CORPUS-BASED TOOLS FOR THE PROCESSING OF MALAY TEXTS

ZURAIDAH MOHD DON
University of Malaya,
GERRY KNOWLES,
Lingenium Sdn Bhd
Malaysia

The translator has to draw on a wide range of linguistic and real world knowledge. Consider a real translation event. A man enters a building in Kuala Lumpur, and the security guard says to him "*Sudah makan*?" The man knows that *sudah* refers to past events and that *makan* means 'eat', so this speech event has something to do with eating in the past. He also knows that *sudah makan* is a complete predicate, and that recoverable subjects can be omitted; and although there is no interrogative syntax, the final rise in pitch suggests that the utterance is a question. This leads to a provisional translation *Have you eaten?* It is polite in this culture to avoid addressing people directly using translation equivalents of English *you*; and so the security guard is being polite. This is confirmed by his friendly body language.

In English, *Have you eaten?* is an invitation, and implies some responsibility for ensuring that the addressee has something to eat. But a man guarding a building is no such position. Many languages have ritual questions equivalent to *How are you?* which are not really questions at all, and speaker and hearer both know that *Sudah makan?* is not a real question. Taking into account the context in which it is uttered, a situation in which the man is coming back after lunch, one of many possible translations is *Welcome back!* To finish the story, the man smiles and enters the building.

Translators have for hundreds of years, and possibly thousands of years, taken for granted the different linguistic levels that have only been included in linguistic theory in recent decades. Syntactic trees and transformations, theme and rheme, semantic relations, speech acts, politeness, formality, cohesion and pragmatic inferences are all part of the stock-in-trade of the competent translator. And in addition, translators have to be aware of cultural differences and sensitivities, and match the cultural effects of the translation to the original. Translations can even

be witty, as in the case of the British commander who conquered Sindh in 1843 and famously sent the message *peccavi* 'I have sinned'.

Translators are expected not only to have an expert knowledge of the source language and its associated culture, but also to be highly accomplished performers in the target language, with the ability to compose a text crafted to a high standard and virtually indistinguishable from an original text. Acquiring the knowledge and developing the skills and expertise required for effective translation can consequently take many years. It is probably one of those high level skills, like the skills of the concert pianist, said to take about 10,000 hours to develop.

There are no shortcuts, but technologies can assist translators not only in their everyday work but also in developing their skills. Conventional dictionaries, evolved from the glosses used in early medieval Europe to help monks with Latin texts, are an obvious example. Modern technologies can go much further in modelling the linguistic knowledge required by the translator and making it available in a usable form. The work reported here is concerned with Malay, the national language of Malaysia. When work began, there were few resources available apart from a range of dictionaries produced by the national language agency, Dewan Bahasa dan Pustaka. The aim was to create a computer readable resource which modelled the native speaker's knowledge of the Malay language. The result is known as MALEX ("MALay LEXicon"). MALEX is essentially a theoretical model of the language, but it has the potential to be developed into tools for the translator.

MALEX consists of a set of related tables and procedures to relate sets of data in different tables to each other. The data is extracted from a corpus of about 3M words of naturally produced Malay text, some of it spoken but most of it written. The first time a new word is encountered in a text it is intercepted and analysed. First any punctuation is stripped off. If it is a new lexical item, a stemmer strips off affixes until the stem matches an existing entry in the table of lexical items. The new item is added to the table with its morphological structure and grammatical class. If no match is found, the new item counts as a new lemma (roughly equivalent to a dictionary headword), and is stored in the lemmas table. A spelling-to-phoneme algorithm is used to generate pronunciation entries, which are stored in a separate table.

These tables include the kind of basic information that anyone who claims to know the language can be assumed to be familiar with. Much of the information is to be found in conventional dictionaries, but a database makes it possible to manipulate the data in ways which are impossible with printed documents. For example, the ability to sort words according to their morphological structure leads to the identification of regular patterns in the morphology and the identification of irregularities. For example, the words *ubat*, *ubatan* and *perubatan* all have 'medicine' as their English translation equivalent. *Ubat* is the simplex form, i.e. the form with no affixes, and refers to the substance taken by the patient. *Ubatan* with the ending *–an* has a more abstract meaning derived from *ubat*, and corresponds to 'medication'. *Perubatan* with the circumfix *per..an* is more

general still, and refers to medicine as a branch of knowledge or academic subject, parallel for example to *pergigian* 'dentistry' derived from *gigi* 'tooth', or *pertanian* 'agriculture' derived from *tani* 'farm'.

Corpora are still relatively new in linguistics, but translators have always used them. The traditional study of the Classics introduced students to the corpus of Latin and Greek texts, together with the skills required to translate them. To translate novels or newspaper articles, it is essential to be familiar with these genres in both source and target languages. Nowadays it is a relatively simple task to compile parallel corpora. Accessing the contents of these corpora is no longer a task for a human being, and can be carried out more quickly and more efficiently by a machine.

The first task is grammatical tagging, which involves associating each word in the text with its grammatical class or 'part of speech'. These tags are used by a parser to ascertain the syntactic structure of phrases and sentences. Grammatical information is of course essential in order to translate. For example, in order to translate *a round shape* and *a round of toast* into another language, we need to know that *round* is an adjective in the first phrase and a noun in the second. This is where Malay gets interesting. The case of *round* is slightly unusual in English, but much more common in Malay, where words tend to slip from one grammatical class to another. For example, there is a common road sign in Malaysia *kurangkan laju* 'reduce your speed', where *kurangkan* (here) means 'reduce' and *laju* 'fast' is normally an adjective. There is a perfectly good deadjectival noun *kelajuan* 'speed', but in the case of the road sign, the adjective slips into a syntactic slot normally reserved for nouns.

Although Malay grammar books provide word formation rules, the fact is that in Malay texts the word formation rules are by no means always carried out, and so may be regarded as optional. Verb forms can be used with no morphological modification as noun modifiers. For example, *goreng* 'fry' is a verb, but it can be used without affixation to modify *nasi* 'rice' in the phrase *nasi goreng* 'fried rice'. This grammatical fluidity has consequences for the processing of texts. To be sure, a large proportion of a Malay text can be parsed using grammatical class information and the properties of individual words, but there remains a large residue which requires access to semantic information and knowledge of the real world.

For the translator this means that for even a simple sentence, it may be necessary to go beyond grammar into meaning, and start afresh with the grammatical encoding of the meaning into the target language. This applies in the case of grammatical classes that have no equivalent in languages like English. For example, Malay has words like *rosak* 'damaged, spoilt' and *pecah* 'broken', which correspond roughly to past participles. Whereas in English, *break* is essentially an action performed by an agent, and *broken* a derived form, in Malay *pecah* is the basic form, and the meaning 'break' derived, using the verbal prefix *meng-* and the agentive suffix *–kan* to form *memecahkan*. In order to handle words like *rosak* and *pecah*, we need to classify words at a semantic

level rather than conventional grammatical class. People who know English and Malay – and especially people who can translate from one language to the other – know perfectly well at an intuitive level how to deal with such words. But from a theoretical point of view it remains an unsolved problem, and work currently underway to extend the MALEX database to include semantic word classes is far from completed.

Perhaps the aspect of the translator's work that takes longest to mature is developing a "feel" for the language and its words. It is also in this area that corpus-based methods have much to offer. It is possible to compare word frequencies in different corpora in order to ascertain which words are used more frequently (or perhaps less frequently) than usual in a particular corpus ("key words"). It is also possible to identify collocations, groups of words that tend to occur together.

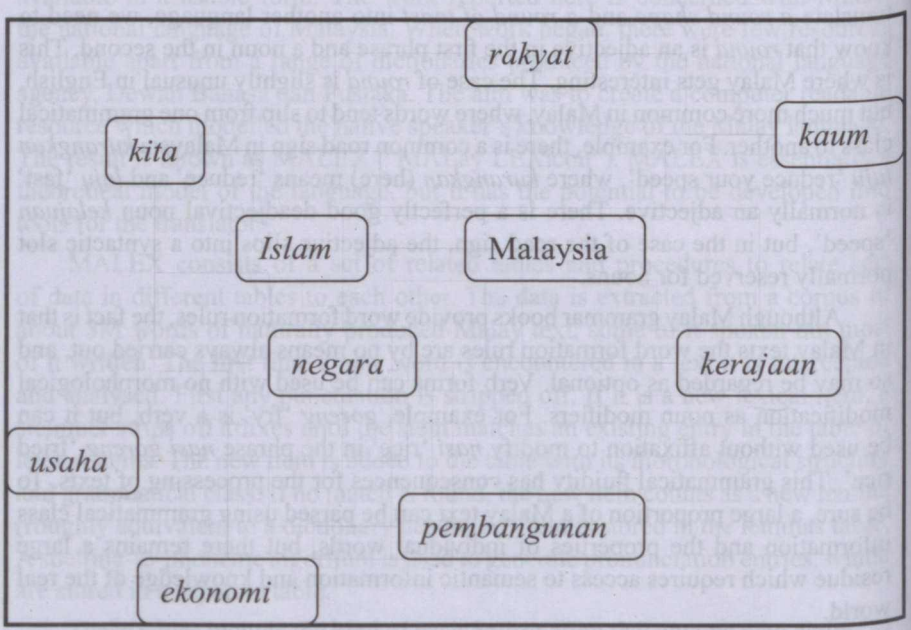

Figure 1: the network of key words.

Source: (Zuraidah Mohd Don, Knowles, & Fatt, 2010)

Figure 1 shows the results of identifying the key words and their collocates in a corpus of the political speeches of Tun Mahathir Mohamed during his time as Prime Minister of Malaysia. The key words collocate with each other and form a semantic network. An interesting subset contains the words *rakyat*

'people', *negara* 'state' and *kerajaan* 'government'. While these terms refer to modern democratic Malaysia, the basic forms *rakyat*, *negara* and *raja* 'king' are of Sanskrit origin and belong to an ancient form of society that can be traced back some 1300 years. At any rate, *rakyat* refers to the people as the ruled, and belongs to quite a different semantic field than *orang* 'person, people, human being(s)'. Similarly, the sense of *pembangunan* 'development' is constrained by its semantic relationship to *usaha* 'effort' and *ekonomi*. Now it is true that someone who knows the language well enough to translate it will be or should be intuitively aware of semantic relationships in political discourse of the kind presented in Figure 1. The point is that intuitive feelings about words can now be expressed as the measured properties of texts.

It is sometimes said that there is no such thing as a theory of translation. If this is true, then translators certainly need a theory of language. A crude word-for-word translation is restricted to lexis, and has no theory of grammar. A literal translation may follow the grammar but have inadequate access to semantics. Translators are not followers of linguistic theory, but have to wait for theoretical linguistics to catch up with their professional practice. After all, it is only since the 1970s that theoretical linguistics has had the categories required to translate *Sudah makan?* into English. An interesting development at the present time is that improved models of the intuitive linguistic knowledge of expert translators is likely to come not from conventional theoretical linguistics, but from advances in corpus linguistics.

REFERENCES

Zuraidah Mohd Don, Knowles, G., & Fatt, C. K. (2010). *Nationhood and Malaysian identity: A Corpus-based Approach*. Text and Talk, 30(3), 267-287.