Computational Linguistics at Universiti Sains Malaysia

Choy-Kim Chuah, Zaharin Yusoff

School of Computer Science
Universiti Sains Malaysia
11800 Penang, Malaysia
{kimc, zarin}@cs.usm.my

Abstract

This paper gives a brief history of UTMK, a computer-aided translation unit, and reports on her projects and research co-operations. After its beginnings as a thesis project on Malay affixation, UTMK's interest moved from machine translation to the development of tools for translation. Today, UTMK's focus is on the development of natural language processing applications and tools (internet browsers, and corpus and dictionary databases). And, continuing with its policy for research collaborations, UTMK is leading a three-country project to pool computing and linguistic resources and expertise on Malay. Due to historical reasons, *bahasa Indonesia* and *bahasa Melayu*, the Malay used respectively in Indonesia and in Malaysia have diverged with differences in vocabulary, pronunciation and spelling. For effective communication, a council was set up in 1972 to standardize the spelling and terminology used in the two countries. Brunei joined this council in 1985. To encourage studies on Malay, texts need to be available. However, resources in digital form are wanting. At a recent meeting, the council proposed to set up a Malay language portal to make linguistic resources from the three countries available on-line, and also to popularise Malay as a South-East Asian language. The participation of non-member countries in the portal is welcomed.

1. Introduction

The availability of powerful computers to store information has spurred researchers in natural language processing to set up not just databases of monolingual dictionaries and corpora, but also to pool data from various sources to set up multilingual dictionary databases. SIGLEX, a Special Interest Group of the Association for Computational Linguistics, is trying to link publicly available electronic dictionaries and corpora for studies in natural language processing (see http://www.cis.upenn.edu/~mpalmer/siglex/online.html).

In a seminar given on 3 October, 2001 at the University of Montreal (http://www-rali.iro.umontreal.ca/Seminar/annonce-03-10-2001.html), it was reported that the *Office de la langue française* in Québec is working with China to construct a multilingual, multimedia terminological database that is internet-oriented (*projet* CHIQ-BTML). Meanwhile, on home ground, the Brunei-Indonesia-Malaysia Council for the Malay language, *MAjlis Bahasa Brunei-Indonesia-Malaysia* (MABBIM), met recently (in September 2001 in Bali) to discuss the setting up of a Malay text initiative.

The rest of the document is organised as follows.

In section 2, after a brief introduction to the humble beginnings of *Unit Terjemahan Melalui Komputer* (UTMK) a computer-aided translation unit in Malaysia, and some of her past and present projects, we discuss UTMK's current projects and her participation in MABBIM's proposed Malay Linguistics portal. The text initiative comes under this portal.

We begin section 3 by drawing attention to the differences and similarities between two variants of Malay, *bahasa Indonesia* ['the Indonesian language'] and *bahasa Melayu*[1] ['the Malay language'], and the reason behind the call for the standardisation of these two "languages". Next, we report on the council which was especially set up and entrusted with this task, and the status of her standardisation efforts.

This council has since its inception grown in members, and so has its vision. To conclude, we reiterate some projects in progress that will contribute to the proposed Malay Linguistics portal, and some steps already taken by Malaysia to popularise Malay.

2. Unit Terjemahan Melalui Komputer

2.1. The beginning and GETA

Computational linguistics in Universiti Sains Malaysia (USM) began as a Masters project on Malay affixation (see Chang, 1978). For guidance, USM consulted the machine translation (MT) research group in Grenoble, France, *Groupe d'Étude pour la Traduction Automatique* (GETA). This cooperation marks the beginning of a long friendship between UTMK and GETA which holds strong until today.

In 1984, after the completion of an English-Malay MT system founded on ARIANE's rule-based system, UTMK was accorded the status of a computer-aided translation unit. Presently, UTMK comes under the School of Computer Science at the university.

2.2. Translation systems and dictionaries

2.2.1. MT System JEMAH

From the experience gained from her earlier MT project, UTMK developed her own rule-based MT system shell in 1988, and in the following year completed the first fully locally built English-to-Malay MT system, JEMAH.

JEMAH is currently used as a prototype application for translating English texts into Malay, and was one of the very few systems at that time in the world that run on the personal computer. The system's vocabulary has about 10,000 words.

[1] Sometimes also referred to as *bahasa Malaysia* ['the Malaysian language'].

2.2.2. Human-aided MT system SISKEP

Besides MT systems, UTMK also developed human-aided MT tools. In 1987, a workstation for translators, SI*stem* S*tesen* KE*rja* P*enterjemah* (SISKEP) was completed. SISKEP's built-in features include a Malay spellchecker (RakanBM) ["rakan" = 'friend'], a Malay thesaurus and terminology dictionaries. The latter dictionaries may be loaded or unloaded as required during translation. The data for the terminology dictionaries were supplied by the nation's Language and Literature Agency, which is better known as *Dewan Bahasa dan Pustaka*, or DBP.

While SISKEP is directed at translators in the local context, i.e. for English-Malay, the original conception of SISKEP was directed at multilingual translators. This is to say that the user is assumed to be a multilingual translator, and when translating between a different language pair, e.g. French-Malay, the appropriate dictionary will be downloaded.

2.2.3. *Kamus Perancis-Melayu Dewan* [= 'The Dewan French-Malay Dictionary']

Among the major difficulties in the preparation of any dictionary are the hiring of experienced lexicographers, and the many man-hours required to complete the project. In the Malaysian context, not only were we faced with the difficulty of finding qualified lexicographers, but finding anybody who is conversant in French and Malay, has a good vocabulary of both languages and is willing to work for a trifling sum. Also, the project was to be completed quickly.

To overcome our problems, we cross-referenced the rudimentary French-Malay dictionary inserted in SISKEP for demonstration purposes with an English-Malay dictionary from our translation systems, and obtained the first draft of a French-Malay dictionary. With corrections, revisions, and the addition of entries and example sentences in French and Malay, by the French Embassy in Kuala Lumpur and DBP, our counterparts in the project, the *Kamus Perancis-Melayu Dewan* (Dewan Bahasa dan Pustaka *et al*., 1996) came into existence.

The project which took about eight years to complete from first draft to printed form, has about 19,000 entries. Much of the work, including the preparation of galley proofs, was carried out on-line. The French-Malay dictionary has since its launching in 1996 been put on the web by GETA, our fourth counterpart in the project (see http://www-clips.imag.fr/cgi-bin/geta/fem/fem.pl).

As the information in the English portion of the dictionary which was used to link French to Malay is not deleted, we technically have in our possession the first draft of a trilingual French-English-Malay dictionary. The information in the English portion has not been fully checked.

2.2.4. *Tesaurus Umum Bahasa Melayu* [= 'The Malay General Thesaurus']

From our various projects, we have at our disposal on-line resources on the Malay lexicon. Using programs, we were able to "generate" the first draft of a thesaurus. Followed by more work and corrections by untrained lexicographers, a Malay thesaurus, *Tesaurus Umum Bahasa Melayu* (Tuan Haji Jumaat *et al*., 1990) was quickly "whipped up".

2.3. Linguistic tools and data compiled

2.3.1. Grammar writing formalisms

In the development of translation systems, various needs arose. One of these was the need for grammar writing formalisms.

At GETA, the first work on grammar formalisms was carried out by Chappuy (1983). Zaharin (1986) who was in GETA at that time, followed-up on Chappuy's work. This research was later followed-up by Tang (1994) at UTMK, and now by a Jordanian foreign student.

2.3.2. Corpus database

To support and encourage linguistic research in Malaysia, especially for Malay, corpus and lexical databases are much needed.

As the country's agency for language and literature, DBP in 1994 embarked to set up such a project, and UTMK's help was recruited. This text initiative by DBP on Malay has as at 31 January 2001, an input of about 71 million words from various text genres (see Table 1 below).

Table 1. DBP's Corpus Database[+]

SUB-CORPUS	PERCENTAGE	WORDS
Newspapers	42.2%	30,138,590
Books	33.0%	23,571,255
Magazines	14.3%	10,237,176
Cataloguing cards	4.4%	3,130,641
Traditional texts	3.2%	2,293,767
Translated books	2.4%	1,709,783
Books on drama	0.3%	191,544
Pamphlets, brochures, etc.	0.2%	169,222
Poems	0%	2,348
Total	100.0%	71,444,326

[+]taken from http://dbp.gov.my/2000/pelawat/bahasa/webdewan2.html; last row inserted by us.

The database for this initiative caters for the storage and retrieval of large texts which may be searched for by string, keyword, or part-of-speech. Besides concordances lists, the database is also augmented with MATA, a MAlay Text Analysis system, which can carry out simple text analyses, and provide statistics on word frequency, sentence length, etc. on selected texts. It even does a simple root extraction of words.

This Malaysian Malay corpus database system is at present only accessible from within DBP (see http://dbp.gov.my/2000/pelawat/bahasa/webdewan2.html) .

2.3.3. Dictionary database

Also in DBP's plan is the computerization of her dictionaries. UTMK's help was once again recruited.

For effective storage and retrieval, the information for each dictionary entry is stored in fields/attributes, e.g. entry, definition, pronunciation, number, gloss, equivalent, etc.. Hence, a search may then be effected on a keyword in a field, or on the field itself. In this way, dictionaries targeted for different groups of users may be extracted from the same dictionary database.

As each dictionary type is kept in a separate database with its own structure, the dictionary structure may be modified as necessary to accommodate new attributes. The information that is to go into a dictionary is extracted automatically using programs. And, via a copy-and-paste into PageMaker files, galley proofs may be produced without manual typesetting. The French-English-Malay dictionary mentioned earlier (see section 2.2.3) was prepared using this methodology.

2.4. Current projects

With increase in communication and business transactions conducted over the net, UTMK's research interest while still on natural language processing is now on the development of tools and applications for such a purpose. Her current projects include the development of Malay internet browsers and generic portals for business over the internet, besides an interest in multilingual dictionaries.

UTMK's research interests also extend beyond the nation's boundaries. Her experience in building databases, and preparing dictionaries and thesaurus, especially in DBP's projects (construction of a corpus database, and computerisation of her dictionaries) makes her an active participant in the planning and setting up of the MABBIM Malay linguistics portal for the region.

For more information on UTMK's research projects, see http://utml-ultra.cs.usm.my/content/rddirection.htm.

3. Malay and the Malay Language Council

3.1. *Bahasa Indonesia* and *Bahasa Melayu*

3.1.1. Two "languages"?

Bahasa Indonesia and *bahasa Melayu* are the respective lingua francas of Indonesia and Malaysia. These two "languages" share common roots in Malay, a language from the Austronesian family. However, because of differing historical backgrounds, these two languages have over time diverged with differences in vocabulary, pronunciation and spelling[2].

For some words, the divergence in pronunciation and/or spelling exhibited is slight, e.g. *uang ~ wang* [= 'money'], *jadwal ~ jadual* [= 'schedule'], *coba ~ cuba* [= 'try'], *resmi ~ rasmi* [= 'official, formal'] and *pikir ~ fikir* [= 'think']. Evidence of their common origin is still obvious (for other examples see Table 2a).

In the following, the divergence is of a different kind. Consider *sepeda ~ basikal* [= 'bicycle'] and *kantor ~ pejabat* [= 'office']. Consider the first example. While the *bahasa Indonesia* and *bahasa Melayu* words differ greatly, the pronunciation of *basikal* closely resembles that of its English equivalent "bicycle". In the second example, *kantor* closely resembles the Dutch word of "kantoor" for "office". The influence of the colonial language over the local *lingua franca* is strong (for other examples see Table 2a).

Table 2a. *Bahasa Indonesia* and *bahasa Melayu*: A comparison of vocabulary

ENGLISH	B. INDONESIA	B. MELAYU
accountant	akuntan	akauntan
activity	aktivitas	aktiviti
bankrupt	bangkrut	bangkrap
because	karena	kerana
case	kasus	kes
comment	komentar	ulasan, komen
English	bahasa Inggris	bahasa Inggeris
management	manajeman	pengurusan
test	tes	ujian
university	universitas	universiti
volcano	gunungberapi	gunung berapi

Other words suffer yet greater divergences with changes in meaning. While "bisa" is ambiguous in *bahasa Indonesia*, "bisa" in *bahasa Melayu* more often than not means 'poison' (see Table 2b below). "Bisa" as a verb does not exist in *bahasa Melayu*.

Divergence can also be in meaning and in spelling. Consider the *bahasa Indonesia*/*Melayu* word for 'official; formal'. The two words differ only in spelling: "resmi" in *bahasa Indonesia* as opposed to "rasmi" in *bahasa Melayu*. While the word "resmi" does exist in *bahasa Melayu*, it is a different word with the meaning of "trait".

For more examples on similar words with different meanings, see Djajasudarma (1996).

Table 2b. *Bahasa Indonesia* and *bahasa Melayu*: A comparison of vocabulary

ENGLISH	B. INDONESIA	B. MELAYU
poison (n.)	bisa (n.)	bisa (n.)
to be able to, can (v.)	bisa (v.)	dapat, boleh (v.)
official, formal (adj.)	resmi (adj.)	rasmi (adj.)
trait (n.)	-	resmi (n.)

3.1.2. Nationalism and one Malay language

With the spirit of nationalism rife in South-East Asia during the post-war period, it is not surprising that the idea of one Malay language was reported to have been proposed as early as the 1950's (http://www.dbp.gov.my/dbp98/mabbim/mabbim.htm), probably in anticipation of the role that Malay would play in the region after independence.

While we have no access to the list of most widely spoken languages in the world for that period, Malay is currently reported to be in 9th place with 148 million

[2] We have excluded from our discussion the variant of Malay used in Brunei.

speakers[3] (see http://www.englishconsulting.com/ engcons8.html). We note that depending on the source, the ranking given may be different. At the DBP website, Malay was reported to be the 5th most widely spoken language in the world with the number of Malay speakers exceeding "200 million"[4] (http://dbp.gov.my/mabm/ mabm.htm).

Today, Malay is the national language of Indonesia, Malaysia and Brunei, and the official language of at least one nation, Singapore.

3.2. The Malay Language Council

3.2.1. The beginning and guidelines

Despite the divergence, *bahasa Indonesia* and *bahasa Melayu* still share many similarities, and are mutually comprehensible to some extent. However, to ensure effective communication at all levels and in all relations, e.g. trade, education, politics, etc., the spelling and terminology used in the two nations need to be standardised.

In 1972, the *Majlis Bahasa Indonesia-Malaysia* (MBIM) was established and entrusted with this responsibility.

In 1975, MBIM finalised the guidelines[5] for the standardisation of pronunciation and spelling. The committee also drew up rules for the transliteration of loan words and the coining of terms[6], which are important for the effective communication of technical knowledge between the countries, a strong motivation behind the proposed Malay Linguistics portal for the region.

We note that the rules for the transliteration of loan words are slightly different in each country. For example, while the English suffix "-ity" is changed into "-iti" in *bahasa Malaysia*, in *bahasa Indonesia*, it is changed to "-itas" (see the *bahasa Indonesia*/*Melayu* words for "activity" and "university" in Table 2a). Colonial influence is clearly deeply-rooted.

Linguistic divergence does not exist just between nations, but also within the country. The pronunciation of speakers of Malay from the east coast of peninsula Malaysia differs quite distinctly from that in the west, and that in the north with that in the south.

In Malaysia, the plan for standardisation was also aimed at standardising the pronunciation of Malays from different regions who in fact speak the same tongue, and at uniting the immigrant population with these local people via a standard national language.

[3] No indication of the year of census was given.

[4] For other differing figures, see http://www.ignatius.edu/Turner/languages.htm. The 1995 Indonesian census gave her population to be about 195 million (http://www.ids.org.my/stats/KeyData/bimpeaga.htm#BIMP-EAGA:).

[5] *Pedoman Umum Pembentukan Ejaan Bahasa Malaysia*.

[6] *Pedoman Umum Pembentukan Istilah Bahasa Malaysia*.

3.2.2. A three-member council

Since its inception, the council has grown. In November 1985, Brunei became a member, and MBIM became MABBIM, a three-member council.

Singapore too joined, but as an observer (see http://dbp.gov.my/ mab2000/mabbim.htm). In the most recent meeting held on 13 March 2002, the Philippines, Thailand, Laos, Vietnam, Cambodia and Myanmar, which do not speak Malay, too attended.

3.2.3. Standardisation efforts

In sub-subsection 3.1.1, we noted the influence of the colonial language over the local lingua franca. Because of this influence and other deeply-rooted reasons, standardisation in the true sense of the word is not possible. A decision taken by the committee may be one of the following two: (a) "agree to be the same", or (b) "agree to differ". In the latter, the committee may choose: (i) to accept a proposed term, but insist that the spelling be different (see "rural communication" in Table 2c), or (ii) to differ entirely in the choice of term (see "consultant" and "resistance" in Table 2c).

Table 2c. *Bahasa Indonesia* and *bahasa Melayu*: A comparison of vocabulary

ENGLISH	B. INDONESIA	B. MELAYU
life span	jangka hayat	jangka hayat
assignment editor (tv/film)	penyunting tugasan	penyunting tugas
excited state	keadaan tereral	keadaan teruja
rural communication	komunikasi pedesaan	komunikasi kedesaan
architecture	arksitektur	seni bina
coin	recehan	syiling
consultant	konsultan	perunding
resistance	tahanan	rintangan
skew	miring	serong

After the terms are finalised, each country prints her own terminology dictionaries. Sixteen terminological dictionaries have been printed by DBP.

By its 37th meeting in 1998, the MABBIM had standardised over 100,000 terms from about 50 fields covering more than 250 sub-fields (http://dbp.gov.my/ dbp98/mabbim/ mabbim2.htm). At this meeting, it was proposed that the results of her meetings be made available to its members and guest members via the net.

As decisions made previously may be reviewed, it is difficult to obtain absolute figures on the degree of unanimity in standardisation. However, from the results of three MABBIM meetings (viz. Nov. 1978, Sep. 1979 and Mar. 1980), we found that where a definite decision has been made, the Indonesians and the Malaysians agreed to differ on 42% (3811) of its decisions. Of these, 5% (445) of the terms differ in spelling, and 37% (3366) differ in the term used.

We note that despite the standardisation efforts of the Malay council, the sister languages still have their differences. In the interest of the languages themselves, comparative studies are necessary to determine to the

extent to which standardisation has been achieved, and the effects of colonial and other influence on the local lingua francas.

4. MABBIM, Malay and the Future

In September 2001, the tri-lateral council met to set up a Malay text initiative. The aim of the initiative is to link via portals, the linguistic expertise and resources (both text and lexicon) available at various language centres, universities, etc. of member countries, or of any non-member country who are interested to participate.

This Malay linguistics portal which will house both oral and written corpora, will be an impetus to the much needed research in comparative Malay studies, and Malay linguistics in general. Details with regards to the type of corpus, dictionary, etc. to be stored, and the assignment of task, and manpower are being worked out by a working committee.

It is in MABBIM's plan to make Malay an official language of South-East Asia by 2005 (see Appendix I of the 37th MABBIM Meeting held in 1998 in Malaysia, which is available in http://dbp.gov.my/mab2000/mabbim.htm).

In support of the internationalisation of Malay, the Malaysian government has as early as July 1997 approved the setting up of the International Council of Malay Language (ICML), or *Majlis Antarabangsa Bahasa Melayu* (MABM) in Malay (http://dbp.gov.my/mabm/mabm.htm). The members for this council which total about three dozens include non-Asian countries like Egypt, Germany, Sweden and Russia.

To popularise Malay, resource centers have also been set up at some institutions of higher learning, such as Ohio University and Leiden University. In 1997, a Malay language center was set up at the Beijing Foreign Studies University.

To contribute to the portal, DBP has a ready corpus database of about 70 million words, and is in the process of getting her dictionaries on digital format. Meanwhile, UTMK with her experience is leading the working committee in the setting up of the Malay portal.

5. Acknowledgements

We would like to thank Rusli from DBP for the last minutes of the MABBIM meeting on the text initiative.

6. References

Chang, M.S. (1978) *Computer System Aids in Natural Language Processing with Applications in Bahasa Malaysia*. M.Sc. thesis, Universiti Sains Malaysia, Penang, Malaysia.

Chappuy, S. (1983) *Formalisation de la description des niveaux d'interpretation des langues naturelles. Etudes menées en vue de l'analyse et la generation au moyen des transducteurs*. Thèse 3ème cycle. Juillet 1983. INPG, Grenoble.

Dewan Bahasa dan Pustaka *et al*. (1996) *Kamus Perancis-Melayu Dewan*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Djajasudarma, T. Fatimah (1996) Kosakata Indonesia—Malaysia: Serupa tapi tak Sama. *Proceedings of the 35th MABBIM Meeting*. 18-22 1996. Bukit Tinggi, West Sumatra. (see http://dbp.gov.my/2000/pengenalan/frame_utama.htm)

Tang, E.K. (1994) *Natural language analysis in machine translation* (*MT*) *based on the string-tree correspondence grammar* (*STCG*). Ph.D. thesis., Universiti Sains Malaysia, Penang, Malaysia.

Tuan Haji Jumaat *et al*. (1990) *Tesaurus Umum Bahasa Melayu*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Zaharin, Y. (1986) *Strategies and heuristics in the analysis of a natural language in machine translation*. Ph.D. thesis, Universiti Sains Malaysia, Penang, Malaysia.