Journal of Universal Language 8
March 2007, 129-159

Orthographic Reforms of Standard Malay Online: Towards Better Pronunciation and Construction of a Cross-language Environment*

Mohd Zaidi Abd Rozan & Yoshiki Mikami
Nagaoka University of Technology

Abstract

The widespread use of communication tools on the Internet has provided greater possibilities as well as increased flexibility for orthographic reforms. As a result, spellings have been established that are closely matched to 'written speech', which emulate verbal expression plus the *elegant* addition of letters from foreign languages. This study will look in particular at Evolutionized Malay (EM), which is cultivated on the Internet. Our analysis of EM reveals two principal functions: 1) EM words performed better in

* The authors would like to express their thanks to Professor Ahmad Zaki Abu Bakar (Universiti Teknologi Malaysia, Malaysia), Professor Robin Nagano Lee (University of Miskolc, Hungary), Dr. Chandrajith Ashuboda Marasinghe (Nagaoka University of Technology, Japan), Dr. Shigeaki Kodama (Asian Language Research Network), and Professor Brian Drier (Nagaoka University of Technology) for their careful attention, comments, and advice and also to Mr. Chew Yew Choong and Ms. Rizza Caminero (GII Lab, Nagaoka University of Technology, Japan) for their kind assistance.

displaying pronunciation properties and 2) EM words are able to form an environment we termed as Cross Language Environment (CLE). It may be that these roles hold a greater degree of attraction and motivates toward higher interaction within a cyberspace community. Surveys on the frequency of the occurrence of EM in the cyberspace of Malaysia are shown based on web forums and blogs pages. It is fascinating to see that EM has quite a substantial number of users throughout Malaysian cyberspace.

Keywords: Orthographic Reforms, Standard Malay (SM), Evolutionized Malay (EM), Cross Language Environment (CLE), pronunciation

1. Introduction

The pervasive nature of electronic communication due to the advancement of Information Communication Technology (ICT) has facilitated the creation of a new culture and lifestyle around the world that are more communicative in nature. Various forms of communication through various sophisticated devices including on-line communication have emerged giving flexibility, simplicity and connectivity to users without the constraints of time, space and form as in the past. The relative lower cost of on-line communication and its effectiveness versus face-to-face communication has contributed to this new development. To name a few, such online systems include computer based synchronous messaging (i.e., Yahoo! Messenger, IRC Chat), asynchronous messaging (i.e., LISTSERV, USENET) and short messaging system (SMS) used for mobile phone messaging.

However, such flexibility and simplicity still fail to provide the necessary functions that can be found in traditional methods of communications. This is because users' communicative options are limited by the nature of the input device, particularly the keyboard or keypad at their disposal. The set of characters physically found on

the keyboard will largely determine the linguistic capacity to produce information by key stroking the keyboard (Crystal 2001).

Production of messages also varies between traditional and electronic channels. Interactivity of online communications has pushed the production of more messages compared to offline channels (even for non-simultaneous modes such as web forums). There seem to be a sense of urgency to compose more messages as the pace of interaction increases. Therefore, while the posting of messages increases, the words that are used will often become compressed or truncated, to save time so as to keep abreast with such rapid transmission as well as to reduce the cost of communication which are sometime charged based on the number of characters in a message.

Looking at the practice of offline letter writing, it is a common scenario to discover certain words that represent a writer's daily aural style to be frequently used, for instance words like, *'wanna'* and *'gonna'*. Such aural style utilized is expected to adhere to standard language rules as both are found in English dictionary. However, in an online channel, nowadays, the ways words are spelled are becoming too flexible and casual, resembling spoken language. For instance, the use of *'wat'* for 'what', *'beeyooteeful'* for 'beautiful' and *'skool'* for 'school' were found in a English electronic discussion forum in Singapore's cyberspace (Ho 2005). According to an article (Baron 2002) that touched on English language, it was surprising to find educated speakers and writers that are not sensitive with the kind of language utilized in a technology driven communication. Here, by means of using *devil-may-care* spelling and punctuation, this truly described how chaotic computer-mediated communication (CMC) can possibly be. In this phenomenon, the spoken style when represented in written mode tends to disobey standard rules; hence, certain changes in terms of spelling and the creation of new words are expected. Most likely, online writing styles have now also infiltrated offline writing styles

(Lee 2002). Interestingly, proof from our observation, offline spellings have also started to show changes as online-styled spelling has started to influence them. This phenomenon can be seen in a case where many students were found to have used online versions of words for offline writings. This case was encountered by Malaysian teachers and was reported in an article titled "*SMS* (Short Messaging System) Ruins Language" (2004). In this article, non-standard spellings that are usually used online and for inscribing into short messaging system-SMS interactions on portable phones were found to be written by students in their Malaysian national level essay examination answers.

A very much-controlled language such as Malay, which has an official and standard version through years of language planning and engineering, also matures with development and progress of the communication channel. Stable and well-established linguistic rules for Malay have been put into practice. For example, the way the language is used during formal meetings is commonly anticipated by conforming to standard linguistic rules and norms that are mutually shared. In an online medium such as a web forum, it is again quite common to encounter unusual words that disobey linguistic rules. Web forum moderators are normally not fussy and impartial on the choice of words used since they usually have a disclaimer saying the responsibility and accountability lies with the users. They often only prohibit the use of abusive words (i.e., flaming), and such notices on ethics and *netizenship* are mostly emphasized in their web site's Frequently Asked Questions (FAQ) or disclaimer notices. As far as we know, notices or statements that stress on the importance adhering to standard grammatical or orthographic rules are not widely or commonly found written within the Malay language mediated web forums except for educational related sites.

In earlier paragraphs, some factors that have produced differences on how people express their words into their offline and

online messages were briefly described. Here, we will explore the phenomena of orthographic reforms for Standard Malay (SM) that is occurring in the Cyberspace. Special attention to the reforms that produced a breed of an unusual form of Malay language, termed here as the Internet Evolutionized Malay or in brief EM. Although there are several non-linguistic elements in EM, this paper only covers the issue of spelling reforms and excludes the usages of emoticons (or paralanguage).

The following sections will cover the background and motivational factors for this study and a brief historical overview of SM spelling reforms. This also covers several phases of evolution for the more casual spoken variation of Malay, Internet Evolutionized Malay (EM), and this is followed by a statistical snapshot of EM occurrence in Malaysian cyberspace collected from forum and blog pages. Finally, the role of EM as providing a better representation of the spoken mode and the creation of Cross Language Environment (CLE) are discussed. A discussion and conclusions are presented in the final section.

2. Background and Motivation

Since the independence of Malaysia in 1957, the establishment and development of Standard Malay (SM) has been under the supervision of a government body named *Dewan Bahasa dan Pustaka* (DBP), also known as the Language and Literary Agency of Malaysia. Standard Malay, Bahasa Melayu or Bahasa Malaysia as it is used interchangeably was established as the official language of Malaysia and is currently spoken by more than 23 million people in Malaysia. However, English is widely used as well as Chinese, Tamil, Arabic and other ethnic tribal languages of the indigenous people of Malaysia like the Iban, Melanau, Bajau, Kadazan and Kelabit. As an official language, Malay is also spoken in Southern

Thailand, Southern Philippines, Singapore, Brunei Darussalam and the rest of the Malay worlds, and in Indonesia, Malay is called Bahasa Indonesia.

Standard Malay (SM) is also widely used online and dominates government domain websites (.gov.my) and vast electronic publications that cover economic, social and political aspects. SM can also be found in other relatively formal writing as well. In recent years, the flexibility in writing, interacting and publishing online contents have transformed the way SM words are being spelled. As Crystal (2001: 2) explains, "When broadcasting enabled selected voices to be heard by millions, there was an immediate debate over which norms to use as correct pronunciation, how to achieve clarity and intelligibility, and whether to permit local accents and dialects". Just as norms for spoken language were debated then (in the 20th century), norms for written language are being examined now.

Based on Crystal's view, the availability of online communication has provided even more creative and liberal ways of interaction. This has brought about the birth of own-standard rules in spellings, and surprisingly still, the interactions are understood by both the writers and readers. The spelling reforms in SM do not occur in isolation; some other languages are in fact facing this phenomenon (Climent *et al.* 2003' Baron 2003). However, the alteration of the established Standard Malay (SM) into an evolutionized form of Malay language has brought about much displeasure, especially among language purists in Malaysia. In (1), we show an excerpt of a message taken from a web forum as a glimpse on how EM words appear online.

(1) Example of Evolutionized Malay Usage

> … dan aku pon balik kampong … sekembalinyer aku dari kampong … aku langsung tak tak call dier sbb aku raser aku tak cukup kuat untuk bermain kata dan persasaan denganyer

> … kenaper setiap akhir kata-kata berakhir dgn "abang tetap sayangkan adek"
>
> '… and I went back to my hometown … after coming back from there … I did not even phone (call) him because I feel that I am not strong enough to play words and feelings with him … why every single word must end with "I constantly care for you" '

Example (1) is an excerpt of a message written using mixed SM and EM words that was randomly selected from a Malay language web forum (http://s3.invisionfree.com/ClubLantui/ar/t156.htm). The EM words in order of appearance are; *pon* ('also'), *sekembalinyer* ('returned'), *dier* ('him/her'), *sbb* ('because'), *raser* ('feel'), *dengannyer* ('with him/her'), *kenaper* ('why') and *adek* ('younger brother/sister'). The SM spelling for the EM words as appeared above are; *pun, sekembalinya, dia, sebab, rasa, dengannya, kenapa and adik* respectively. It is visible from the excerpt that the writer of the message has written words differently in comparison to SM words.

Asmah says, "The Malays, as a race, would rather die than lose their language to a foreign one" (cited in Smith 2003: 63). This statement shows that the Malay language is highly honored by Malay people, and for example, manipulating by means of altering its spelling norms is judged as unwise. In addition to the unpleasant effect on SM, which is used as a mother-tongue language for Malay people, another concern will be on the unexpected changes in SM that could affect its function as a *lingua franca* within diverse ethnic groups in Malaysia; which served as the Malaysian national identity as well. Here, the negative consequences affecting SM sovereignty might be the cause for extreme concern.

For the past several years, the Malaysian media has actively publicized reports on EM, which deal with the negative aspects of

its existence in a superficial manner. Generally, most attention has focused on the event in which EM has badly affected SM morphology and spelling. Awang Sariyan, a Professor of Malay Linguistics at one of Malaysia's leading universities, stressed that "standard Malay (SM) is gradually becoming a victim since online technology provides autonomous writing freedom and most people seem to react positively to it" (Awang, personal communication, October 4, 2005). He also mentioned in a paper that 'the main concern would be the worsening of SM in terms of its status and quality as the emergence of new words are being used together with offline SM'.[1]

Looking from another perspective, the literature about the online evolution of English have stated many facts regarding the importance of understanding the whole chain of the reforms. (Kies 1999) stated that, although online communication is different from face-to-face language, most instructors fail to realize or even take advantage of the differences. Such differences exist due to factors like, situational, personal or topical (Tanaka 1998) that are influencing the way people communicate, whether in oral or written mode. It is also important and good for teachers to understand how students communicate and interact efficiently online especially when they are utilizing a virtual space (i.e., web forums, online chat, etc.) to deal with classroom tasks (McNaught, personal communication, April 13, 2006). In agreement with those points, Awang, from his sociolinguistic perspective, admitted that the issues of EM usages will produce potential research topics (Awang, personal communication, October 4, 2005).

On the other hand, such dynamic characteristics of spelling reform can be perceived as a fresh start injecting new sustainable

[1] The reference is to page 3 of Awang's paper "Strategy to Uplift the Standard Malay" circulated at the Meeting of Strategy to Uplift Standard Malay, hosted by the Ministry of Education in Putrajaya, Malaysia on December 21, 2004.

force into SM in this cyber age. While the dominance of other foreign languages (especially English) is highly visible online, still the EM that is vigorously used is the product of and a modified version of SM. This has shown that a new breed of language evolving from SM is being nurtured online without any sign of diminishing. Since it is impossible to control these new spellings as they progress online, based on our observation, there are reasons to believe that the utilization of EM will provide significant research findings and become useful for appreciating the factors affecting online communication and virtual communities. On top of that, a recommendation by Climent *et al.* (2003) about the challenges faced by machine translation system in parsing non-standard spellings in web documents that could affect the system performance itself, as many automated translation systems are not designed for this particular variation.

This EM phenomenon is not identical to another spoken (sometimes written) discourse style called *bahasa rojak* that usually appears by substituting a whole word with a non-Malay word (instead of just a few letters within a word).[2] Rais (2005) said, *bahasa rojak* is similar to a breed of *half snake and half eel*. He defined *bahasa rojak* as a language that uses mix words, for example, from Malay and English to generate phrases. For instance, the phrase 'sure *heboh*' means 'surely uproarious' and is formed by combining the English word 'sure' with a Malay word *'heboh'*. In SM, it is against the rules to mix words from different languages in a phrase, particularly when alternatives for the foreign words are available in SM. The proper phrase should be *'pasti heboh'* as the word 'sure' means *'pasti'* in Malay. The same phenomena were

[2] The word *rojak* originally appeared from a name of a famous dish that comprises of fruit or raw vegetable-salad that are mixed together in bowl with special sweet sauce, grounded peanut and chilli paste. Ingredients may vary among dishes and such varieties are usually taken as analogy to the mixture of different language in a text (Malay-English, Malay-Chinese, and others).

reported by Su (2003) in which several English expressions or words written in Latin characters have appeared within a Chinese character environment on electronic bulletin boards (BBS).

Before going into symbols or signs representing sounds and its spelling, it is necessary to adhere to rules for representing sounds (more precisely phonemes) and their written symbols (gloss). In this paper, sounds are shown in International Phonetic Alphabet (IPA) symbols between slashes (e.g., /ə/), while written symbols (letters) are shown here as follows: <e> or <a>.

The following section of this paper will provide a brief history regarding the issue of SM spelling reform in the non-virtual world. This is followed by information from Gani's short article on the analysis of the SM Internet evolution, and our snapshots on EM occurrences for Malaysian cyberspace based on Language Observatory (LO) project data (Mikami *et al.* 2005). Our main objective is to answer two research questions; 'How are Evolutionized Malay (EM) words able to represent the pronunciation of spoken Malay in a better way?' and 'How is Evolutionized Malay (EM) able to create a Cross Language Environment?'

3. Brief Historical Background on Standard Malay (SM) Spelling

In the year 1904, Wilkinson introduced the first spelling system for Malay language due to the wide use of Latin script for Malay inscription (Asmah 1989). This system was extensively used in Malaya (former Peninsular Malaysia before Malaysian independence), Singapore and Brunei. Prior to the use of Latin script, a writing system for Malay called *Jawi* was widely used, particularly for religious and literary traditions. *Jawi* script has been originally derived from Arabic script; however, due to some phonetic

properties in Malay, six *abjads* were created in addition to the 29 Arabic *abjads* (including one superscripted letter '*hamza*'). These were created by making slight modifications to the original corresponding Arabic *abjads* (Amat 1996). The derivations involved only adding an extra dot or dots in the upper, lower or inner positions of the *abjads*. More information is available in Daniels and Bright 1996. Since the establishment of Wilkinson system, the Malays have been writing their language in two completely different writing systems, *Jawi* and Latin (*Rumi)*.[3]

The Wilkinson system had to undergo major changes after being practiced for 20 years, when Zainal Abidin Bin Ahmad (Za'aba) started a reformation effort. Za'aba was a well-known grammarian who devised a plan to replace the vowel grapheme <u> with <o> in final closed syllables when the final consonant is represented by <k, h, ng> or <r>. He also replaced <i> with <e> in final closed syllables, where /k/ or /h/ is the final consonant (Asmah 1989). Some example of words affected are *'sepuluh'* ('ten'), *'ketuk'* ('knock'), and *'bilik'* ('room'), which were changed into *'sepuloh'*, *'ketok'* and *'bilek'*, respectively.

Even though at that time, Za'aba gave no special explanation for such changes, if we look properly at those reformed words, the changes allowed the spelling of the words to better reflect their pronunciation. However, it was believed that the phonetic realization (Asmah 1989) was to be the main reason for the change as opposed to Wilkinson's intention, which was to create coherencies of vowels with orthography. This is similar to the case

[3] Jawi is still used in Malaysia and some South East Asian countries. However, the popularity is lessening due to various factors. Simplicity for writing using Latin script instead of Jawi SM is one example. Johor Heritage Foundation (YWJ), a state-owned organization, has been extending great efforts to attract citizens especially young people to get involved in Jawi writing activities. For this, events that demonstrate the beauty of writing Jawi are held regularly in the state of Johor in the southern region of Malaysia (Abd Rozan 2005), where this organization is located.

of English and French as opposed to Spanish and Finnish, where the latter have a close grapheme-to-phoneme transcription (Divay & Vitale 1997).

Za'aba's proposals were adopted in schools for the teaching of Malay language from year 1930 until 1972 parallel with several other episodes of spelling reforms.[4] However, due to a comprehensive review in 1972, launched by DBP, a new spelling system that overruled Za'aba's was put into practice.

Anyhow, his proposal was devised more than 80 years ago, long before the existence of online communications. We cannot know how Za'aba would respond to the radical changes in SM spelling caused by the Internet if it had occurred while he was still around to deal with it. Perhaps if Za'aba's conventions were still in place, the evolution of Malay Language online might be just be a side issue, as his main point is to provide spellings of words that better reflect their pronunciation, as is happening now in online communications.

4. Evolutionized Malay (EM) on the Internet

Communication technologies are always compatible with SM and it is always possible to utilize only those text-based features to communicate in the Malay language. This is because SM words are inscribed using the 26 letters of Latin alphabets, known in Malay as *Rumi*, plus some non-alphabetic characters used mainly for reduplicated words, such as the hyphen and apostrophe for some Arabic terms (Li *et al.* 2005). Simply said, any English language-compatible machine is compatible for SM. A simple text-transmission protocol that is only based on the American Standard

[4] Spelling reforms have occurred not only for Malaysia but also for Indonesia. Efforts to replace British and Holland influences within Malaysian Malay and Indonesian Malay respectively were launched as a standardization plan for both countries.

Code for Information Interchange (ASCII) character sets is sufficient for SM communication, without even using the more complex extended version of ASCII which is the UNICODE.

In some cases, online variants occur because non-Latin based languages need to be transliterated into Latin based words in order to be operational within text communication technologies. One good example is the creative use of numbers to represent Arabic *abjads* that are not represent able by Latin script. For instance, in a transliterated Arabic word *'so2al'* ('question'), the presence of an Arabic numeral '2' is to represent the *abjad 'ain'* that functions as a glottal stop (Palfreyman and Khalil 2003). In the old Malay spelling, a glottal stop is represented by an apostrophe, for example in *'so'al'*, but nowadays it is simplified as '*soal*' based on SM style. Other examples are the use of Arabic numeral '7' to represent the *abjad 'ha'* such as in *'wa7ed'* ('one') and a numeral preceded by an apostrophe, '7', to represent '*kha*' such as '*7ebar*' ('news'). In comparison with SM, such similar sound to represent *'kha'* is written as *'khabar'* ('news').

Based on Linguasphere statistics by Dalby (1999), there are more than 160 million Malay language speakers all over the world and the majority are in four South East Asian countries, i.e. Brunei, Indonesia, Malaysia and Singapore. SM is an official language in Malaysia, Indonesia and Brunei. However, the pronunciation among Malay speakers varies between different countries (El-Imam and Don 2000). The greatest differences are between the pronunciation of Indonesian Malay and SM. For instance, *'gula'* ('sugar') is pronounced */gula/* by Indonesian speakers, while SM speakers pronounce it as */gulə/*.

The motivation for the emergence of EM is based in the informal nature of daily face-to-face conversation. This type of conversation is significantly used in spoken but not in written mode. However, because of the exclusion of face-to-face experiences and audio visualizations in the current online media, the modifications of the

Standard Malay spellings, as an effort to represent spoken words in a written manner took place. As a result, this colloquial style conversation has gained a written mode in addition to its spoken mode. This written mode that bears a resemblance to *spoken word* (*written speech*) style is what we have termed here as the Internet Evolutionized Malay (EM).

A typical socialization process in the verbal mode takes place when the sharing of experiences and information occurs by means of face-to-face meetings (f2f) or telephone conversations. According to Bradner *et al.* (1999), the *chat mode* can be evaluated as being much like a conversation. Since the wordings in EM are *chat mode*-like, the spelling structures are very much similar to the phonetic of spoken words. Here, the ability to describe the spoken element in written conversation is one advantage of EM. In this case, the spoken words that typed by a composer are meant to convey the aural effects similarly expressed in colloquial styles. The following paragraphs will explain this type of evolution occurring in SM.

Gani (2000) has identified the stages of SM evolution on the Web based on the orthographic characteristics of words. In Table 1, examples of SM transformation into its newly formed words are shown. The transformations occur in three phases. Evolution process 1 occurs due to the need to shorten words (economization). Here the words are always contracted and only the stressed syllable(s) are written down (sometimes with all vowel letters omitted). Still, native readers surprisingly understand the meaning of these newly formed words. Note that the economization factor is not covered in this paper. For evolution process 2, the transformation is motivated by the need to be similar to spoken style. Here, some words are shortened, but the number of syllables is quite similar to the original words. Evolution process 3 is similar to the second but with slight changes in spellings such as removing the last one or two letters and substituting new letters. Gani perceived this final phenomenon as melodically driven utterances. The possible reason for this is the

need to utilize friendlier expressions to indicate close contact for greater intimacy among speakers.

Similar to the evolutions in the previous paragraph, Herring (2001) also suggested three features mostly found in non-standard English e-mails. They are (1) economization factor to ease typing effort, (2) representation of typed-text emulating spoken style and

Table 1: SM Evolution on the Internet (adapted from Gani, 2000)

SM Words	Newly Formed Words	Evolution	Transformation Factor
Bagi ('give')	*Bg*	1	Word economization
Berhenti ('stop')	*Benti*		
Duduk ('sit*')	*Duk*		
Begini ('like this')	*Gini*		
Ambil ('take')	*Amik*	2	Similar to spoken style
Hantar ('send')	*Antar*		
Baca ('read')	*Bace*		
Serius ('serious')	*Seryuz*	3	Melodically driven utterance
Lepas ('ago*')	*Lepaz*		
Lupa ('forget')	*Luper*		
Cerita ('tell story')	*Citer*		

**Only one literal meaning without listing other available homonyms*

(3) creative expressions by a composer of the e-mail. In SM evolution, the melodically driven utterances suggested by Gani may hold a similar function with such creative expressions mentioned by Herring. Generally, both terms emit social signals that actively display the presence of *someone* in a community, and this is an important aspect in social discourse. Humans are social animals and the ability to be part of an online community is a critical and important toward a feeling of social presence (Xu 2005).

5. Snapshots of Evolutionized Malay (EM) Occurrences in Web Forums and Blog Pages

An experiment to determine the occurrences of EM by automatically counting the number of web pages in Malaysian cyberspace was performed. This experiment was based on two technical processes: (1) crawling cyberspace initiated by seed URLs and thus fetching web pages (Boldi *et al.* 2004) and (2) running a language identification engine to determine three aspects of the web pages, its LSE triplet (namely Language, Script and Encoding system). The huge data are stored in clusters of 40 servers at Nagaoka University of Technology under the management of the Language Observatory (LO) project (Mikami *et al.* 2005). The system based on N-gram technology (Suzuki *et al.* 2002) has been extensively used in the past four years in a project for LSE identification of huge domains.

To automatically identify the language of the fetched web pages, training data must be fed into the system. As for this moment, our identifier can identify more than 330 languages based on the overall number of trained data available that are essential for the system to learn and distinguish automatically. For this specific test, training data comprised of EM texts were also used. The output results were lists of Uniform Resource Locator (URL), the LSE and its matching codons (%). Filters were then applied to the output in order to retrieve only the highest matching percentage per page (return one language per URL). In this test, types of languages (based on ranks) and their occurrences (number of web pages) in web forums and blogs in the Malaysian country code (.my) Top Level Domain (ccTLD) were analyzed.

Our result in Table 2 shows the dominancy of English language pages in both web forums and blogs, with more than 50% coverage. EM ranked second with 14.3% and 27.4% for web forums and blogs respectively. As EM is often used for non-formal interactions, it is

not surprising to find its large occurrences in these two categories. Please note that the number of unique pages retrieved and analyzed from the .my domain came to a total of 3,410,914 pages. This total figure includes diverse types of pages, not only web forums and blogs which covers a portion of 21.1% of the whole pages for .my domain.

Table 2. Language and Occurrences in Web Forums and Blogs of Malaysian ccTLD (.my)

Language Rank & Number of Pages (%)						
Rank	**Web Forums**			**Blogs**		
	Language	Number of Pages	Percentage	Language	Number of Pages	Percentage
1	English	416,442	62.6%	English	32,844	59.7%
2	Evolutionized Malay (EM)	95,213	14.3%	Evolutionized Malay (EM)	15,047	27.4%
3	Chinese (Mandarin)	7,720	1.2%	Malay*	497	0.9%
4	Malay*	4,713	0.7%	Others	6,612	0.1%
5	Others	140,906	21.2%			
	Total	664,994			55,000	

**This includes Standard Malay and Indonesian Malay.*

The approach we took in classifying forums and blogs pages for this analysis was wholly based on the website's domain name. For example, if the word '*forum*' is found in the domain name, it will be classified as a web forum and if the word '*blog*' appears, as a blog. However, this method has less accuracy for producing classification compared to the *HTML template* and to a more advanced method called *genre analysis* (Santini 2006).[5] In this analysis, we are able to show at least a minimum number of the occurrences of EM for these two groups; there are in fact more EM pages in web forums and blogs than shown by the data provided in Table 2.

[5] This information was given to the authors via personal communication with K. Yamamoto, September 20, 2006.

6. Evolutionized Malay (EM) Characteristics

6.1 Contributing to Better Representation of Pronunciation

The sounds of phrases and words read from a passage and transmitted and collected by our ears are very different in comparison to the process of reading the same passage by sight. This is because written texts are not capable of efficiently bearing essential information possessed by its spoken counterparts and may lead to numerous losses of verbal information. For example, when the Malay language is uttered, we are able to hear that some syllables stand out above others. However, just by looking at the text, it is normally difficult to determine the start of a high pitch or stress marks since humans will generally read by pronouncing word-by-word or phrase-by-phrase based on their own style. Hence, they will mainly think that they are decoding the message in the text correctly. However, it is normal to find that different people disagree over the same message.

Another alternative to this problem would be to represent spoken phrases in written-text mode. What defines written-text mode is that the spelling of the words are constructed based on the way the words are exactly uttered, or also known as *written speech* or *spoken text* (Crystal 2001).

Based on evolutions 2 and 3 in Table 1, we found that EM has a distinct attribute in which it allows the construction of a specific word representing its exact intonation. We are not claiming that EM can replace sound-generated phrases, but the way words are spelled in EM indicates to us that it possesses *tonal* features that are not presentable in standard Malay spellings. While EM carries specific phonetic marks that act as attributes in transmitting special cues, this often signals how important certain information are conveyed. For instance, a reader reads words in a text and tends to emulate, or mimic, it by keeping to the suitable *rhythm* or *beat* (Abd Rozan &

Mikami 2006). This is because expressions are compounded superficially in EM words. This imitates a situation in which the reader is listening to the sounds of the words pronounced, and thereby promotes better pronunciation.

In Table 3, International Phonetic Alphabets (IPA) fonts are used as a transcription to describe the sounds when pronounced. To compare the pronunciation style, three types of articulation modes are shown: Written, Spoken and EM. For the Written style, words are spelled exactly as listed in the SM Dictionary; whereas for the Spoken and EM words, the spellings are normally altered. For example, the word '*kenapa*' 1 was transformed to *'kenape*' in spoken style (as a guide for pronunciation, spelling is actually non-existent) and *'kenaper'* in EM style.

Table 3. Word Pronunciation Style Based on Three Articulation Mode

SM Dictionary Words	**Articulation Mode**		
1. *Kenapa* ('Why')	**Written**	**Spoken***	**EM**
Spelling	*Kenapa*	*Kenape*	*Kenaper*
IPA Spelling	kənʌpʌ	kənʌpə	kənʌpə
Similarity of intonation pattern	Less similar to spoken	-	Similar to spoken
2. *Cerita* ('Story')			
Spelling	*Cerita*	*Ceite*	*Citer*
IPA Spelling	tʃərɪtə	tʃəɪtə	tʃɪtə
Similarity of intonation pattern	Less similar to spoken	-	More similar to spoken
3. *Macam ini* ('Like this')			
Spelling	*Macam ini*	*Camni*	*Camnie*
IPA Spelling	mʌtʃʌm ɪnɪ	tʃʌmnɪ	tʃʌmnɪ
Similarity of intonation pattern	Less similar to spoken	-	Similar to spoken

**Words are intentionally spelled this way, for pronunciation guidelines, as Spoken is not a written language.*

Similar to words no. 2 and no. 3, all these new orthographies represent a more expressive manner of *written speech*. The pattern that appears in IPA spelling shows that words in EM and Spoken share more closer resemblance in sound. By contrast, a typical Written style demonstrated a lesser degree of similarity compared with both EM and Spoken styles. A clear reason for this is that Written style (Standard Malay) does not regularly emulate the exact "sound" of the written words when pronounced.

For example, the word *'bace'* and *'baca'* (both meaning 'read') are spelled in EM and SM, respectively. Here, the <e> in *'bace'* is /ə/ whereas the ending <a> in *'baca'* is /a/. By substituting it with the letter <e> such as in EM, it is clear that the pronunciation of the letter 'a' is actually 'e', such as for 'e' in last syllable of the English word 'dinner'. The usage of EM words contributes to the certainty of pronouncing the ending letter <a> where the sound is rather /ə/ and not /a/.

For evolution no. 3, most words with a final syllable ending with <a> are changed to <er>. For example, the word *'kenapa'* ('why') is transformed into *'kenaper'* and *'jika'* ('if') to *'jiker'* and yet they closely follow spoken Malay, except for the presence of a letter <r> in the words. This is a very interesting phenomena, where Malay speakers habitually perceived this letter <r> as non-voiced. Based on our observation, these transformations are related to the English language influence. We dedicate the next subsection to explain our observations of these phenomena for both the occurrences of <er> and <z> in evolution no. 3 occurring in the last one or two letters of a word.

6.2 Formation of Cross Language Environment (CLE)

In the real world, as seen in (2), the development of Standard Malay involved three phases, starting with Archaic Malay evolving into Classic Malay and finally into Modern Malay (known as

Standard Malay-SM). Sanskrit was the most influential language in the derivation of Archaic, followed by the influence of Arabic and British English on Classic and Modern Malay, respectively. This explains why, there are many SM words that originated from Arabic, English, and Sanskrit. For example, the word *'bahasa'* ('language') was derived from the Sanskrit word *'bhasa'* and *'sengsara'* ('misery') from *'samsara'*. Arabic words *'kursiyun'* and *'shukran'* evolved into *'kerusi'* ('chair/stool') and *'syukur'* ('thankful') in Malay. Words like 'time', 'game' and 'torchlight' (the meanings are self-explanatory) from English language are spelled *'tem'*, *'gem'* and *'toclait'* with the same meanings, respectively, in Malay. However, to ensure that the loaned words comply with SM, strict rules from spelling conventions were applied to the original foreign derivations before they were registered as SM words. As a result, the syllables and shapes for selected loaned words from other languages were either purposely altered or preserved as is.

The previous paragraph explained some loaned words that are used in the real world as described in the upper row of (2). The lower row shows the occurrences in the virtual world where the use of Internet has affected the spelling of SM.

(2) Influential Factors on Malay in Real and Virtual Worlds

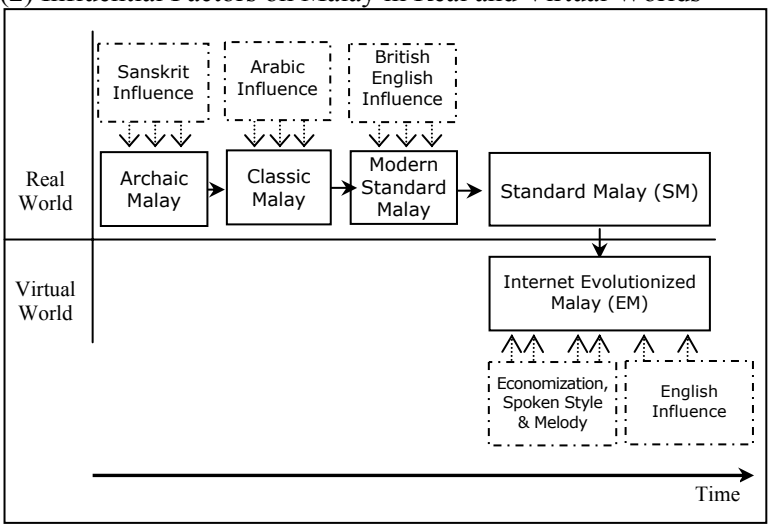


Table 4 describes the consonants of SM whose articulations have been influenced by English. However, this table only shows the effected consonants rather than all nineteen primary and eight secondary SM consonants (please refer to the complete table in El-Imam & Don 2000). It is interesting to note that English influence only occurred in some consonants and vowels of SM. A single exception occurred in the case of the word *'serius'* ('serious') which was transformed to *'seryuz'* where the vowel 'iu' was replaced by 'yu'; the reason for this anomaly is under investigation by the authors. The two basic types of alterations are substitution or addition of one letter. Again, for SM words ending with the letter <s>, substitution with <z> always took place. However for the letters <p>, <b>, <m>, <t>, <n>, <k> and <ng>, the addition of the letter <z> after those letters normally occurred.

Table 4. Consonants of SM that are Influenced by English language

Place & Manner of Articulation	**Bilabial**	**Alveolar**	**Velar**
Oral Stop	/p/ e.g., SM *hara/p/* (‘hope’) e.g., EM *hara/pz/* /b/ e.g., SM *aza/b/* (‘punishment’) e.g., EM *aza/bz/*	/t/ e.g., SM *pena/t/* (‘tired’) e.g., EM *pena/tz/*	/k/ e.g., SM *peli/k/* (‘strange’) e.g., EM *peli/kz/*
Nasal (stop)	/m/ e.g., SM *dala/m/* (‘in/inside’) e.g., EM *dalam/mz/*	/n/ e.g., SM *pada/n/* (‘match’) e.g., EM *pada/nz/*	/ng/ e.g., SM *aban/g/* (‘brother’) e.g., EM *aban/gz/*
Fricative		/s/ e.g., SM *lepa/s/* (‘let go’) e.g., EM *lepa/z/*	

Table 5 describes the influence of English sounds on two vowels in SM, /i/ and /ə/. Again, only the effected vowels are shown. The two basic types of alterations are, i) addition and ii) substitution + addition, of one letter. For SM words ending with the letter <i>, the addition of <e> took place. However, for words with the final letter <a>, substitution with <e> together with addition of the letter <r> normally occurred.

Table 5. Vowels of SM that are Influenced by English

Tongue Position	**Front**	**Central**
High or Closed	/i/ e.g., SM *har/i/* (‘day’) e.g., EM *har/i/e*	
High-mid or Half-closed		/ə/ e.g., SM *bung/a/* (‘flower’) e.g., EM *bung/ə/r*

The final syllable for words ending with <er> is rather rare in SM, but from our data it appears that, the frequency of occurrences is rather high as EM users tend to type <er> to replace words that end with <a>. Based on our observation, these two letters are derived from English because it resembles the appearance of many words in English registries. For example, the word '*sayer*' ('I/me') was originally derived from '*saya*', and the last syllable of English words like *'later'* and *'catcher'* have the same *'er'* sound as *'sayer'*. This also holds true with all the EM words ending with <er>, for instance, '*biler*' (when) and '*bagaimaner*' (how).

From language universality point of view, the commonalities between EM and English can be found in both the orthography and in the sound of the last syllable of many word occurrences. However, this pattern is uncommon in SM. Only a handful of words in SM are found to have <er> in the last syllable, which produces two different sound varieties. For instance, in the word *'pamer'* (display), the last syllable is pronounced similar to <ma> in an English word 'may'. In contrast to that, the word *'koreografer'* (choreographer), which was actually a foreign word in SM, has the same final syllable sound with its corresponding English word. Another aspect is the regular use of plural indicators such as <s> or <z> for Malay words, even if such indicators are not in favor of the grammatical rules of Malay language.

From Table 6, we can see that the number of operations needed to transform SM to EM words is mostly minimal. Edit distance (also known as Levenshtein distance) is useful to indicate the similarities of two words based on the number of operations for deletion, insertion or substitution of one single character (Gilleland 2006). Here, the highest edit distance is two units, and the lowest is one. In general, this shows that only a simple step is needed to change the appearance of SM words into EM words, and this is generally done by adding or substituting only one letter or maximum of two letters.

Table 6: SM, EM Words and Edit Distance

SM Words	EM Words	Edit Distance
Lepas ('let go')	*Lepaz*	1
Serius ('serious')	*Seryuz*	2
Asyik ('spellbound')	*Asyikz*	1
Batas ('limit')	*Bataz*	1
Dayus ('unmanly')	*Dayuz*	1
Mampus ('dead')	*Mampuz*	1
Saya ('I/me')	*Sayer*	2
Bila ('when')	*Biler*	2
Pula ('also')	*Puler*	2

Looking at the characteristics of EM as presented in subsection 6.1 and 6.2, the *striking* appearance of EM words by just a slight modification in their spellings may hold a greater degree of attraction and could be a motivation for higher interaction by the creation of a special environment. This has led to what we have termed a Cross Language Environment (CLE). Here, the SM spelling is altered to create special effect and to add some flavor of *exoticism* and *weirdness*. In whatever condition, the initial part of the words can still be recognized and found in the SM dictionary. The transformations of SM to EM mostly affect the ending of a word, at which point verbal stress is normally applied. The sounds produced by this consonant are stressed in speech and given stronger impressions not only on the *ear* but also on the eye.

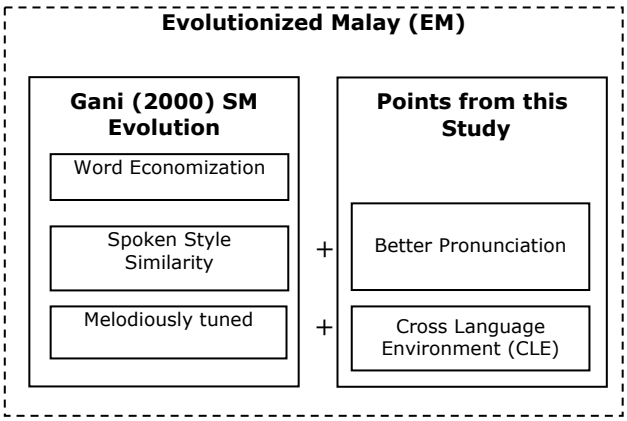
The sense of *flavor* from other languages shows that some pre-knowledge or understanding in the utilized language (particularly English) is rather necessary in order to create such words. For the readers, it may not be unpleasant even if they do not have a basic English background since the words are comprehensible and since their main components are still in Malay. Another possibility for such word modification is that, the author of a message will think in

this way, "I write what I hear, but there is option in spelling, and I choose a spelling that could attract readers because of its exotic appeal". Here the author is trying to express a new concept symbolized by a new spelling to be shared with and by all members as this is a kind of social presence indicator.

7. Discussion and Conclusion

The aim of this paper is to provide discussion in the effort to understand EM as the product of the SM online evolution. Here, we further elaborate on Gani's (2000) statements about the SM evolution by providing detailed analysis of our two novel points, which are a better pronunciation and Cross Language Environment (CLE) as shown in (3).

(3) Evolutionized Malay (EM) from the Combination of Gani SM Evolution with Elements of Better Pronunciation & Cross Language Environment



First, we have explained that the orthographic properties of EM work well by virtue of having a spelling style that is capable of representing spoken Malay pronunciation in a better way. It has long been understood that the oral mode was only possible for people communicating across a close distance, but here it is shown that it is possible to draw many people with similar interests into a virtual space and have them *talking* in their *oral style* but in written mode. This has the potential to provide the kind of platform that motivates sharing and knowledge creation during the course of interaction.

Second, the formation of a Cross Language Environment (CLE) by having *exotically* spelled words blended with letters inspired from the English language could provide a greater level of attraction. One important point to note is that the birth of EM is of a unique kind, because only a few letters within a word are affected as compared with the earlier explained *bahasa rojak*. The display of new orthographies as an *'eye-candy'* could excite members in a community and they may feel like participating or just lurking. EM could act as a catalyst for a greater tendency toward interaction: if the surroundings are dull, no one is likely to join. People will likely sense that it is a wonderful space to build closer contact.

Although further research is required, particularly on the psychological aspects of these two points in online communication, still the proliferation of EM is unquestionable.

Based on the findings presented in this paper, there are two implications towards typology and the universality of languages. First, the commonality of spellings particularly for the last syllable of EM words, which goes along with English words (e.g., <er> in *'biler'* and 'later' and <z> in *'makanz'* and *planz* 'plans') exhibit that the influence of English (either as its standard form or online form) as a dominant language is real. It shows, at least in this paper, that a *pop* culture on the web for written Malay is mostly prompted by the user's *know-how* of the English language. These aspects are closely related to the universality of such features found in EM and

English language.

Secondly, although the semantic meaning is the same for EM and SM words (e.g., *'sayer'* with *'saya'* (I/me) and *'kenaper'* with *'kenapa'* (why)), the utilization of EM may well represent better expressions as it resembles *spoken style*. Such dissimilarities in spellings are important and could affect language typology, particularly of SM and EM words, which have the same semantic meaning but contributing towards the differences in perceived expression.

We would also like to mention two of our next study topics to advance our research. First, another important step in our research is to investigate user experiences, particularly in relation to the perceived desirability for using and manipulating EM on the Web. This should also include experiments on the differences of emotion expressions conveyed between the usage of SM and EM. This could yield significant results and could project many useful Language-Discourse-Emotion (tripartite) correlations.

Second, a survey on the coverage of EM in cyberspace must be further improved. This should cover greater populations such as determining the breakdowns of EM utilization in three other Malay language user countries: Singapore, Brunei and Indonesia. This should also include coverage for any possible varieties of EM rooted from different dialects all over Malaysia. In the future, we are planning to perform EM identification for these three ccTLDs, generating more in-depth analysis covering almost all possible secondary domains; next, we hope to report a detailed analysis of EM dispersion.

The driving force of EM in online communications is a strong research topic because understanding it will provide insights allowing us to formulate wiser steps in response to the '*threat'* which EM is often viewed. The issues are surrounding the growth of EM are socially and communicatively challenging because they can provide strong support for further SM development or even a

negative reaction.

References

Abd Rozan, M. 2005. *Malaysia Trip Report: January 2005*. Language Observatory Project. Japan Science & Technology Agency.

Abd Rozan, M. & Y. Mikami. 2006. Bahasa Sembang in Web Forums: Knowledge Management for Piles of Atopian Discourse. *International Conference on Web Information Systems & Technologies* 119-122.

Amat, J. 1996. *Language Planning: History of Jawi Characters*. Selangor: Language and Literary Agency of Malaysia.

Asmah, H. 1989. The Malay Spelling Reform. *Journal of the Simplified Spelling Society* 9-13.

Baron, N. 2002. 'Whatever': A New Language Model? *Proceedings of Convention of Modern Language Association, New York, December 27-30.*

Baron, N. 2003. The Language of the Internet. In A. Farghali (ed.), *The Stanford Handbook for Language Engineers* 59-127. Stanford, CA: CSLI Publications.

Boldi, P., B. Codenotti, M. Santini, & S. Vigna. 2004. UbiCrawler: A Scalable Fully Distributed Web Crawler. *Software: Practice & Experience* 34.8, 711-726.

Bradner, E., W. Kellogg, & T. Erickson. 1999. The Adoption and Use of BABBLE: A Field Study of Chat in the Workplace. In S. Bødker, M. Kyng, K. Schmidt (eds.), *Proceedings of the Sixth European Conference on Computer Supported Cooperative Work-ECSCW 99*, 139.

Climent, S., J. More, A. Oliver, M. Salvatierra, I. Sanchez, M. Taule, & L.Vallmanya. 2003. Bilingual Newsgroup in Catalonia: A Challenge for Machine Translation. *Journal of Computer-mediated Communication* 9.1. Available at URL <http://jcmc.indiana.edu/vol9/issue1/climent.html>.

Crystal, D. 2001. *Language and the Internet.* Cambridge: Cambridge

University Press.

Dalby, D. 1999. *The Linguasphere Register of the World's Languages and Speech Communities Volume* 1. Wales: Linguasphere Press.

Daniels, P. & W. Bright. 1996. *The Worlds Writing System*. Oxford: Oxford University Press.

Divay, M. & A. Vitale. 1997. Algorithms for Grapheme-phoneme Translation for English and French: Applications for Database Searches and Speech Synthesis. *Computational Linguistics* 23.4, 495-523.

El-Imam, Y. & Z. Don. 2000. Test-to-speech Conversion of Standard Malay. *International Journal of Speech Technology* 3, 129-146.

Gani, M. 2000. Standard Malay Internet Evolution. Available at URL <http://www.bahasawan.org/isian/SM-net.htm>.

Gilleland, M. 2006. *Levenchstein Distance, in Three Flavors*. Available at URL <http://www.merriampark.com/ld.htm>.

Herring, S. 2001. Computer-mediated Discourse. In D. Tannen, D. Schiffin, & H. Hamilton (eds.), *Handbook of Discourse Analysis* 612-634. Oxford: Blackwell…

Ho, C. 2005. Many in One-One in Many: Towards Construction of Community in an Electronic Discussion Forum. *Journal of Language and Learning* 3.1, 1-35.

Kies, D. 1999. *Metalearning: Learning How to Learn in the New Media*. Available at URL <http://papyr.com/web99/>.

Lee, J. 2002. I Think, Therefore IM. *New York Times*, September 19, E1.

Li, H., A. Mahani, & B. Teoh. 2005. A Grapheme-to-phoneme Converter: A Rule-based Approach. *Conference Paper Presented in Proceedings Oriental-COCOSDA 2005*. Jakarta.

Mikami, Y., P. Zavarsky, M. Abd Rozan, I. Suzuki, M. Takahashi, T. Maki, I. Ayob, P. Boldi, M. Santini, & S. Vigna. 2005. The Language Observatory Project (LOP). *In Special Interest Tracks & Posters Proceedings of the 14th International Conference on World Wide Web* 990-991.

Palfreyman, D. & A. Khalil. 2003. A Funky Language for Teenzz to Use: Representing Gulf Arabic in Instant Messaging. *Journal of Computer-*

mediated Communication 9.1. Available at URL <http://jcmc.indiana.edu/vol9/issue1/palfreyman.html>.

Rais, Y. 2005. Writers' Talk: Between the Carried and the Pursuit. Available at URL <http://www.bharian.com.my/Misc/DBP/Artikel/Bahasa/20050602165216/ArtBahasa>.

Santini, M. 2006. Some Issues in Automatic Genre Classification of Web Pages. Presented in *JADT 2006—8èmes Journées internationales d'analyse statistique des données textuelles à l'université de Besançon.*

Smith, K. 2003. Minority Language Education in Malaysia: Four Ethnic Communities' Experiences. *International Journal of Bilingual Education and Bilingualism* 6.1, 52-65.

SMS (Short Messaging System) Ruins Language. Available at <URL http://aplikasi.kpkt.gov.my/akhbar.nsf>.

Su, H. 2003. The Multilingual and Multi-orthographic Taiwan-based Internet: Creative Uses of Writing Systems on College-affiliated BBSs. *Journal of Computer-mediated Communication* 9.1. Available at URL <http://jcmc.indiana.edu/vol9/issue1/su. html>.

Suzuki, I., Y. Ikami, A. Ohsato, & Y. Chubachi. 2002. A Language and Character Set Determination Method Based on N-gram Statistics. *ACM Transactions on Asian Language Information Processing* 1.3, 270-279.

Tanaka, S. 1998. *Varieties and Universals of Language*. Nagoya: Chunichi-Bunka Co., Ltd.

Xu, Y. 2005. Creating Social Presence in Online Environment. In B. Hoffman (ed.), *Encyclopedia of Educational Technology*. Available at URL <http://coe.sdsu.edu/eet/articles/creatsp/start. htm>.