Research Article

Kelantan and Sarawak Malay Dialects: Parallel Dialect Text Collection and Alignment Using Hybrid Distance-Statistical-Based Phrase Alignment Algorithm

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

[1]Faculty of Information Communication and Technology, Universiti Tunku Abdul Rahman, Kampar, Malaysia
[2]School of Computer Sciences, Universiti Sains Malaysia, Penang, Malaysia
[3]Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, Malaysia
tienping@usm.my[2]

Article History: Received: 10 November 2020; Revised: 12 January 2021; Accepted: 27 January 2021; Published online: 05 April 2021

Abstract: Parallel texts corpora are essential resources especially in translation and multilingual information retrieval. However, the publicly available parallel text corpora are limited to certain types and domains. Besides, Malay dialects are not standardized in term of writing. The existing alignment algorithms that is used to analayze the writing will require a large training data to obtain a good result. The paper describes our 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 of the parallel text. The proposed approach has a better precision and recall than the state-of-the-art GIZA++. In the paper, the alignment obtained were also compared to find out the lexical similarities and differences between SM and the two dialects.

Keywords: Malay dialects,parallel text,word alignment

1. 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 aregion.Dialectometry, a sub-component of dialectology, is "the measurement of dialect differences, i.e. linguistic differences whose distribution is determined primarily by geography".Many studies in dialect look at the phonological and phonetic differences between dialects. Heeringa(2004)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 (Wieling, et al., 2016). Dialects can also vary in the writing. For instance,Wieling et al. (Wieling, Montemagni, Nerbonne, & Baayen, 2014) investigate the differences in lexical between Tuscan dialects that is spoken in the area of central Italy and standard Italian. On the other hand, Grieve (2016) highlighted the regional variation in written American English.

Malay is a good case study for dialectometry as it presentsmany dialects. SM is from Johor, Riau dialect. The Malay dialects in Malaysia can be grouped based on their geographical distribution (Colins, 1989). Peninsular Malay dialects have been classified differently in the literatures (Onn, 1980; Asmah, 1991).This paper investigatestwo dialects: Kelantan Malay dialect (KD)from Peninsular Malaysia, and Sarawak Malay dialect (SD)from East Malaysia.In Malaysia, most of the works in dialectometry focus on the phonology aspect (Asmah, 1977; Abdul, 2006). In this paper, we look at dialectometry from the perspective of writing, particulary in lexical differences. The study of the lexical differences is interesting becausenative speakers communicate also through writing, besides speech,often in social media such as blogs and forums.

2. Methods For Building Parallel Corpus

Many parallel corpora have been created for various purposes. However, it happens often that the existing parallel corporado notfit the requested purposeof the user, or the user simply cannot afford to pay for the language resource.Therefore, the only solution is to build the parallel corpus.

2.1 Parallel corpusacquisition

The Web as a parallel corpus means that one webpage written in a source languagehas 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, 2003). A search tool will locate webpages that might have parallel translations by using different strategies,such as the structural relation between a parent webpage

and its sibling webpage, or heuristic information such as the date, file size comparison, and language markers in the HTML structureto reduce the scope of the search. An English-Malay parallel text was also constructed from the news articles (Yeong et al., 2019).

When the required data is not available on the Web, researchers need to either locate the data in different supports or construct a corpus from scratch. One interesting example is the Basic Travel Expression Corpus (BTEC) (Takezawa et al., 2002). The corpus contains more than 200 thousand common phrases and sentences in Japanese-English extracted from travelling phrase books. The initial project waslater 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 (Takezawa et al., 2007). The speech was transcribed and translated by an interpreter from Japanese to English or English to Japanese.

There were a few works that constructeddialect parallel corpora. Almeman et al. (2013) reported a parallel Arabic dialects speech corpora. 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 (Azham Hussain, et al, 2019). Another work is the parallel speech corpus for Japanese dialects (Yoshino et al., 2016). 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.

2.2 Data alignment

Alignment in machine translation involves identifying corresponding words between two sentences of different language that are translations of each other. Alignment algorithms can be divided to distance-based, statistical-based, neural networks, and heuristics.The distance alignment such as Levenshtein distance is used for string matching. The matching of two strings can be viewed as asequencealignment.From the perspective of alignment, the algorithm finds the maximum number of sequential alignments that can be formed.

The statistical approach is one of the most usedapproach 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.The E step thenestimates 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.Machine translation that based on phrase unitwas proposed by Koehn et al. (2003) 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.

Recently,many studies showed that neural networksproduce very good results insolving 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 anencoder-decoder architecture with attention mechanism(Bahdanau et al., 2014)was used for sequence-to-sequence modeling.The word/phrase alignmentin encoder-decoder networks can be visualized through the attention matrix.

The distance-based alignment algorithm, particularly Levenshtein distance algorithmis 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. Building Malay Dialect Parallel Text Corpus

In this paper, we propose to build a Malay dialect parallel text corpus by recording dialect dialogue, and then transcribing and translating the dialogue. The methodology used here is similar to Takezawa et al. (2007). The process goes through three main steps: recording dialect dialogues, transcribing the dialogues, and then translating the dialect transcriptionmanually to SM.

3.1Recording dialect dialogues

The dialogue recordings were conducted innoise free roomsat 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 seated in different roomsto avoid the speech to mix during recording. 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 recording is set at 16kHz/16bits per sample.Refer to Table 1.

Table 1. 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

The native dialect speakers then transcribed the speech in his/her dialect. The speakers will listen to the recording and then write them in words in his/her dialect and then translated to SM.Each dialogue consists of 200-400 sentences. Only 12 of the total 30 dialogues in KD were transcribed and all 8 dialogues in SD were transcribed as listed in Table 2. 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 sentence.

Table 2. Samples of the transcription and translation of the recording

KD and SM parallel sentence	**SD and SM parallel sentence**
KD:*Tehadiktawahebeykehtokletokgulo*	**SD:** *Zuladasikkitaknangga dalam Astro.*
SM:*Tehadik rasa tawar kerana terlupaletakgula.*	**SM:** *Zuladatakkamumenonton dalam Astro*

3.3 Aligning transcribed dialect words and phrases

The alignment of words and phrases is executed after acquiring the parallel sentences. We propose a hybrid distance-statistical-based phrasealignment algorithm that uses Levenshtein distance and statistical approach to align words and phrases automatically. The alignment algorithm was improved fromKhaw and Tan (Khaw & Tan, 2014) to include phrase matching. See Figure 1.

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

Figure1. Hybrid distance-statistical baseddialect phrase alignment algorithm

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 thetarget 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 before the distance of the words is calculated using Levenshtein distance. The parallel sentences used in the example are '*saya bawanasi.*' and '*kawebawaknasi.*' (English: I brought rice).Refer to Figure 2.

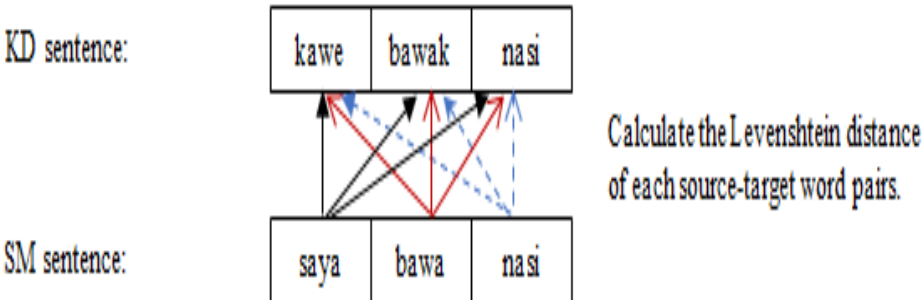

Figure2. Levenshtein distance comparison for a word in SM to all KD words

The Levenshtein ratio is then calculated for each source and target word pair using equation [1]. The word pair that has the lowestLevenshtein 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.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 tuples are listed below:

- (KD, SM): (mano, mana), (abe, abang), (naka, nakal), (pula, pulau), (anok, anak)
- (SD, SM): (pake, pakai), (pulo, pulau),(mberi, memberi), (ngisi, mengisi), (nyesal, menyesal)

$$ratio_{Lavens\ htein} = \frac{distance_{\ Levens\ htein}\ (w_s, w_t)}{lengt\ h(w_s)} \quad [1]$$

$$a(w_s, w_t)' = argmin\ ratio_{Lavens\ htein}\ (w_s, w_t) \quad if\ ratio_{Lavens\ htein}\ (w_s, w_t) < threshold \quad [2]$$

Step 2: Aligning non-similar words using pigeonhole principle

At this point, there might besome words in the target language (SM) that are not aligned to any word in the source language (Malay dialect). The source language wordthat 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*'.Some examples of unique dialect words extracted from the alignments are listed in (dialect, SM) tuples below.

- (KD, SM): (bokali, mungkin), (oyak, kata), (cakno, peduli), (hok, yang), (katok, pukul)
- (SD, SM):(molah, buat), (madah, beritahu), (sik, belum), (kamek, saya), (mun, kalau)

Step 3: Refiningalignment 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].

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

$$= \text{argmax}\ \frac{C(w_s, w_t)}{C(w_s)} \quad [4]$$

In equation [3], $w_s$is the source word andw_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.

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

A word can be translated using more than a word (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 set to two words, 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 (3)where W' is the most probable phrase.

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

A phrase formation threshold can be used to determine whether a phrase should be formed. If the (bigram) probability of a sequence is lower than the threshold, we assume it is not a valid sequence. A development set data can be used to estimate the threshold. We identified 55 of phrases of length two in KD, while 19 of phrases of length two are found in SD.Some examples of phrase obtained from this step are listed in (dialect, SM) tuple below:

- (KD, SM): (manih letting, sangatmanis), (tawahebe, sangattawar), (sesokdo'oh, sangatmiskin), (air batu, air sejuk), (sakni, tadi)
- (SD, SM): (duakigek, dua), (macamney, bagaimana), (ndakbrani, takutnya), (gineyginey, walaubagaimanapun), (tek dah, telahpun)

4. 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(Och and Ney, 2000).The calculation of the Levenshtein distance is time-consuming as it has the time complexity ofO(|VS|*|VT|*m*n), where |VS| is the size of the source vocabulary, |VT| is the size of the target vocabulary, m is the size of the source word and n is the size of the target word.After computing theLevenshtein distance, many alignmentswere 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|).

$$Precision = \frac{Number\ of\ Correct\ Alignment}{Total\ Number\ of\ Reference\ Alignment} \quad [6]$$

$$Recall = \frac{Number\ of\ Correct\ Alignment}{Total\ Number\ of\ Proposed\ Alignment} \quad [7]$$

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 thirty percent 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 3.

Table 3. 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 average precision and recall of the alignment between Malay dialect and SM obtained fromour proposed approach were 0.9542 and 0.9502 for KD, and 0.9503 and 0.9432 for SD. The overallresults 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 (Grimes et al., 2012). Example of

many-to-many (KD, SM) alignment in tuple obtained are: (tawahebe, sangattawar), (sesokdo'oh, sangatmiskin), and (manih letting, sangatmanis).

The alignment algorithm also clusters variants of the same word together. These variants in Table 4existbecause there is no standard orthography in the dialects.

Table 4. Examples of dialect word 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 5 shows the size of KDvocabulary and SD vocabularyextracted from the parallel text. The vocabularyis 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 vocabularyare 3237 and 2676 respecitively.The number of non-similar (unique) words in KD and SD are about12%. This indicates that about 10 percent 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 5. 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. Malay Dialect Lexical Analysis

This section examines the lexical similarities and differences between SM and Malay dialect through the analysis of similar wordsfound in word alignment.Many of the findingsare supported by the studies in Malay phonology and phoneticsindirectly in the literature.Phonology 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.

5.1 KDlexical 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 (Tan & Ranaivo-Malancon, 2009) 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.

Table6. 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 aSM 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 SMword 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'. If base word starts with a 'h', the letter 'h' is dropped.	*berlatih* *tertelan* *berikat* *berhulur* *terangkat* *terhanyut*	*belatih* *tetele* *berikat* *berulo* *terakat* *teranyut*	to train swallowed belted is giving upraised adrift
9.	**'e' of Prefix 'se-' Deletion**	The letter 'e' in the prefix 'se-' of a SM wordis 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.	yakahharapan	yokoharape	is it? hope
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**	In SM, the perfective marker *sudah* occurs before an intransitive verb.	*Diasudahmakan.*	*Diamakandoh.*	He has already

	Marker Position	In KD, the same perfective marker written as *doh* occurs after an intransitive verb.			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.	*Diasangatletih.*	*Dialetihsangat.*	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.

5.2 SDspelling analysis

In our analysis of SD, we found that thegraphemesin SD is the same as inSM. 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 finalletters of a word, 5 are performed on the prefix of a word. It does not show any changes in word order.

Table 7. 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**	Theletters '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

6. 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 ten percent 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.

7. Acknowledgements

The authors sincerely thank all the respondents who participated in this research. The authors also thank Universiti Sains Malaysia for the financial support [Project Number = 304.PKOMP.6316283].

References

1. Abdul, H.M. (2006). SintaksisDialek Kelantan. Kuala Lumpur: Dewan Bahasa dan Pustaka.
2. Almeman, K., Lee, M., &Almiman, A. A. (2013). Multi dialect Arabic speech parallel corpora. Proceedings of ICCSPA. Sharjah, United Arab Emirates: IEEE.
3. Asmah, H.O. (1977). The Phonological Diversity of the Malay Dialects. Kuala Lumpur: BahagianPembinaan dan Pengembangan Bahasa, Dewan Bahasa dan Pustaka.
4. Asmah, H.O. (1991). Aspek Bahasa dan Kajiannya: Kumpulan SiriCeramahPeristilahan. Kuala Lumpur: Dewan Bahasa dan Pustaka.
5. Azham Hussain, S.V Manikanthan, Padmapriya T. and Mahendran Nagalingam. "Genetic algorithm based adaptive offloading for improving IoT device communication efficiency". Wireless Network, August, 2019. DOI: 10.1007/s11276-019-02121-4.
6. Bahdanau, D., K. Cho and Y. Bengio. (2014). Neural machine translation by jointly learning to align andtranslate. In Proceedings of ACL-IJCNLP.
7. Colins, J.T. (1989). Malay Dialect Research in Malaysia: The Issue of Perspective, Bijdragen tot de Taal-, Land- en Volkenkunde: 235-264.
8. Grieve, J. (2016). Regional Variation in Written American English. Aston University: Cambridge University Press.Colins, J.T. (1989). Malay dialect research in Malaysia: the issue of perspective, Bijdragen tot de Taal-, Land- en Volkenkunde: 235-264.
9. Grimes, S., K. Peterson, K. and X. Li. (2012). 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.
10. Heeringa, W. (2004). Measuring dialect pronunciation differences using levenshtein distance. Netherlands: Ph.D. thesis, Rijksuniversiteit Groningen.
11. Koehn, P., F.J. Och and D. Marcu. (2003). Statistical Phrase-based Translation. In Proceedings of the Human Language Technology Conference: pp. 127–133). Edmonton.
12. Och, F.J. (2000), A comparison of alignment models for statistical machine translation. Proceeding of the 18th International Conference on Computational Linguistics, Saarbrucken.
13. Onn, F.M. (1980). Aspects of Malay phonology and morphology: a generative approach. Bangi, Universiti Kebangsaan Malaysia.
14. Wieling, M., S. Montemagni, J. Nerbonne and R.H. Baayen. (2014). Lexical differences between Tuscan dialects and standard Italian: Accounting for geographic and sociodemographic variation using generalized additive mixed modeling. Language, 90(3): 669-692.
15. Resnik, P. and N.A. Smith. (2003). The web as a parallel corpus. Computational Linguistics, 29(3): 349-380.
16. Takezawa, T.S. (2002). Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In Proceedings of the International Conference on Language Resources and Evaluation: pp. 147-152.
17. Takezawa, T., G. Kikui, M. Mizushima and E. Sumita. (2007). Multilingual spoken language corpus development for communication research. Computational Linguistics and Chinese Language Processing Journal, 12(3): 303-324.
18. Yeong, Y.-L, T.-P. Tan and K.H. Gan. (2019). A hybrid of sentence-level approach and fragment-level approach of parallel text extraction from comparable text, Procedia Computer Science, 161: 406-414.
19. Yoshino, K., N. Hirayama, S. Mori, K. Itoyama and H.G. Okuno. (2016). Parallel speech corpora of Japanese dialects. Proceedings of LREC 2016: pp. 23-28), Portorož, Slovenia.