# AUTOMATIC IDENTIFICATION OF CLOSE LANGUAGES: case of Malay and Indonesian

*Bali Ranaivo – ranaivo@cs.usm.my*
*Siti Khaotijah Mohamad – sitijah@cs.usm.my*
*Computer Aided Translation Unit*
*School of Computer Sciences*
*Universiti Sains Malaysia*
*11800, Penang, Malaysia*

**Abstract - Résumé**

The Malay used in Malaysia and the one used in Indonesia are very closed to each other since they both have the same common origin, the bahasa Melayu. Despite this fact there exist differences which are to be found in the spelling, lexical and morphological levels. Any and all correction, translation, understanding of a Malay text without knowing whether it is Malaysian or Indonesian may lead to serious errors or to the use of irrelevant program. For these reasons we intend to create a program able to determine whether a text is Malay or Indonesian using the said differences according to the digressive order of their reliability. This research being at its very beginning, we shall, in this article, only present the descriptive part of the differences existing between Malay and Indonesian.

*Le malais utilisé en Malaisie et celui utilisé en Indonésie sont très proches puisqu'ils sont issus de la même langue, le Bahasa Melayu. Malgré cette parenté, il existe des différences qui se situent aux niveaux orthographique, lexicale et morphologique. Toutes correction, traduction, compréhension d'un texte malais sans savoir s'il est malaisien ou indonésien peut conduire à des erreurs importantes ou à l'application de programme inadéquat. Pour toutes ces raisons, nous projetons de créer un système capable de déterminer si un texte est malais ou indonésien en utilisant ces différences selon leur ordre de fiabilité: du plus sûr au moins sûr. Cette recherche n'étant qu'à son début, nous ne présentons dans cet article que la partie descriptive des différences qui existent entre le malais et l'indonésien.*

**Key words - Mots clefs**

Automatic Language identification, Malay, Indonesian.

*Identification automatique des langues, malais, indonésien.*

## 1. Introduction

For several centuries, the Malay language or Bahasa Melayu, has been used as the lingua franca of various ethnic groups in the Malay archipelago. Today it is known as Bahasa Melayu (BM) in Malaysia (after 1963) and as Bahasa Indonesia (BI) in Indonesia (after 1928) and it is still known as Bahasa Melayu in Singapore and Brunei Darussalam. BM and BI originate from the same language and remain mutually intelligible. However, because of different socio-political developments in Malaysia and Indonesia during the last one hundred years or so, the two varieties of Malay spoken in the two countries have developed differently, each influenced by different factors resulting in many noticeable differences.

For example, because of different colonizer, the Malay language of Malaysia contains more English loan words and the Indonesian, Dutch loan words. Besides, the impact of Islam among the Malays in Malaysia is much more entrenched in BM than in BI. In Indonesia the use of Arabic is more limited. Instead, western and local vocabularies have constantly enriched the

Indonesian language because of the influence of dialects.

With the rapid development of Malaysian and Indonesian web sites and the expansion of multilingual text corpus, the number of documents writen in Malay (BM and BI) increase every day. Then a question arises on how to identify the closed languages like BM and BI and how to process those documents automatically.

Therefore our aim is to build one rapid and efficiency program that able to determine whether one written text said Malay text is Malaysian or Indonesian. We could reach our object if we classify and order those differences.

Before enumarating the differences between BM and BI, we'll cite the works that have already done to identify Malay texts.

## 2. Language identifiers for Malay texts

We query on the Web by using the search browser Google "Malay language identification" and "Indonesian language identification". For the first query we got 4,480 answers and for the second 10,900 answers. As usual, we got a lot of answers related or not to our queries. But among all those miscellaneous answers we find few language identifiers for Malay and for Indonesian. One of them is the Xerox's language identifier[1] that recognise 47 languages of which Malay and Indonesian. The demo is available on line. We have tested it by using some of the criteria used for us to identify Malay from Indonesian. For example, we submit the sentence *Saya sukar bepergian*. The system recognised it as a Malay text and not Indonesian. In Malay, the equivalent word of *bepergian* is *berpergian*. Without testing all systems that exist to identify one of the two languages, BM or BI, we can say that most of them don't really make the difference between BM and BI. The criteria they use can recognise Malay language meaning BM and BI.

[1] http://www.xrce.xerox.com/cgi-bin/mltt/LanguageGuesser

## 3. Spelling system

In 1972, Malaysia and Indonesia announced officially the implementation of the New Romanized Spelling System. Before this date, the two countries had their own spelling system influenced by English in Malaysia and by Dutch in Indonesia.

BM and BI have respectively 9 vowels and 26 consonants. The grapheme < e > represents two phonemes: /ə/ (schwa) and /e/.

| | | |
|---|---|---|
| Vowels | One letter | a, e, i, o, u |
| | Two letters | ai, au, oi |
| Consonants | One letter | b, c, d, f, g, h, j, k, l, m, n, p, q, r, s, t, v, w, x, y, z |
| | Two letters | gh, kh, ng, ny, sy |

**Figure 1: Graphemes system in BM and BI**

If we can say today that this reform spelling is a success, few parts of those Dutch and English spelling systems still remain.

### 3.1. Combined letters

In BI, the spelling of some proper names used under Dutch colonization is still on. This concerns the sounds /u/, /dz/ and /ʧ/. Before the reform spelling, they were written differently in Malaysia and in Indonesia.

| | BM | BI | New Reform Spelling |
|---|---|---|---|
| /u/ | u | oe<br>*Soekarno* | u<br>*Sukarno* |
| /dz/ | j | dj<br>*Dardjowidjojo* | j<br>*Darjowijojo* |
| /ʧ/ | ch | tj<br>*Darmasoetjipta* | c<br>*Darmasucipta* |

Some of borrowed words from Dutch show specific ending when they are retranscribed into BI. For example, all words ending with the trigrams < oar > find in Malay texts come from Dutch: *supletoar* 'supplement', *repertoar* 'repertoire', *trotoar* 'pavement'. BM use other words for the same meaning: *penambah, himpunan, laluan jalan kaki.*

### 3.2. Apostrophe ' and inverted comma ‘

Apostrophe and inverted comma are used equally in BM and BI. The rule doesn't precise when those marks could be used. It says only that "an apostrophe is the punctuation mark used to show the omission of a letter or letters". So in BM and BI we can find those kinds of reductions in one text:

*I'gris* (= *Inggeris*) English
*P'cis* (= *Perancis*) French
*Ali 'kan* (= *akan*) *kusurati.*
I'll write to Ali.

In Indonesia, the apostrophe can be also used to show the omission of numerals:
*1 januari '88* (= *1988*)

In Malaysia, there are two words in which the apostrophe or the inverted comma is used: Za'ba or Za'ba[2] and Dato'[3]. They are the vestiges of former spelling system: the apostrophe is used to mark the glottal stop.

### 3.3. Reduplicating symbol '2'

The Malay language has different forms for reduplicating words (cf. Asmah Haji Omar, 1975:185-223, "Reduplication in Malay"). One of those forms is the whole reduplication, where the whole word is repeated twice by inserting a hyphen between the two. Sometimes those whole reduplicated words are written differently. The word is written once, the hyphen disappears, and the number '2' follows the word. The difference between BM and BI is the position of this numeral. In BM, '2' occupies the same baseline as the word reduplicated. In BI, it is raised as the square symbol '²'.

| BM | BI | |
|---|---|---|
| *buku-buku* | | books |
| *buku2* | *buku²* | |

### 3.4. Spelling of numerals: the use of full stop and comma'

In BM and BI, the use of full stop and comma are inverted. If in BM, the comma is used in digits to indicate numbers in thousand, million and so on, in BI, it is the full stop that plays this kind of role.

BM

| | |
|---|---|
| two thousand | 2,000 |
| two thousandths | ke-2,000 = *kedua ribu* |
| two thousand ringgit | $2,000.00 |

BI

| | |
|---|---|
| two thousand | 2.000 |
| two thousandths | ke-2,000 = *kedua ribu* |
| two thousand rupiah | Rp2.000,00 |

## 4. Morphology

Because of their relationship, the difference between BM and BI at morphology level is very slight. We have found only two differences. The first one is about the rules with the prefix {beR-} and the circumfix {beR-an}. The second is about abbreviations.

### 4.1. Affixation with {beR-}, {beR–an}

The affixation with {beR-} and {beR-an} depends on the initial phoneme type of the root. If it is a /r/, the /r/ of the prefix and the

[2] It's the name of a Malaysian grammarian; we can find also his name written: Zaaba.
[3] It's the synonym of *Datuk*, a honorific Malaysian title.

/r/ of the root fuse and become one /r/. In other contexts, the prefix is just put before the root without any changing. But in BI there is another context for those two affixes. With some roots in which the first syllable is closed with /r/, the prefix has the form "be-".

1) {beR-} → be / __/r/
{beR-}+ RUMAH ⇒ berumah

2) {beR-} → ber / __other context
{beR-}+IBU ⇒ beribu
{beR-}+SURAt ⇒ bersurat

Those two rules are common for BM and BI. The third rule is specific to BI.

3) {beR-} → be / __/CVr/
{beR-}+SERTA ⇒ beserta

### 4.2. Abbreviations

Malay people seem enjoying to shorten words. They use them very often so for someone who learns Malay it becomes very fuzzy to understand the meaning of the sentence. For our research, we will use those words to disguish BM and BI.

Examples of abbreviations in BM
*KDN = Kementerian Dalam Negeri*
The Ministry of Domestic Affairs
*Pernas = Perbadanan Nasional*
National Consortium
*tadika = taman didikan kanak-kanak*
Kindergarten

Examples of abbreviations in BI

*ABRI = Angkatan Bersenjata Republik Indonesia*
Army of Republic of Indonesia
*Akabri = Akademi Angkatan Bersenjata Republik Indonesia*
Academy of Indonesian Army
*pemilu = pemilihan umun*
general election

## 5. Lexicon

| | |
|---|---|
| Data | >< 10 000 words |
| Redundant | >< 7 694 words |
| Wordlist | >< 2 306 words |
| Same BM/BI words | >< 1 605 |
| BI words only | >< 701 |

**Figure 2: Data statistic**

In our data that we have retrieved from Media Indonesia Online, dated on August 8, 2002, they used 2 306 words to write the news with total of words is about 10 000 words. From 2 306 words, we got about 701 words that used only in BI. This mean 1605 words are the same words used in BM. Therefore we use this 701 words as our main guide to identify the difference of lexicon between these two languanges.

The lexicon differences can be grouped under five main aspects:

**1)** Words with different spellings

| BM | BI |
|---|---|
| *televisyen* | *televise* |
| *stesen* | *stasiun* |
| *kuiz* | *kuis* |
| *kerana* | *karena* |
| *bahawa* | *bahwa* |
| *Ogos* | *Agustus* |
| *akhbar* | *akbar* |
| *bahagian* | *bagian* |

BI used the Dutch spelling while BM used English.

**2)** Words with different meanings

| | BM | BI |
|---|---|---|
| *pejabat* | office | government officer |
| *sulit* | secret | difficult |
| *boleh* | can | permisable |
| *polisi* | policy | police |
| *bisa* | venom | can |

It is true that a much larger number of words in BM and BI have the same meanings. However, there are also many words with

various meanings and the first or more important meaning in BM often differs from that in BI.

**3)** Different words for one thing /concept

| BM | BI | Meaning |
|---|---|---|
| *ikan bilis* | *ikan teri* | anchovies |
| *tuala* | *handuk* | towel |
| *bilik* | *kamar* | room |
| *dibenarkan* | *diizinkan* | permitted |
| *rancangan* | *acara* | programme |
| *maklumat* | *informasi* | information |
| *ponteng* | *absen* | absent |
| *itik* | *bebek* | duck |

Most of the 701 words that have been mentioned before are included in this type of differences.

**4)** Different loan words

BM: *abdomen, abstrak, absurd, birokrasi*

BI: *adhesi, administrasi, adopsi, mobil, direktur*

BM has adopted many Briitish words while BI has incorporated many Dutch words.

**5)** Different title as an indicating rank, status

BM: *En. Zakaria Idris, Pengurus Besar*

BI: *Sdr. Zakaria Idris, Direktur Utama*

In Malaysia there are many titles to addressing someone depend on their rank, status and age but may be not so many in Indonesia.

**Building proper lexicon**

We can see through those lexicon differences that each language has developed its own lexicon by borrowing, shorting words or derivating. Let's call this kind of lexicon, "proper lexicon of BM (or BI)". We are planning to get those proper lexicons by mapping the *Kamus Dewan* and the *Kamus Besar Bahasa Indonesia.* We'll extract from this comparison three lists of words: 1) list of words belonging to BM and BI, 2) list of BM words and 3) list of BI words. Our work will make easy if we can use the electronical form of the Indonesian dictionary.

If we want to know exactly the language of one text, the best way for it is to check if most of the words of this text belong to the dictionary of one language. Instead of using the whole dictionary, it's more reliable to use only part of this dictionary meaning the words that belong only to this language. It's in this point of view that we want to create the proper lexicon of BM and BI.

## 6. Conclusion

From what we have said, we can divide the criteria identifications into three groups: criteria at the characters level, criteria at the morphology level and criteria at the lexicon level. Between the three, it's of course the last one that can be really reliable for language identification. But we think that before looking up in the dictionary, the result get from the two other levels may direct the program to open the appropriate lexicon that can reduce the time for identification.

In our article, we raise the problem of the identification of closed languages. We hope that through our research, the differences between Malay and Indonesian will appear clearly and that our language identifier will be an unavoidable tool for any system working on Malay texts.

## REFERENCES

Asmah Haji Omar, *Essays on Malaysian Linguistics*, Dewan Bahasa dan Pustaka, 1975.

Asmah Haji Omar, *The Linguistic Scenery in Malaysia*, Dewan Bahasa dan Pustaka, 1992.

*General guidelines for Malay spelling*, Dewan Bahasa dan Pustaka, 1992.

*General guidelines for the formation of terms in Malay*, Dewan Bahasa dan Pustaka, 1992.

Ismail Bin Dahaman, *Pedoman Ejaan dan Sebutan Bahasa Melayu*, Dewan Bahasa dan Pustaka, 2000.

Ismail Bin Dahaman, *Pedoman Ejaan Rumi Bahasa Melayu*, Dewan Bahasa dan Pustaka, 2000.

*Kamus Besar Bahasa Indonesia*, Departemen Pendidikan dan Kebudayaan, 1990.

*Pedoman Umum Ejaan Bahasa Indonesia Yang Disempurnakan*, Edisi kedua, berdasarkan Keputusan Menteri Pendidikan dan Kebudayan Republik Indonesia, 1987.

Samsuri Ikip Malang, *Tata Kalimat Bahasa Indonesia*, Sastra Hudaya, 1985.

Van Dyck, A.-M., Malherbe, V., *Parlons Indonésien*, L'Harmattan, 1997.

Vinsensius Berlian Vega SN, Bresnan S., "*Indexing the Indonesian Web: Language Identification and Miscellaneous Issues*", http:///www.comp.nus.edu.sg/~vinsensi/pubs/www10-IndexingIndonesianWeb.pdf.