TOWARDS A LANGUAGE INFORMATION CENTRE FOR MALAY

Zaharin Yusoff
Computer-Aided Translation Unit
Universiti Sains Malaysia
11800 Penang, Malaysia
[e-mail: zarin@cs.usm.my]

Introduction

Malay is not only the national language of Malaysia, Indonesia and Brunei but it is also spoken in Singapore, southern Thailand and southern Philippines. It is indeed a major official language of the ASEAN community. However, despite the language being widely spoken and written, much is still to be done in the way of linguistic research and more so in the development of language processing systems and tools. In fact, very little of the literature and reference material in Malay is available in raw electronic form to support research and development work, let alone in forms that are readily accessible and manipulable on the information highway.

In an effort to remedy this situation, the *Computer-Aided Translation Unit* at Universiti Sains Malaysia in Penang has embarked on a joint effort with *Dewan Bahasa dan Pustaka* (Academy of the Malay Language) in Kuala Lumpur to set up a kind of Language Information Centre for Malay which will be placed on Internet and hence accessible to all nationally as well as internationally The Centre will contain various language databases pertaining to Malay, for example a corpus system, a wide variety of dictionary systems, and possibly a very general lexical database with as much information as possible about all words in Malay. The Centre will also make available various language processing tools that may contribute towards further development of language processing tools and applications. Other applications that may promote the usage of Malay are also of concern, in particular translation systems to and from Malay.

With the existence of such a Centre, general users, educators and learners of Malay will find it convenient to retrieve information and hence further enhance usage of the language, something which is a matter of policy for the country. More important for the linguistic community, the ready availability of the language databases would help quicken the pace of linguistic research, which at the moment is rather slow due to the difficulty in obtaining the data relevant to a given research. Although a very ambitious project and hence runs the risk of taking far too long to complete, some beginning for the general lexical database for Malay will certainly go a long way in providing valuable raw data for linguistic research. The language processing tools will be a major boon to many as they will help cut down on development time for building application systems. Needless to say, the availability of on-line translation systems for Malay will help many users in a wide range of domains ranging from research to commerce. On the other hand, it is the need to develop a machine translation system that prompted the *Computer-Aided Translation Unit* to spearhead the various projects leading to the conception of the Language Information Centre. Writing grammars for machine translation systems requires substantial linguistic data and only such an information centre can provide the requirements for effective development work.

The project is indeed a large one and it may be a while before the Centre is fully functional, but a basic infrastructure is being built at the moment and this paper reports on some of the work done towards achieving this aim. At this stage, the main concerns are at the technical level as well as problems in data collection and project management. Concerns about commercial and legal matters will no doubt arise in due course but they will have to wait until more concrete policies are formulated at the government level.

Basic Physical and Administrative Infrastructure

As have been alluded to in the introduction, the proposed Language Information Centre for Malay is not necessarily an administrative centre but rather an electronic centre. The language databases, processing tools and application systems are placed in machines that are accessible on the world-wide network, in particular Internet. In fact, the Centre may be

a virtual centre, in the sense that the machines on which the databases, tools and applications are placed may be physically distributed all over the country (or world for that matter), but that there is a common directory which guides the user to the application requested for. The actual access that goes across the country would of course be transparent to the user. This may in general slow down access time, but there is merit in placing the systems with their respective developers or data owners to ensure proper maintenance either for the software or for data coherence. Naturally there will be administrators assigned to the Centre, but they will be mainly responsible for ensuring communication between the various parties involved so that the Centre does not go 'down'. At any rate, in as far as the user is concerned, the Centre is logically a single unit that can be visualised as in the diagram further below.

The underlying idea here is that the language databases, processing tools and applications that have been, are being or will be implemented are first identified and then collected into the Centre. This collection may be physical, where the systems are handed over to the administrators of the Centre, or it may be virtual, where the owners register their systems at the Centre and allow access to users logging on to the Centre. Systems registered at the Centre but are situated at remote sites will have to be linked to the Centre via Internet or some other compatible connection. An automated directory of the various software and applications is prepared in such a way that a user needs only to log on to the Centre but obtains access to everything registered irrespective of whether the system requested is physically local or remote to the Centre. Users connect to the Centre via Internet (provided by MIMOS and called JARING), by dial-up, or through some other commercially available connections. Logging in may be via telnet or via certain client modules provided by the Centre which reside on the user machine that look for the directory at the Centre or directly for the appropriate application.

In terms of administrative organisation, one sees that there are basically three groups of people involved in the Language Information Centre:

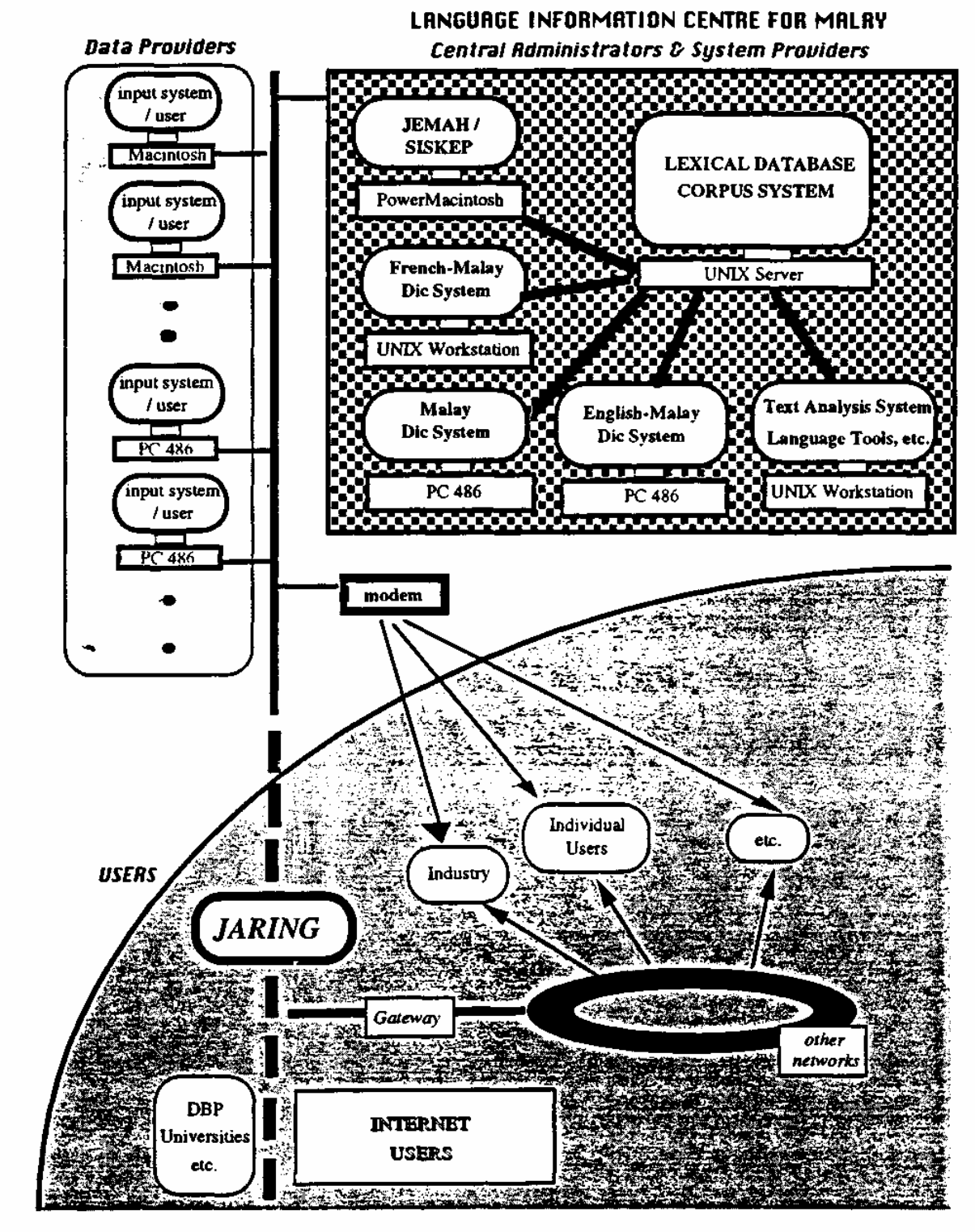
Central Administrators
These are the people who are at the front-line of providing service to users. They register clients as well as provide them with the necessary connection to the Centre, very much in the same way as local Internet administrators provide service to subscribers (in Malaysia, MIMOS takes this role). On the other end, they also register providers of databases, tools and applications as well as provide them with the appropriate links to the Centre to ensure accessibility if the systems are to be left at the respective remote sites. The central administrators are also responsible for coordinating all efforts in maintenance for both software and data. Above all, they have to make sure that the Centre is always 'up', which entails servicing all software and hardware at the Centre, and in particular the network links to and from the Centre.

System Providers
Databases, tools and applications may be handed over outright to the Centre in which case the systems will be handled by the central administrators at the Centre. The system providers may however be requested to maintain the systems for both software and data. The other possibility is that the systems are left at the original site of the system providers, in which case maintenance of the software and data for the systems will have to be their responsibility. Since the link between the Centre and remote sites of registered systems are critical, these remotes sites are administratively considered part of the Centre (which is as indicated in the diagram).

Data Providers
Language systems are typically data intensive, and so data providers are critical to the whole set up. Collection of linguistic data is not only a very time consuming process but it also has its own deep complexities especially when the data set to be collected requires interpretation at higher levels, like semantics, discourse, pragmatics, etc. If dictionary publishers require many lexicographers working full time to keep dictionary entries up-to-date, the same can be said about linguistic data. As such, various teams of data providers have to be formed for the various language systems registered at the Centre. Data providers have the responsibility of inputting new data, checking for coherence as well as keeping the data up-to-date at all times. The teams may deal directly with the central administrators or via the system providers depending on who have the responsibility over data maintenance.

The data providers may work remotely and so they will have to be connected to the Centre (also as indicated in the diagram).

As shown in the diagram below, others are essentially users, and they may be connected to the Centre via Internet, by dial-up, or through gateways from other networks. Apart from the application systems and tools which may reside on particular machines or use certain operating systems, the rest should be platform-independent.



Proposed Services

The kernel of the Language Information Centre will contain principally Malay language databases. The following have been or are being worked on at the moment:

- Corpus System:
  storage and retrieval system for texts in Malay ;
- Various Dictionary Systems:
  - Malay.
  - English->Malay,
  - French->Malay;
- General Lexical Database (prototype);
- Text processing tools:
  - Malay thesaurus,
  - English-Malay Terminology,
  - Malay spell-checker;
- Etc.

An elaboration on some of the major ones will be given a little further on. Another class of services pertains to tools for language processing that can be used for development of applications or for further development of tools. Amongst those which have been or are being implemented locally are the following:

- MATA:
  Malay text analysis system that produces statistical output about a given text;
- GWL:
  grammar-writing language for analysis and generation [Tong89a];
- JEMAH:
  integrated environment for developing machine translation systems [Tong89b];
- Tree Editor:
  for building a corpus of abstract linguistic representation structures,
  e.g. morpho-syntactic structures, functional structures, logical structures,
  multilevel structures, etc. [Vauquois78];
- STCG editor:
  editor and parser/generator for a bi-directional grammar formalism,
  namely the String-Tree Correspondence Grammar |Zaharin87];
- Etc.

There are of course many tools that can be placed in the Information Centre, not only those developed locally but also those already made available as shareware on the world-wide Internet. In terms of translation tools, we have the following which are readily available:

- SISKEP:
  a machine-aided human translator workstation for English->Malay [Tong87];
- UDMT:
  an English->Malay translation aid that basically combines JEMAH and SISKEP.
  and classed as a User-Driven Machine Translation system [Somers-et-al90];
- English->Malay JEMAH:
  a automated translation system for draft English->Malay translation;
- Etc.

Details on these may be obtained from the references provided. The main thing to note here is that the Information Centre is not to be simply an archive of systems and tools that a user makes copies from but one that is service-oriented. In other words, the Centre responds to queries by retrieving from one of its databases or by calling upon one of its tools and applications. A user may ask for some particular information which is obtainable from one of the language databases (e.g. all example sentences making use of a particular terminology) or he may request for a particular service (e.g. translation).

For the above, the Centre would require some sort of information broker to understand and subsequently respond to the query. The ultimate is to have a natural language interface that understands the query and then looks up the directory of services to pick up and call the relevant application. While waiting for the construction of such an interface, a query in a standard format should suffice, something very similar to those employed in various electronic data interchanges available commercially. What is important at the moment is the directory of services and the possibility of remote calls to the relevant services. The following discusses some of the major services to be made available at the Centre.

Corpus System

A corpus is one of the basic ingredients in linguistic research. Within this project, the original motivations for collecting Malay texts were threefold: corpus study for grammar writing in machine translation research, concordance for lexicographical work, and the need to keep an archive of examples of Malay literature of the 20th century. The usage of the corpus has however expanded to many other forms of linguistic study since the first two million words were put into the system.

Although there are quite a few corpus systems already available commercially or as shareware, an entirely new corpus system was built for this project [Zaharin-et-al89]. This is mainly to cater for the type of lexicographical work done here as well as the types of demand from the linguistic research conducted. There is also a need to ensure that modules for different functionalities may be readily added to the system in case new user requirements arise, some of which may be unique to Malay.

The Corpus System is written in C and runs under UNIX (Solaris) on the SUN. Its basic function is to provide concordance search for any string of characters. The search key can also be a combination of strings, root-words and syntactic categories. The system is combined with a Malay text analysis system that can provide various statistical output, like word count, output of new words, etc. Data-wise, the system currently holds texts from novels and text books totalling 10 million words, with the target being 200 million within 5 years. Some of the texts have been contributed to the European Corpus effort.

Once part of the Language Information Centre, users may either log on to the system via the directory of services, or be provided with a client module that locks on directly to the system engine. A smaller version may also be provided as a language processing tool that can be incorporated into other language processing systems.

Dictionary System

There are already a few dictionary systems involving Malay that have been built for PCs. for instance the one for the large English->Malay dictionary recently published by *Dewan Bahasa dan Pustaka,* and the one for the French->Malay dictionary which is currently in its final phase of compilation (it is also put up as a WEB server). These systems are also used for automatically generating smaller dictionaries and for experiments into generating inverse dictionaries.

Given that the logical structures for many dictionaries are very similar, the project is developing a general Dictionary System that can be used to quickly build dictionary systems for new dictionaries [Zaharin94a]. Given a new dictionary, the system is designed in such a way that the administrator only needs to input the logical structure of the dictionary and then import the data which is to be prepared in a standard format.

The Dictionary System is also written in C and runs under UNIX (Solaris) on the SUN. All words in the dictionary are indexed so that retrieval is very flexible. Flexibility is very much needed because the system is built not only to facilitate the setting up of on-line dictionaries but also for experiments in lexicographical work (e.g. as an aid for compiling new dictionaries from existing dictionaries).

As in the case for the Corpus System, once part of the Language Information Centre, users may either log on to the Dictionary system via the directory of services or be provided with a client module that locks on directly to the system engine. The main Dictionary System will also have a directory of dictionaries available at the Centre. A smaller version may also be provided as a language processing tool that can be incorporated into other language processing systems.

Lexical Database

The ultimate aim is to have a general lexical database that stores all information about all words in Malay. The principle is that if an information can be linked to a word then that information should be stored in the lexical database in some form or other. Once done, any information that can be related to a word should be retrievable from the database. As such, the database should be able to support queries ranging from simple and direct ones like *"what language does the word 'orang-utan' originate from?",* to indirect ones such as *"what are the possible semantic features of the indirect object of a 3-place argument verb if its direct object is animate?"*. The second query here also indicates that retrieval should be dependent on the content of the information set rather than its form, which is why information is best kept in coded form as opposed to the text form as in dictionary systems. In a dictionary system, one may have to ask *"what words are defined by means of the word 'take'?",* but in a lexical database, any feature of the concept 'take' can be made use of - the fact that there is an object involved, the location of the object, the owner or source, etc. For the lexical database to be as general as alluded to here, there should be no *a priori* list of possible queries supported, meaning that any form of query should be permissible as long as the information is contained in or is deducible from the facts in the database.

The size of the data is expected to be very large with about 155,000 words for Malay, where each word may have anything up to 1000 attributes. The data also covers a very wide range of knowledge in about 30 different subdomains of language and linguistics. No pretension is made here on the feasibility of completing such a project in a short or mid-range period. The massive size of the data points to the estimation that data collection alone may take decades, and this will be without complete certainty of the accuracy, consistency and coherence of the collected data, in particular over time where beliefs may evolve. The availability of the required level of expertise also has to be considered, in an effort to ensure that data collection will be carried out by competent personnel and in a consistent and effective manner. Problems in software implementation require the corresponding efforts in research and development where storage and retrieval time need to be carefully balanced. Worse, there is even very little idea, at the outset, as to the exact list of attributes to be associated to each lexical entry, and in what form should the data representation be to ensure efficient storage and retrieval given the very general requirements. Needless to say, the list of generic data as well as the contents pertaining to each entry will be subject to very frequent modifications in the development stage, and this also applies to any global or logical view of the structure of the data taken at any one point in time, which unfortunately translates to the data representation design.

Nonetheless, the project needs to begin, or otherwise it may never be completed. Such a project will force a situation where data needs to be collected and hence the corresponding studies be carried out, which will surely lead to a better position than no linguistic results or no data at all. Having a global long-term goal, albeit very ambitious, can be argued to be better then building smaller independent systems and hoping that they can be successfully integrated at a later stage. A global effort necessitates attention to be paid to issues of consistency and coherence. One very important feature that will consequently be inherent in the resulting product is that, in carrying out a long-term project such as this one, it is an absolute necessity to ensure that all data is completely recoverable at any stage of the development. This is not not only to cater for the frequent modifications expected but mainly to be able to service other smaller applications while awaiting completion.

More details pertaining to the development of the proposed general lexical database for Malay can be obtained from [Zaharin94b]. Issues discussed there include the general goal, difficulties in data definition and data representation, principles of system design and the ensuing system implementation. Efforts in data collection are also addressed along with a few points on project management.

Language Processing Tools

With the advent of Internet, more and more language processing tools (often together with source code) are being made available as shareware, and this is over and above those which are commercially distributed. Here, the Centre cannot hope to store all such software, but perhaps it can offer a directory of the locations and descriptions for such software. The more common ones (e.g. rootword extractors, morphological analysers and generators, etc.) may be physically stored at the Centre, and in particular those which are locally built.

Perhaps what is more important is to provide a more service-oriented point, where queries may be answered rather than being just a library to copy from. A considerable number of users are interested in using the tools for a very limited period of time (e.g. analysing a particular text) or that they are not in a position to handle the tools themselves, for often that requires some knowledge in computing. A central point in terms of availability, but geographically local with respect to the user, is also required because many users are not that familiar with browsing through Internet.

Whatever the case may be, language processing tools are very necessary to be made readily available in order to help quicken the pace in the development of language application systems. Such tools may not only avoid the duplication of work but may make possible the development of systems which may otherwise have been abandoned due to lack of expertise for developing a particular component.

Machine Translation

A Language Information Centre as proposed here should go a long way in helping towards the development of machine translation systems. To begin with, the Corpus System would provide the necessary texts for recognising patterns for writing grammars in the analysis and generation modules. The Dictionary System and the Lexical Database would no doubt provide the specialised lexicons in all phases of the system, and if the data in the Centre are general as well as large enough, it is highly likely that the dictionary modules can even be generated automatically.

Once stable, the machine translation systems can be put up as a service at the Centre. The simplest way is to leave the system in some easily locatable folder where users may run the system themselves. A better way is to put it as one of the functionalities of the Centre and so a user may first choose the appropriate translate function followed by sending the text to be translated. Translation in bulk may perhaps be done in a manner similar to the method adopted by some commercial electronic data interchanges, where a message bearing a header and the text to be translated is placed in some translation box and then the result is collected later on once a message from the Centre is received telling the user that the job has been completed. The header may contain information such as the address of the sender, the source and target languages, the subject domain and perhaps the choice of a particular machine translation system. A sort of translation broker can periodically scan through the messages in the translation box, send each message to the corresponding translation system, collect and store the results in some out box, and then send appropriate messages to the users. This assumes that all machine translation systems offering their services at the Centre must be in a form that can be activated via some function call.

An interesting possibility is to allow any machine translation system anywhere in the world to register its service at the Centre. Such systems, wherever they may be located, can be considered as part of the Centre like the other application systems indicated by the diagram given earlier. The maintenance of such systems clearly cannot be handled by the Centre, but the latter can perhaps send occasional messages (automatically) to check whether the remote systems are operational or whether the network links are up, failing which the said system may be temporarily taken off the directory of services. In fact, with proper interfacing, all these as well as user requests for translation may be operated

entirely via electronic mail, which is probably the most reliable and widespread form of electronic communication at this moment in time.

Concluding Remarks

The success or failure of natural language processing systems depends not only on useful and reliable processing software but also, and perhaps more so, on large available data. Many interesting tools have been developed and distributed widely but tend to be left unutilised because they lack sufficient data to be considered for immediate use. The development of many large systems have long suffered because there is not enough data to carry out the much needed research and development efforts to trigger crucial advances in the domain. Many specialised lexicons have been built, but the solution is clearly not there because they tend to be rebuilt over and over again with new applications. There is indeed a need for a centralised point for the storage and retrieval of general and in particular application-independent data that can used and reused over all domains of language application. This situation is very true for the Malay language where the development of natural language systems can be said to be very slow. For instance, such data is very much required for writing grammars in machine translation systems involving Malay, but data in electronic form, whether in its raw form like the corpus or structured data as found in a lexical database, have been unavailable for many years. The Language Information Centre for Malay as proposed in this paper is clearly a step towards, if not the ultimate answer for, remedying the situation.

A project of this size requires very careful planning in terms of costs, manpower and time. Apart from the development of the main software, support tools have to be built to help in the preparation and the actual data collection effort, not to mention the development of by-product systems to sustain continual interest and perhaps even to finance parts of the project. From the research and development point of view, it is best to view any effort towards establishing such an information centre at this moment not as a solution in itself but rather as a beginning that opens up many avenues for uncountable potential advances in the domain of language and linguistics. It is not so much what is done now, but what is important is to get the project started with any means possible, and then improve through experience - but of course one must always try to be on the lookout for intermediate results and by-products. No doubt the intricacies and problems involved in establishing a language information centre that can serve a wide range of language applications is potentially as good a testbed for research and development as machine translation has been in the field of linguistics and natural language processing in general.

REFERENCES

[Somers-et-al90] Harold Somers, John McNaught, Zaharin Yusoff. *A user-driven interactive machine translation system,* proceedings of the Seoul International Conference on Natural Language Processing, Seoul, November 1990 (4ms).

[Tong87] Tong Loong Cheong, *The engineering of a translator workstation,* Computers and Translation, vol. 2, USA, 1987.

[Tong89a] Tong Loong Cheong, *A data-driven control strategy for grammar writing systems,* Machine Translation, 4(4), USA, 1989, pp.177-1193.

[Tong89b] Tong Loong Cheong, *An integrated development environment for machine-aided translation,* proceedings of the 1st National Computer Science Conference, Kuala Lumpur, 1989, pp. 69-78.

[Vauquois79] Bernard Vauquois, *Description de la structure intermediaire,* communication presentée au colloque de Luxembourg, April 1978 (GETA document).

[Zaharin87] Zaharin Yusoff, *String-Tree Correspondence Grammar: a declarative grammar formalism for defining the correspondence between strings of terms and tree structures,* proceedings of the 3rd Conference of the European Chapter of the Association of Computational Linguistics, Copenhagen, April 1987, pp.160-166.

[Zaharin-et-al89] Zaharin Yusoff, Ng Kok Wan, Teh Hock Soon, *A corpus database,* International Workshop on Machine Translation and Computational Linguistics, Universiti Sains Malaysia, 18-22 Disember 1989 (11p).

[Zaharin94a] Zaharin Yusoff, *Struktur kamus dan penggunaannya.* Prosiding Seminar Perkamusan Melayu, Dewan Bahasa & Pustaka, Disember 1994.

[Zaharin94b] Zaharin Yusoff, *A lexical database for Malay,* proceedings of the International Conference on Linguistic Applications, Penang, July 1994 (pp.74-83).