory-in-deep-learning-and-nlp. [Accessed 8 April 2017].

[61] X. Zhang, J. Zhao and Y. LeCun, "Character-level convolutional networks for text classification," in *In*

Advances in neural information processing systems, 2015.

[62] Y. Kim, Y. Jernite, D. Sontag and A. M. Rush, "Character-aware neural language models," in *Conference in Advancement of Artificial Intelligence*, Texas, 2015.

[63] J. Y. M. Khaw and T. P. Tan, "Hybrid Approach for Aligning Parallel Sentences for Languages without a Written Form using Standard Malay and Malay Dialects," in *IALP*, Sarawak, 2014.

[64] L. Tian, F. Wong and S. Chao, "Word Alignment Using GIZA++ on Windows," *Proceedings of Thirteenth MT Summit, Xiamen, China,* 2011.

[65] S. Grimes, K. Peterson and X. Li, "Automatic Word Alignment Tools to Scale Production of Manually Aligned Parallel Text," *In Proceedings of the Eighth International Conference on Language Resources and Evaluation. Istanbul, Turkey,* 2012.

[66] T. P. Tan and B. Ranaivo-Malancon, "Malay Grapheme to Phoneme Tool for Automatic Speech Recognition," in *Third International Workshop on Malay and Indonesian Language Engineering*, Singapore, 2009.

[67] A. Adelaar, "Structural diversity in the Malayic subgroup," in *The Austronesian languages of Asia and Madagascar*, A. Adelaar and N. Himmelmann, Eds., Routledge, 2005, pp. 202-226.

[68] A. Y. Abdul Aziz, "Analisis koda berdasarkan kekangan dalam dialek Kelantan," *GEMA Online™ Journal of Language Studies,* vol. 12, no. 4, pp. 1127-1144, 2012.

[69] A. Thavisak, "Malay dialects in Thailand," in *Proceedings of The International Symposium on Language and Linguistics*, Bangkok, Thailand, 1988.

[70] U. Umar, "Language and writing system of Bangkok Melayu," in *In Proceedings of the International Conference on Minority Languages and Writing Systems*, Beijing, China, 2007.

[71] M. O. Farid, "Aspects of Malay phonology and morphology," University Microfilms International, 1980.

[72] C. H. Wu, "Patterns of sound correspondence between Taiwanese and Germanic/Latin/Greek/Romance lexicons, Part I," *Sino-Platonic Papers,* no. 262, August 2016.

[73] C. H. Brown, E. W. Holman and S. Wichmann, "Sound correspondences in the world's languages: Onlines upplementary materials," *Language,* vol. 89, no. 1, pp. s1-s76, 2013.

[74] J. T. Collins, "The phonology of Tioman Malay and the reconstruction of Proto-Malay," *Dewan Bahasa,* vol. 29, no. 5, pp. 369-383, 1985.

[75] B. S. Teoh and C. K. Yeoh, "Fonem vokal nasal dalam dialek Kelantan," in *Bunga rampai fonologi Bahasa Melayu*, F. M. Onn, Ed., Petaling Jaya, Fajar Bakti Sd. Bhd., 1988, pp. 91-98.

[76] M. Yunus, The Malay Sound System, Kuala Lumpur: Fajar Bakti Sdn. Bhd, 1980.

[77] N. Yupho, "Consonant clusters and stress rules in Pattani Malay," *Mon-Khmer Studies,* vol. 15, pp. 125-137, 1989.

[78] M. H. Hamzah, J. Fletcher and J. Hajek, "Word-initial voiceless stop geminates in Kelantan Malay: Acoustic evidence from amplitude/F0 ratios," in *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS)*, Glasgow, UK, 2015.

[79] L. Galea, A. Hermes, A. Gatt and M. Grice, "Cues to gemination in word-initial position in Maltese," in *In Proceedings of the 18th International Congress of Phonetics Sciences (ICPhS)*, Glasgow, UK, 2015.

[80] J. Madzhi, "Fonologi Dialek Melayu Kuching (Sarawak)," in *Dewan Bahasa Dan Pustaka*, Kuala Lumpur, 1998.

[81] H. Aimi Synaza, "Politeness strategies used by speakers of two Malay dialects," in *University of Malaya*, Kuala Lumpur, 2012.

[82] J. T. Collins, Dialek Melayu Sarawak, Kuala Lumpur: Dewan Bahasa dan Pustaka, 1987.

[83] J. K. L. Wong, The Sabah Malay dialect: Phonological structures and social functions, Kota Kinabalu, Malaysia: Centre for the Promotion of Knowledge and Language Learning, Universiti Sabah Malaysia, 2000.

[84] Yusriadi, Dialek Melayu Ulu Kapuas Kalimantan Barat, Kuala Lumpur: Dewan Bahasa dan Pustaka, 2007.

[85] C. Goddard, The languages of east and southeast Asia, USA: Oxford University Press, 2005.

[86] H. M. Caseli, A. M. Silva and G. V. Nunes, "Evaluation of Methods for Sentence and Lexical Alignment of Brazilian Portuguese and English Parallel Text," in *In Proceedings of the SBIA 2004 (LNAI)*, Berlin, Heidelberg, 1994.

[87] W. O'grady and J. Archibald, Contemporary linguistic analysis: An introduction, Canada: Pearson Canada, 2015.

[88] M. P. Perea, "Dialectometry: A new treatment of dialectal morphological data," 2007.

[89] T. Kendall and V. Fridland, "Mapping the perception of linguistic form: Dialectometry with perceptual data," 2016.

[90] T. Leinonen, "Factor analysis of vowel pronunciation in Swedish dialects," *International Journal of Humanities and Arts Computing,* vol. 2, no. 1-2, pp. 189-204, 2008.

[91] K. Tyshchenko, "Metatheory of Linguistics," 2000. [Online]. Available: http://linguist.univ.kiev.ua/museum/book/index.html. [Accessed 8 April 2017].

[92] F. Petroni and M. Serva, "Measures of lexical distance between languages," *Physica A: Statistical Mechanics and Its Applications,* vol. 389, no. 11, pp. 2280-2283, 2010.

[93] S. F. Chen, "Aligning Sentences in Bilingual Corpora using Lexical Information," in *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, 1993.

[94] L. T. Lim and E. K. Tang, "Building an Ontology-based Multilingual Lexicon for Word Sense Disambiguation in Machine Translation," in *Proceedings of the PAPILLON-2004 Workshop on Multilingual Lexical Databases*, Grenoble, 2004.

[95] C. G. Clopper and D. B. Pisoni, "Some acoustic cues for the perceptual categorization of American English regional dialects," *Journal of Phonetics,* vol. 32, no. 1, pp. 111-140, 2004.

[96] C. G. Clopper and J. C. Paolillo, "North American English vowels: A factor-analytic perspective," *Literary and linguistic computing,* vol. 21, no. 4, pp. 445-462, 2006.

[97] J. Prokić, J. Nerbonne, V. Zhobov, P. Osenova, K. Simov, T. Zastrow and E. Hinrichs, "The computational analysis of Bulgarian dialect pronunciation," *Serdica Journal of Computing,* vol. 3, no. 3, pp. 269-298, 2009.

[98] T. D. Do, V. Le, B. Bigi, L. Besacier and E. Castelli, "Mining a Comparable Text Corpus for A Vietnamese-French Statistical Machine Translation System," in *Proceedings of the 4th EACL Workshop on Statistical Machine Translation*, Athens, 2009.

[99] M. Mohammadi and A. N. Ghasem, "Building Bilingual Parallel Corpora based on Wikipedia," in *International Conference on Computer Engineering and Applicatons*, Bali, 2010.

[100] P. Resnik, M. B. Olsen and B. Diab, "Creating a Parallel Corpus from the Book of 2000 Tongues," in *Computer and the Humanities*, 1999.

[101] T. Mayer and M. Cysouw, "Creating a Massively Parallel Bible Corpus," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, Reykjavik, Iceland, 2014.

[102] K. Jones, S. Strassel, K. Walker, D. Graff and J. Wright, "Multi-language speech collection for NIST LRE," in *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia, 2016.

[103] B. Szmrecsanyi, Grammatical variation in British English dialects: A study in corpus-based dialectometry, Cambridge: Cambridge University Press, 2013.

[104] A. Fruttaldo, "Crawling in the deep: A corpus-based genre analysis of news tickers," in *Proceedings of Corpus Linguistics*, Lancaster University, UK, 2015.

[105] M. Y. Maris, The Malay sound system, Kuala Lumpur: Fajar Bakti Sdn. Bhd., 1980.

[106] C. K. Ajid, Dialek geografi Pasir Mas, Kuala Lumpur: Penerbit Universiti Kebangsaan Malaysia, 1985.

[107] A. S. Hasnudin, "Politeness strategies used by speakers of two Malay dialects," University of Malaya, Kuala Lumpur, 2012.

[108] D. Lindermann, "Bilingual lexicography and corpus methods. The example of German-Basque as language pair," *Procedia Social and Behavorial Sciences,* vol. 95, pp. 249-257, 2013.