COMPUTATIONAL LINGUISTICS IN MALAYSIA

Zaharin Yusoff
Computer-Aided Translation Unit
School of Computer Sciences
Universiti Sains Malaysia
11800 Penang, Malaysia
(zarin@cs.usm.my)

1. Historical Perspective

Computational linguistics in Malaysia began in early 1977 with the implementation of a morphological analyser for Malay for a Masters thesis at Universiti Sains Malaysia (USM) in Penang. The external examiner was Professor Bernard Vauquois from Groupe d'Etudes pour la Traduction Autonatique (GETA) in Grenoble, France. What started as an academic event led to a very long term collaboration between GETA and USM starting with the development of a prototype English-Malay machine translation (MT) system using GETA's ARIANE-78 system that further led to the development of various linguistic tools and language databanks as well as basic research on grammar formalisms. The collaboration continues to this day with the support of USM and the French government. The Computer Aided Translation Unit (known as UTMK) was set up in USM in 1984 and spearheads the collaboration from the Malaysian end. UTMK has slowed down considerably on MT but continues work on various other aspects of computational linguistics, their latest effort being on a search engine based on natural language processing techniques.

Another group began in Universiti Teknologi Malaysia (UTM) in 1988 with the arrival of the Japanese CICC project spearheaded by Fujitsu from the Japanese end bringing with them the Atlas-II system. A multi-lingual translation system project was set up involving several east Asian countries, with the Malaysian members being UTM and Dewan Bahasa dan Pustaka (DBP), the Malay Language Academy. The Malaysian component of the project was called Kanta, which lasted for about six years culminating in the establishment of the National Institute of Translation Malaysia (known as ITNM). ITNM has since stopped work on MT and so has UTM. UTM operated from Kuala Lumpur during the Kanta project but the main campus has moved to Johor Bahru.

A small but stable group exists in Universiti Kebangsaan Malaysia in Bangi within the Faculty of Information Technology. Their work centres on the development of language tools for Malay, such as natural language query, morphological analysers, etc. Work in other universities and research institutions is very limited and sporadic. There was an attempt at MT at Universiti Malaya in Kuala Lumpur back in the mid-eighties and some work on computer-aided language learning was carried out for a while at Universiti Institute Technologi MARA in Shah Alam.

To date, except for a handful (≤ five) which toyed with the idea of commercialising small language tools (e.g. Malay spell-checker), no private company has put in any serious effort or investment into computational linguistics. Government support other than for the Kanta project is indeed very low, and researchers have to compete with all other domains and disciplines for the limited R&D grants made available by the Ministry of Science, Technology and Environment (c.f. US$250 million for the entire period of the 7th Malaysia plan (1996-2000). The level of appreciation for the field has remained very low throughout its 23 year history in the country, especially given the rush for quick results and profitability by both the public and private sectors.

With the above, capacity within the country in terms of numbers and level of expertise in computational linguistics has never reached the required critical mass. If anything, interest in the field has been on a decline after its small peak in the late eighties and early nineties.

2. Computer-Aided Translation Unit

The Computer-Aided Translation Unit or *Unit Terjemahan Melalui Komputer* (UTMK) is the longest running research unit working in computational linguistics in the country. Maintaining a core team of 6-10 researchers and about the same number as contract personnel, it has also always been the largest. Apart from GETA (now GETA-CLIPS), which is UTMK's closest and longest running foreign collaborator, it has also worked with other computational linguistic research units/teams in UMIST, Bangkok, Prague, Sofia, Nijmegen, Kyoto, etc. UTMK's closest local collaborator is DBP, a partnership that dates back to 1982.

As mentioned earlier, UTMK began with the development of a prototype English-Malay MT system (for chemistry text) using GETA's ARIANE-78 system. The work was completed in 1985, but efforts to expand the prototype to an industrial system was not successful due to lack of financial support and too high expectations (translation of text books in very quick time). In parallel, a MT system shell called JEMAH was developed in LISP on the Macintosh (ARIANE-78 ran on the mainframe) and the English-Malay prototype application was completed in 1987.

As an immediate product for translation, a machine-aided human translation (MAHT) system called SISKEP was developed on the IBM PC in 1986 and a Macintosh version in 1987, the design being inspired by Melby's Mercury system. The system gained some measure of popularity and was in fact put on the market for almost a decade. Apart from the screen design and other language specific utilities, the basic components of SISKEP included a bilingual dictionary look-up with root-word extraction, Malay spell-checker (the first in the country) and thesaurus look-up (which led to the publication of the first ever edition of a Malay thesaurus), and these basic components were bundled in a desktop accessory called RakanBM that could work with any word processsor on the Macintosh or on Windows. The considerable success of the project led to the conception of a user-driven MT system in 1989 (together with UMIST), which is essentially SISKEP with JEMAH running in the background, where users may choose to toggle between MAHT facilities and full MT depending on their translation needs.

Realising the need to collect linguistic tools and in particular data for Malay in order to get any further with MT or any computational linguistic application for that matter, UTMK embarked on building linguistic tools and language databanks. A Malay text analysis system called MATA (including concordance, frequency counts, statistical analyses, etc.) was developed in 1983 on the mainframe and was later ported to the PC and then UNIX in 1990 to be included in a Corpus System for DBP. The system now holds Malay texts with a total of more than 12 million words. A Dictionary System was also developed for DBP in 1992. The system is a generic system that can generate dictionary systems as applications, and 5 DBP dictionaries have been implemented using the said system, including a first ever French-Malay dictionary (a collaborative work involving the French Embassy, DBP, GETA and UTMK). There was an attempt to build a general Lexical Database for Malay (Treasure Box: anything one wants to know about Malay) but the project never really got off the ground due to lack of funds. These projects occupied the period up to the end of 1995.

At the level of grammar, UTMK has always worked on grammar formalisms, namely the String-Tree Correspondence Grammar (STCG) based on the concept of Structured String-Tree Correspondence (SSTC). This also entailed work on graphic grammar editors, parsers and in particular a bilingual SSTC corpus bank to fuel various experiments in example-based MT as well as automatic generators of analysis and synthesis programs. One major setback to such efforts is the lack of formal linguistic studies for Malay, a situation that has also forced UTMK to work directly on Malay linguistics.

As of 1996, circumstances forced UTMK to venture into more commercial undertakings. One such project is the development of an EDI engine (parsing/generation of message types are needed). Others include a search engine (a commercial product) based on word ontology and distance, a multilingual chat system (real-time MT for a very restricted language), and a generic internet portal (…the relation to computational linguistics is still unknown…).