Design and Construction of a Machine-Tractable Malay-English Lexicon

Chiew Kin Quah*[1], Francis Bond† and Takefumi Yamazaki◆

*R & D Department
NTT MSC Sdn Bhd
No. 43000, Jalan APEC,
63000 Cyberjaya,
Selangor Darul Ehsan,
Malaysia
annieq@nttmsc.com.my

†NTT Communication
Science Labs
NTT Corporation
2-4, Hikaridai, Seika-cho,
Soraku-gun, Kyoto
619-0237, Japan
bond@cslab.kecl.ntt.co.jp

◆NTT Cyber Space Labs
NTT Corporation
1-1 Hikarinooka
Yokosuka-shi, Kanagawa
239-0847, Japan
yamazaki@nttnly.isl.ntt.co.jp

Abstract

In this paper, we introduce a machine-tractable Malay-English lexicon currently being developed at NTT MSC, Malaysia. The lexicon is designed to satisfy three criteria: Firstly, to develop and document detailed syntactic features useful for both analysis and generation. Secondly, to use a well-developed semantic ontology, in our case the semantic classes from the 2,710 classes used in the machine-tractable Japanese-English Goi-Taikei ontology. Thirdly, to get as wide cover as possible. The lexicon currently contains around 91,000 Malay-English pairs. Each entry consists of nine major fields, which include such information as numeral classifiers associated with common nouns, and meta-codes to show honorific use, register and origin. In addition, English and Chinese translations and comments are provided for future use in machine translation systems and also as an aid for non-Malay speakers. A version of the dictionary, which does not show all fields, is available on-line.

1. Introduction

In this paper, we introduce KAMI (KAmus Melayu-Inggeris) --- the machine-tractable Malay-English lexicon being developed by the Multilingual Translation Research Project Team of NTT MSC Sdn Bhd. There are several hard-copy Malay-English paper dictionaries, the best being the *Kamus Perwira* (Kelana & Lai Choy, 1998) containing 40,000 basic words with the addition of several tens of thousands of derivatives. There are also some machine-tractable lexical resources. There are two resources available for research purposes: the European Corpus Initiative Multilingual Corpus I (ECI/MCI), which includes a small Malay corpus and the CICC Malaysian Lexicon developed by the Center of the International Cooperation for Computerization (CICC, 1995).

There are three Malay lexicons that can be accessed on-line: Dr Bhanot's Malay-English Cyber-Dictionary (http://www.malaysia.net/cybercom/dictionary/); the French-English-Malay (FeM) Dictionary, which supplies English and Malay translations of the French entries (http://www-clips.imag.fr/geta/services/dicoweb/dicoweb_en.html); and the Malay-English-Finnish (MEF) Cyber-

[1]Also a full time lecturer at the Department of Malay Language and Translation, Faculty of Language Studies, National University of Malaysia.

Dictionary (http://koti.mbnet.fi/~amika/dictionary/malay.htm). However, the MEF only has 358 words. All three lexicons take advantage of their electronic form to allow searches from either English or Malay (or French or Finnish).

Dictionaries can be judged on three criteria: quantity of information, quality of information, and effectiveness of presentation (Landau, 1989: 306). The CICC lexicon has good coverage, with around 50,000 Malay entries, and quite detailed syntactic and semantic information set out in the documentation. Unfortunately, the quality is poor. In many entries, the Malay index terms and English translations are misspelled and/or the syntactic and semantic tags are mistyped. During the project, the syntactic/semantic tags were revised, and tags from both the old and new sets are mixed. Because of these errors, the resource is of limited use. On the other hand, Bhanot's cyber lexicon is much smaller, containing 10,000 translation pairs and no syntactic information. The overall quality is high. The FEM lexicon is based on a French-English Lexicon, with Malay added, as described in (Lafourcade, 1997). It is similar in style to a paper dictionary, with entries having an index word, part-of-speech and examples of use in French, English and Malay.

In the next section, we will introduce KAMI, our machine-tractable Malay-English lexicon. Our aim is to build a lexicon primarily for use in NLP applications, in particular, machine translation. Ideally we would like to use the same dictionary for both the analysis and generation of Malay. The lexicon is also being used to produce a Japanese-Malay lexicon, by using the English and Chinese entries to link to a Japanese-English lexicon (Bond *et al*., 2001).

To further these aims, we use a well-developed semantic ontology: the 3,000 semantic classes used by the machine translation systems ALT-J/E and ALT-J/M (Ogura *et al.*, 1999). In addition to depth of information, we consider coverage to be very important, although at present we have only around 91,000 Malay-English pairs.

A version of the dictionary, which does not show all fields, is currently accessible on-line at <sangenjaya.arc.net.my>.

After describing the structure of each entry, we then discuss some of the issues that arose during the construction. Finally, we give some of our future plans for the lexicon.

2. Design of Malay-English Lexicon Entries

In this section we describe the logical structure of each entry of the lexicon. Each entry has nine fields, as shown (with examples) in Figure 1. More detail is given about each feature in the following subsections. All the features are documented in HTML, so that the documentation is portable and on-line. Each feature has a unique code, so that they can be searched for easily. Particularly detailed information is provided for the syntactic tags. There are only two required fields: the Malay index word and its part-of-speech.

The total number of Malay-English entries now stands at 91,426, with 67,658 Malay index words. 79% of the Malay words have only one translation, 14% have two English translations and 4.1% have three translations The average number of translations is 1.35.

2.1 Malay Index Word

This is the headword, used to look up the word during analysis, or generate it during generation. If there are spelling variants, they will be listed here, with the preferred form given first (see Section 3 for further discussion). Because this is a bilingual lexicon, multi-word expressions are allowed as index words. Approximately 36,000 entries are multiword: e.g. *tidak boleh dibahagi* "indivisible"**.**

No	Explanation	Example(s)	Comment
(1)	Malay Index Word	*gajah*	required
(2)	Malay Root Word		only if index is derived form
(3)	Part-of-Speech	Noun	required
(4)	Syntactic Features	CL=*ekor*	list of features
(5)	Semantic Features	536:beast	list of features
(6)	English Translation	*elephant*	list of translation equivalents
(7)	Eng. Definition	*a kind of animal*	description of English translation
(8)	Chinese Translation	*zou*	EUC encoded characters
(9)	Meta-Tags		list of relevant meta-tags

Figure 1. **List of fields in KAMI**

2.2 Malay Root Word

If the index word is a derived form, then the root is entered here. In traditional Malay lexicons, this is used as the headword (see Section 3 for further discussion).

2.3 Malay Part-of-Speech

This is the part-of-speech of the Malay index word. We have a hierarchy of parts-of-speech. There is an initial split into 5 groups: nouns, verbs, adjectives, adverbs and function words. Within each word class, further details are given, for example the noun word class is further divided into five distinct types: common nouns, proper nouns, numeral classifiers, pronouns and titles.

Parts of speech are labeled with a three-letter code. All the noun codes begin with N to indicate that the word/phrase in field one is from the class of noun. For ease of processing, codes (the second and third letters) from each of the five distinct types begin with different letters. The following two letters in each code will indicate what type of noun index word is, e.g., NC refers to common noun or NP refers to proper noun. Hence the part-of-speech codes are NNC for common nouns and NNP for proper nouns. Altogether we distinguish 38 different parts of speech.

2.4 Syntactic Features

More detailed syntactic information is given in this field, either as single codes, or codes with values.

Some examples of codes are given here. NHR is used to show a multi-word noun with the head on the right (the default in Malay is for the head to be on the left). CL is used to mark the default classifier associated with each noun. This code must be accompanied by a value, which is the classifier, given as a string. Malay is a numeral classifier language, and nouns normally cannot be directly modified by numerals. Instead a classifier is used, similar to *piece* in *two pieces of paper*. The classifier is useful for generation in NLP applications, and also for second language learners. Classifiers are listed for 18,900 Malay nouns (22%). We are trying to put in the minimum amount of syntactic information needed to correctly generate Malay.

2.5 Semantic Classes

Semantic information is stored in this field. The most common information is semantic classes from the Goi-Taikei ontology (Ikehara *et al.*, 1997). This is a hierarchical ontology of some 3,000 semantic classes, organized with is-a and has-a relations. The classes can be used to disambiguate words with multiple senses. For example *perang* has two distinctive meanings: “brown” (as in colour) and “war”.

Separate entries will be created for these senses, the first marked with [2352:color] and the second with [1755:war].

In some cases we wished to distinguish semantic classes not given in the Goi-Taikei ontology. For example, [bladed weapon] is a natural class in Malay, and has its own classifier *bilah*. In such cases we have added semantic classes to the ontology (so far we have only added [bladed-weapon] and [substance]).

2.6 English Translation

This field lists one or more translation equivalents for the Malay index word or phrase. If there is more than one translation, and using different semantic classes cannot disambiguate the translations, then they will all be listed, with the translation judged to be most common given first. For example, *gerombolan* can be translated as 'band', 'gang' or 'group'. Because *group* is the most general translation, we list it first. Of course, multiple Malay words may have the same English translation: e.g. *pencernaan* and *penghadaman*, both have 'digestion' as their English translation.

2.7 English Definition

When we can't immediately determine a good translation equivalent, we add a definition in this field. For example, many plant and animal names have no exact equivalent. For these, we leave the translation field blank, and translate as the Malay word in single quotes, followed by the definition in brackets: for example, *mempinas* (a kind of fish), *mempitas* (a kind of tree).

2.8 Chinese Translation

We also include Chinese translation equivalents, which we hope will be useful in future Malay/Chinese machine translation research, as well as for the Chinese speakers in Malaysia (25.45% of the 22.2 million population) (www.ids.org.my/idsonline/KeyData/population.htm).

2.9 Meta-Tags

Meta tags include usage tags such as **archaic, vulgar, honorific, taboo**; etymologic tags showing the origin of loan words (marked using the ISO language code of the language of origin) and dialect tags for words used mainly in Indonesian, Javanese and Buginese.

2.10 Effectiveness of Presentation (On-line Lookup)

In order to present the lexical information effectively, we have created a web-based look up tool that allows the user to look up any field of the lexicon: the Malay index word, the English and Chinese translations, or even the syntactic and semantic features. It is thus possible, for example, to list all words counted with a given classifier, or all members of a given semantic class. This makes the lexicon a potential monolingual research tool, as well as a multi-lingual lexicon.

3. Issues in the Construction of the Malay-English Lexicon

The base of our Malay-English lexicon was a dictionary produced by a Malaysian translation company. This included English, Malay and Chinese, with some part-of-speech information and numeral classifiers. We reformatted the dictionary so that each entry had the nine fields described above, plus an ID field and a field for comments by the lexicographer. The entire dictionary is kept as a single text file. It is around 8 Mbytes, which is small enough to edit using the text editor Emacs, which we use in

both Windows and Unix environments. During construction, we used a variety of one-off perl scripts to add, change and reformat information automatically.

3.1 Orthographic Issues and Standardization

During the construction of the Malay index words, we found many spelling variations. For example, *gemala* and *komala* are spelling variations of a single word, with the meaning "magic stone". Rather than create multiple entries with all information identical except the index, we allowed a single record to have multiple index forms, with the preferred form, i.e. (standard) Malay, listed first. The standard Malay used in this construction is based on the more accepted dialect of Malay (see Quah, 1997), which originated in the Riau islands (now part of Indonesia).

Another related issue is that some words we found in on-line resources are not found in the various dictionaries used as reference. There can be two reasons for this: the first is the word does not exist in the standard Malay vocabulary but is somehow "borrowed" from another language and the second is the word does exist but is rather recent and has not been incorporated into the printed dictionaries. The former is common, where many words found during checking are of Indonesian, Javanese or Buginese origins that have not been accepted as standard Malay but are used by Malay language speakers in Malaysia. The latter is more problematic as there is no way of knowing if the words have been standardized with out looking for them in a large up-to-date Malay corpus.

3.2 Morphological Issues

Another issue that arose while compiling, checking and editing the lexicon was how many derivations need to be listed. It is inefficient to list semantically transparent regular derivational forms, although they may be needed in a bilingual dictionary if their translations are irregular. Malay is an agglutinative language, with many derivations arising through affixation (Abdullah Hassan, 1974; Heah, 1989; Asmah Hj Omar, 1988). Affixation is highly productive, for example, the verb *guna* [use] in its root form generates many derivations via affixation:

a) Verbs: *mengguna* "to use", *diguna* (passive form of *mengguna*), *menggunakan* "to use for certain purpose", *digunakan* (passive form of *menggunakan*), *mempergunakan* "to use something to gain benefits", *dipergunakan* (passive form of *mempergunakan*), *menggunai*/*menguna-gunai* "to charm, to use love potion on someone"
b) Nouns: *pengguna* "consumer, user", *penggunaan* "usage, utilization, consumption", *pergunaan* "the use of something", *gunaan* "applied as in applied science", *kepengunaan* "consumerism", *kegunaan* "use, benefit", *guna-guna* "charm, love potion"
c) Adjectives: *gunawan, seguna* "to have the attributes of kindness"

To add to the complexity, the affixation process in Malay allows a maximum of three layers for any root word, for example *berkeseorangan* "to suffer from loneliness" has undergone three layers of affixation. The first layer is the root word *orang* "person" prefixed by *se-* (*seorang* "alone"); the second layer is the circumfixation of *ke-an* (*seorang* is 'sandwiched' between *ke-* and *–an* to create *keseorangan* "loneliness"). In the third layer the derived word from the second layer *keseorangan* is prefixed by *ber-* to become *berkeseorangan* "to suffer from loneliness" (Quah, 1998).

As space is not so much of an issue for an electronic lexicon, we decided to err on the side of excess, and add derived words if they potentially made the translation process easier. Any words that can be generated by rule-based processes can always be deleted in the same way.

3.3 Pronunciation

Sometimes, ambiguous words are distinguished by their pronunciation, such as *semak*: one /*semak/* is pronounced with a schwa (e-pepet, as in *saber*) and the other /*semak/* is pronounced with an e-taling (as in *error*). The meanings are "bush" and "check/inspect" respectively. Similarly, *kelah*: one */kelah/* is pronounced with an e-taling to mean "picnic" while two other */kelah/* have e-pepets; one means a kind of freshwater fish and the other means "to make a complaint or accusation". In most cases, the distinction by pronunciation involves words with e-pepet or e-taling because these spellings are not marked in the romanized spellings (although they are in the Jawi script derived from Arabic). In our lexicon, only the semantic classes indicate the correct pronunciation of such words.

4. Future Work

We would like to extend the lexicon further. Firstly, we need to complete the existing entries: in particular some entries are missing their semantic classes and Chinese translations. Secondly, we want to include more morphological and frequency information. Thirdly we will add more entries to the existing lexicon. Finally, we would like to make a fuller version available on-line.

References

Abdullah Hassan. 1974. *The morphology of Malay*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Asmah Hj Omar. 1988. *Susur galur bahasa Melayu*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Bond, Francis, Takefumi Yamazaki, Ruhaida binti Sulong, and Kentaro Ogura,. 2001. Design and construction of a machine-tractable Japanese-Malay lexicon. In *Seventh Annual Meeting of the Association for Natural Language Processing (NLP-2001)* pp 62-65, Tokyo..

CICC, 1995. Malaysian Basic Dictionary. Tech Report 6 –CICC—MT54, Center of the International Cooperation for Computerization, Tokyo.

Cowie, J, E. Ludovik and R. Zacharski. 1998. An autonomous, web-based, multilingual corpus collection tool. In the *Proceedings of the International Conference on Natural Language Processing and Industrial Applications*. At http://cls.nmsu.edu/~raz/langrec/nlpia.htm.

Heah, Carmel Lee Hsia. 1989. *The influence of English on the lexical expansion of Bahasa Malaysia.* Kuala Lumpur: Dewan Bahasa dan Pustaka.

Ikehara, Satoru, Mahahiro Miyazaki, Satoshi Shirai, Akio Yokoo, Hiromi Nakaiwa, Kentaro Ogura, Yoshifumi Ooyama & Yoshihiko Hayashi. 1997. *Goi-Taikei: A Japanese Lexicon*. Tokyo: Iwanami Shoten. 5 volumes/CDROM.

Kamus Dewan. 1993. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Kelana, C. M. and Lai Choy. (compilers). 1998. *Kamus Perwira: Melayu–Inggeris*. Selangor: Penerbitan Daya.

Lafourcade, Mathieu. 1997. Multilingual Dictionary Construction and Services - Case Study with the Fe* Projects. *Proc. PACLING'97*, September 2-5 1997, Meisei University, Ohme, Tokyo, Japan, vol. 1/1, pp 173-181.

Landau, Sidney. 1989. *Dictionaries, The Art and Craft of Lexicography*. New York: Cambridge University Press.

Ogura, Kentaro, Francis Bond, and Yoshifumi Ooyama. 1999. ALT-J/M: A prototype Japanese-to-Malay Translation System. In *Machine Translation Summit VII*, Singapore. 444-448.

Quah, Chiew Kin. 1997. *The translation of English academic texts into Malay with special reference to the translation of English affixes*. Unpublished PhD dissertation. University of Surrey, Guildford, England.

_____. 1998. *Translating English affixes into Malay*. Bangi, Malaysia: Fakulti Pengajian Bahasa, Universiti Kebangsaan Malaysia.