Design and Development Procedure of an English-Malay Parallel Corpus

Tengku Sepora Tengku Mahadi, Helia Vaezian, Mahmoud Akbari

Universiti Sains Malaysia

Abstract: The current article aims to introduce a corpus compilation project at the School of Languages, Literacies and Translation of the Universiti Sains Malaysia.[1]. The research project included constructing a 500,000 word English-Malay parallel corpus of legal texts, developing an English-Malay translation memory of legal texts from the corpus and finally building a corpus-based glossary of legal terminology. The steps we followed before embarking on our corpus compilation project including the decisions over the size of the corpus as well as the number and the length of the texts to be included in the corpus are explained here. The paper further elaborates on the procedures followed to construct the corpus and develop the other products of the research project namely the Translation Memory and the Glossary.

1 Introduction

When corpora first entered Translation Studies as a discipline, their application was limited to the studies on the language of translation versus non-translated language (Baker, 1993). They however soon found their ways into other areas within the discipline. For instance, they were used in translation evaluation (Bowker, 2000) and in studies on

[1] http://www.usm.my/bi/

translator's style and ideology (Baker, 2000). Furthermore, Corpora came to be used in translator education. In this context, corpora were appreciated as valuable tools for both learners and teachers. They, among other thing, were shown to enhance learner's source text understanding (Bowker, 1998) and their understanding of specialized terms (Gavioli & Zanettin, 1997).

Considering all the benefits of corpora, many universities around the world have invested considerable funds on various corpus compilation projects from large-scale projects on building national corpora to smaller projects on building various parallel and comparable corpora. The present paper elaborates on a corpus compilation project at the School of Languages, Literacies and Translation of the Universiti Sains Malaysia. Despite being totally independent and self-sufficient in itself, the current research took off more as a feasibility and pilot study for a considerably huge project of constructing the Malay National Corpus. It has made use of new technology and a fresh corpus building method to have an appropriate estimate of the length, human resources and the funding needed among others for the development of the Malay National Corpus.

2 What is a corpus?

The term "corpus" is the Latin word for "body" and it refers to any body of text. McEnery and Wilson (2001) define corpus as "a finite-sized body of a machine readable text, sampled to be maximally representative of the language variety under consideration" (p. 32). For Tognini-Bonelli (2001) a corpus is "a computerized collection of authentic texts amenable to automatic or semiautomatic processing or analysis" (p.55). According to her (ibid), "the texts are selected according to explicit criteria in order to capture the regularities of a language, a language variety or a sub-language". The representativeness of the language included in the corpus is one of the most important issues with regards to corpora. The authenticity of texts included in a corpus is also of prime importance. Tognini-Bonelli (2001: 55) defines this issue in the following words "All the material included in the corpus, whether spoken, written or gathered along any intermediate dimension, is assumed to be taken from genuine communication of people going about their normal business". Last but not least, all the texts in a corpus are in machine readable format.

3 Types of corpora

There are various types of corpora from monolingual general reference to bilingual parallel corpora. Within the field of Translation Studies, however, some types of corpora have proved more useful than others. Parallel or translation corpora along with comparable corpora are two of the most used corpora in Translation Studies.

Kenny defines *parallel corpus* as "a body of texts in one language along with their translation into another language" (2001: 62). Parallel corpora can be either unidirectional or bidirectional. *Unidirectional parallel corpus* is defined by Olohan (2004: 24) as "a corpus containing source texts in language A and target texts in language B". She further states that *bidirectional parallel corpus* is a corpus containing "source texts in language A and target texts in language B, and source texts in language B and their translations into language A" (ibid).

Comparable corpora have been defined by Teubert (2004) as corpora in two or more languages with the same or similar composition. Olohan (2004) distinguishes three types of comparable corpora namely monolingual, bilingual, and multilingual comparable corpora. Olohan defines monolingual comparable corpora as corpora consisting of “translations and comparable non-translation in the same language” (ibid: 35). She further refers to bilingual or multilingual comparable corpora as “corpora of comparable original texts in two or more languages" (p. 35).

4 Corpus compilation process

The process of corpus compilation like any other purposeful activity requires some planning. According to Atkins et al., issues such as the size of the corpus, the range of the language varieties, the time span to cover and the level of encoding details to be included in electronic form are among the subjects compilers must consider prior to embarking on corpus building process (cited in Pearson, 1998).

In fact, it can be claimed that it is the purpose of the corpus which determines the design of the corpus and all the following decisions. So the very first step in every corpus compilation project would be defining the purpose or the intended function of the corpus. Defining the purpose of the corpus would further provide answers to questions regarding the favorable size of the corpus and the period as well as the genre(s) the corpus is to cover. Other issues such as whether the corpus includes written or spoken transcription and whether full texts or samples of texts are used are also determined by the purpose for which the corpus is intended. Last but not least, it is necessary to consider issues such as the time and funding available to meet the intended purpose. As Meyer (2002) states in determining how large a corpus should be, it is necessary “to compare the resources that will be available to create it (e.g. funding, research assistants, computing facilities) with the amount of time it will take to collect texts for inclusion, computerize them, annotate them, and tag and parse them" (p. 32).

5 Corpus analysis tools

Corpus analysis tools enable users to manipulate the information contained within a corpus in a variety of ways. Bowker (2002: 47) defines corpus analysis tools as computerized tools that help translators to "manipulate and investigate the contents of the corpus".

Word-frequency list is the most basic feature of the corpus analysis tools which makes it possible for user to discover how many different words are in a corpus and how often each of them appears. Concordancer is another basic feature of most corpus analysis tools. This feature as defines Bowker (ibid: 53) is "a tool that retrieves all the occurrences of a particular search term in its immediate context and displays these in an easy-to read format". Most corpus analysis tools have a feature to compute collocations, i.e. co-occurrence patterns of words. This feature, in fact, determines "whether two words are collocates by comparing the actual co-occurrence patterns of pairs of words against the pattern that would have been expected if the words were randomly distributed throughout the text" (ibid: 64).

6 Applications of corpora in Translation Studies

The use of corpora in translation studies is relatively new. It, arguably, dates back to 1993 when Baker first discussed the importance of applying corpus evidence to study the nature of translated texts per se. Baker stated that translation as a unique communicative event must be recorded and explored and corpora can well serve the purpose by providing “theorists of translation with a unique opportunity to observe the object of their study and to explore what it is that makes it different from other objects of study" (ibid: 135).

Based on theoretical statement, Baker (1996) suggested applying corpus tools to investigate four features of translated texts or translation universals namely *simplification*, *explicitation*, *normalization* and *levelling out.* In line with this approach, a considerable number of studies within the discipline of Translation Studies have focused on applying corpora to study translation universals during the recent years.

Translation evaluation is another area within Translation Studies which has enjoyed the benefits of corpora during recent years. Corpora, by their very nature, are representation of language as it really is used; they thus can provide evaluators with solid criteria to evaluate and assess translations. Bowker (2000) is possibly the first translation scholar to adopt a corpus-based approach to translation evaluation. Drawing on advantages of corpora over the conventional resources available to evaluators i.e. dictionaries, parallel texts, subject field experts and intuition, Bowker asserts that corpora have the potential to overcome many shortcomings inherent in traditional translation evaluation. She further uses an evaluation corpus comprising three subcorpora namely Quality Corpus, Quantity Corpus and Inappropriate Corpus to test the applicability of corpora in translation evaluation. The results of her comprehensive study show that the evaluation corpus “allows evaluators to both identify and correct a larger number of errors, and to do so in a more objective way" (ibid: 206).

Corpora further came to be used in studies on translator’s style and ideology. An example of corpus-based research on translator's style is the one carried out by Baker

(2000). Baker uses corpora to show how linguistic habits of two translators differ and how they are reflections of translators' styles (ibid). She uses type/token ratio, average sentence length, and the translator's usage of the lemma *say* to look for the differences in the styles of two translators (ibid).

Translator Education has benefited from the introduction of corpora into the discipline as well. Bernardini (1997) is among the first scholars to ask for the integration of corpora into translation classrooms. She back in 1997 suggested that the traditional translation teaching should be complemented with large corpora concordancing to let the trainees acquire skills they need to become (professional) translators. According to Bernardini (2004: 20), while parallel corpora can "act as an expert system, drawing the learner's attention to (un)typical solutions for typical problems found by mature, expert translators", comparable corpora can provide learners with information as for "typical turns of phrases, collocations, terms and their lexico-syntactic environment". Zanettin shares the same view when he states "learning to use corpora as translation resources should (also) be part of the curriculum of future translators and become part of their professional competence" (2002: 10).

Last but certainly not least, corpora proved useful to computerized translation systems. As laviosa states "computerized systems such as translation memories and example-based and hybrid MT programs have also been enhanced by the statistical and lexico-grammatical analysis of corpora" (2003: 105).

7 The corpus compilation project at the school

The present paper aims to introduce a corpus compilation project at the School of Languages, Literacies and Translation of the Universiti Sains Malaysia. The research project included constructing a 500,000 word English-Malay parallel corpus of legal texts. This section elaborates on our research objectives and a number of important decisions we made prior to embarking on our corpus compilation project.

7.1 Purpose of the research

The objectives of the present project included: 1) compiling a bilingual corpus to be used in a) future corpus-based translation and linguistic researches and b) translation classrooms as a new translation resource for students; 2) building a Translation Memory to be applied by professional and student translators at the school and finally 3) generating a corpus-driven glossary.

7.2 Why parallel corpus?

After discussions among the research team members, it was decided that a parallel corpus would best serve the purpose of the project. The languages included in our parallel corpus are English and Malay with the direction of translation from English into Malay.

Parallel corpora are valuable resources when it comes to bilingual lexicography. They are also invaluable resources for translation researchers. Such corpora are very rich sources of terminological and phraseological information and can be applied by translation researchers to gain contrastive knowledge about the source and target languages. They also have a lot to offer to translators. Parallel corpora can be used by translators to see what past translators have done and how they have done it. As

Machniewski (2006) states corpora of translations provide translators with insights into strategies employed by past translators. Apart from that, parallel corpora offer the possibility to search for examples of translations in their context and so they can provide a useful supplement to decontextualised translation equivalents listed in dictionaries. Moreover, parallel corpora can be turned into translation memories rather easily. Parallel corpora, by their very nature, are repositories of source texts along with their translations and as such if aligned they can be used as translation memories.

7.3 Text type

It was decided to draw on legal genre in the present project due to several reasons. First, the School of Languages, Literacies and Translation of the Universiti Sains Malaysia has an internal translation service center with important clients from the industry and the government with a considerable number of texts for translation which fall within the domain of legal genre. A corpus of legal texts and a legal translation memory, thus, would be very useful for the center. Second, it was believed that finding original English texts and their translated Malay texts in legal genre would be easier compared to other genres. Third, there were already a considerable number of legal texts translated from English into Malay readily available in soft copy format in the translation service center of the school. Last, it was believed that a corpus-driven English-Malay legal dictionary would be highly needed in the market and that the present project could be a step towards the development of such a dictionary.

7.4 Size of the corpus

Being a short term research project with limited time and funding, we decided to keep the size of the corpus to a maximum of 500,000 words with each subcorpus containing approximately 250.000 words.

8 Corpus construction phase

The present section elaborates on the process of preparing the texts for inclusion in our parallel corpus.

8.1 Image file preparation

Scanners have been the common tools used for converting hardcopy to softcopy format for a long time. There are, however, some drawbacks in using scanners to convert hardcopy to softcopy format in big projects like the one described here; the method is not very convenient as the process is not completely automatic and it is time consuming.

Half of the current project was done using this conventional method and the average speed was 90 A4 pages per hour for converting hard copy to image file format. Later on, the project team was inspired by a video posted on Youtube[2] a couple of months ago. The video is a showcase on a project by two Japanese researchers who have apparently designed a new generation of superfast scanners which can scan with the speed of 200 pages per minute (for further information please visit: http://www.youtube.com/watch?v=CX9dnoEc53Y).

We decided to develop a much simpler version of that scanner based on the youtube video; therefore, we mounted a Canon camera on top of a cubic structure with the lens downward and placed the book at the bottom and started shooting and flipping

[2] www.youtube.com

the pages. This method remarkably increased our speed to an average of 700 A4 pages per hour.

8.2 Converting image files to machine readable format

Omnipage is a well-known OCR software which is usually supplied with every scanner purchased and its performance can be considered quite acceptable for an average user. It, however, may not be as practical when it comes to more serious projects like building a huge corpus. There are two major drawbacks to it, namely its rather low speed and accuracy.

For the purpose of this project, we employed READIRIS Pro which is a powerful and professional OCR software. Application of READIRIS Pro remarkably increased our speed and the accuracy of the converted files.

8.3 Verification process

Although the accuracy of READIRIS Pro software is quite acceptable, we still had to check the texts before building the parallel corpus for two main reasons: 1. correcting some strange symbols and misspelled words in the file which could be less due to the software inability in recognizing the letters and more due to the low quality of the print; 2. removing any discrepancy at sentence and paragraph level between the source and the target texts such as segments left untranslated or additions of sentences or paragraphs to the translation.

The strange symbols and misspelled words in the text were quickly located by the help of MS word spelling function and were replaced by the correct word with reference to the original document. The discrepancies were also removed by reading through both texts in order to have perfect parallel texts.

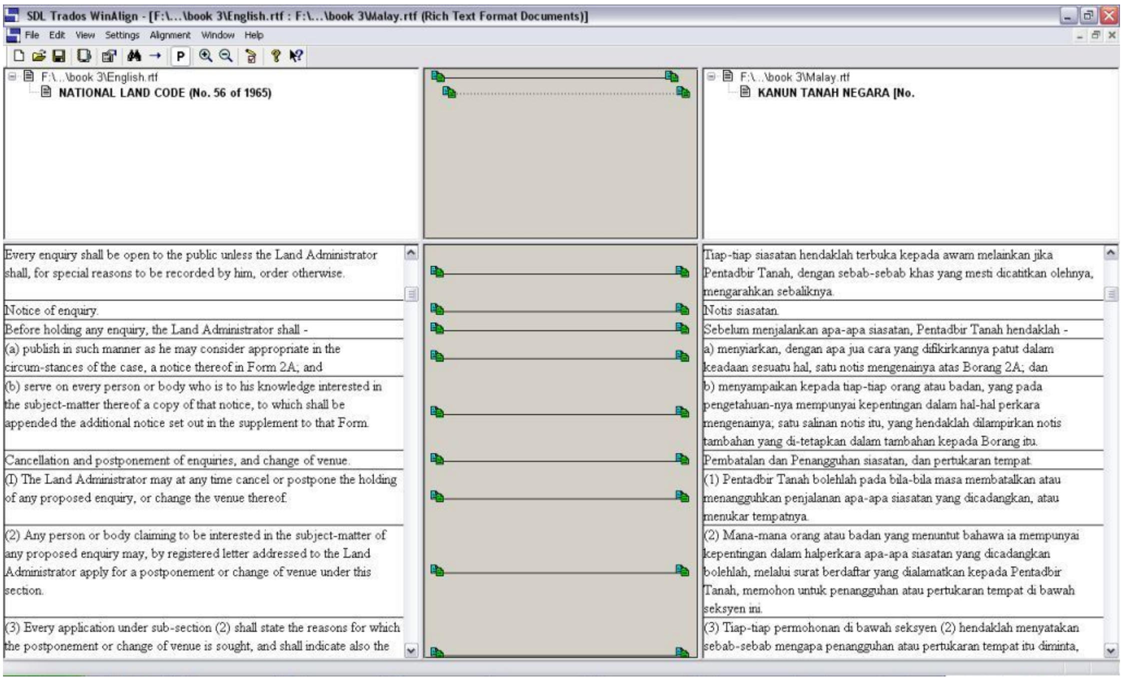
After verifying each and every pair of source and target texts, we merged all the texts of each language together and saved them in RTF and Plain text formats to form the two components of our parallel corpus. In these formats, they are readily available for utilization by students, teachers, and researchers.

9 Building the TM

Translation Memories (TM), as defined by Bowker, are repositories of “source texts segments explicitly aligned with their target texts counterparts” (2002, p 92). They, in fact, can be considered as data banks from which translators can retrieve already translated segments that match a current segment to be translated. There are generally two ways to build such repositories: using empty TMs to carry out translations and using existing translations as input for TMs.

The first way is described by Somers (2003) as the simplest method to build TMs since each sentence is automatically added to translation memory database as one goes along the process of translation. The second method is to “take an existing translation together with the original text and have the software build a TM database from it automatically” (ibid, p. 34). The only issue which needs to be considered in the second method is that of alignment which is defined by Bowker (2002, p. 109) as “the process of comparing a source texts and its translation, matching the corresponding segments, and binding them together as translation units in a TM”.

Automatic alignment tools can be used to carry out the above process. The tool which was employed for the purpose of this project is Trados WinAlign. It is one of the powerful tools in the SDL Trados suite with useful features for alignment. It determines parts of the source and target language files which belong together and puts them side by side. The users can have an interactive part in the alignment process. They are able to optimize the alignment results through modifying alignments and editing text segments directly. After running the WinAlign software, adding the texts and clicking on the *align file names* button, the software aligns the text as you can see in figure 1. Then we have to check the alignment for any misaligned segment. It is worth mentioning that our checking revealed 100% accuracy of the aligned segments.


SDL Trados WinAlign - [F:\...\book 3\English.rtf : F:\...\book 3\Malay.rtf (Rich Text Format Documents)]
File Edit View Settings Alignment Window Help
P

F:\...\book 3\English.rtf
NATIONAL LAND CODE (No. 56 of 1965)

F:\...\book 3\Malay.rtf
KANUN TANAH NEGARA [No.

Every enquiry shall be open to the public unless the Land Administrator shall, for special reasons to be recorded by him, order otherwise.	Tiap-tiap siasatan hendaklah terbuka kepada awam melainkan jika Pentadbir Tanah, dengan sebab-sebab khas yang mesti dicatitkan olehnya, mengarahkan sebaliknya.
Notice of enquiry.	Notis siasatan.
Before holding any enquiry, the Land Administrator shall -	Sebelum menjalankan apa-apa siasatan, Pentadbir Tanah hendaklah -
(a) publish in such manner as he may consider appropriate in the circum-stances of the case, a notice thereof in Form 2A; and	a) menyiarkan, dengan apa jua cara yang difikirkannya patut dalam keadaan sesuatu hal, satu notis mengenainya atas Borang 2A; dan
(b) serve on every person or body who is to his knowledge interested in the subject-matter thereof a copy of that notice, to which shall be appended the additional notice set out in the supplement to that Form.	b) menyampaikan kepada tiap-tiap orang atau badan, yang pada pengetahuan-nya mempunyai kepentingan dalam hal-hal perkara mengenainya; satu salinan notis itu, yang hendaklah dilampirkan notis tambahan yang di-tetapkan dalam tambahan kepada Borang itu.
Cancellation and postponement of enquiries, and change of venue.	Pembatalan dan Penangguhan siasatan, dan pertukaran tempat.
(I) The Land Administrator may at any time cancel or postpone the holding of any proposed enquiry, or change the venue thereof.	(1) Pentadbir Tanah bolehlah pada bila-bila masa membatalkan atau menangguhkan penjalanan apa-apa siasatan yang dicadangkan, atau menukar tempatnya.
(2) Any person or body claiming to be interested in the subject-matter of any proposed enquiry may, by registered letter addressed to the Land Administrator apply for a postponement or change of venue under this section.	(2) Mana-mana orang atau badan yang menuntut bahawa ia mempunyai kepentingan dalam halperkara apa-apa siasatan yang dicadangkan bolehlah, melalui surat berdaftar yang dialamatkan kepada Pentadbir Tanah, memohon untuk penangguhan atau pertukaran tempat di bawah seksyen ini.
(3) Every application under sub-section (2) shall state the reasons for which the postponement or change of venue is sought, and shall indicate also the	(3) Tiap-tiap permohonan di bawah seksyen (2) hendaklah menyatakan sebab-sebab mengapa penangguhan atau pertukaran tempat itu diminta,

Figure 1 Sentence alignment in WinAlign

After checking the alignment, we saved and exported the project which in fact resulted in the construction of the TM. Later on, as new texts were available for inclusion in the corpus and TM, we went through the same process explained above and finally merged all the created TMs to form a single Translation Memory.

10 Glossary compilation

To extract the legal terminology from our parallel corpus, we made use of SDL MultiTerm Extract application. Upon running the application and choosing *new project* from *SDL* MultiTerm Extract file menu, we are provided with five options in New Project Wizard, one of which is the *Dictionary Compilation project* option. TMX, TMW, and TTX are all among the supported file formats for this operation. After clicking on *Dictionary Compilation project*, we are able to add our TM. By clicking ‘Finish’ button, the project is created using the default settings.

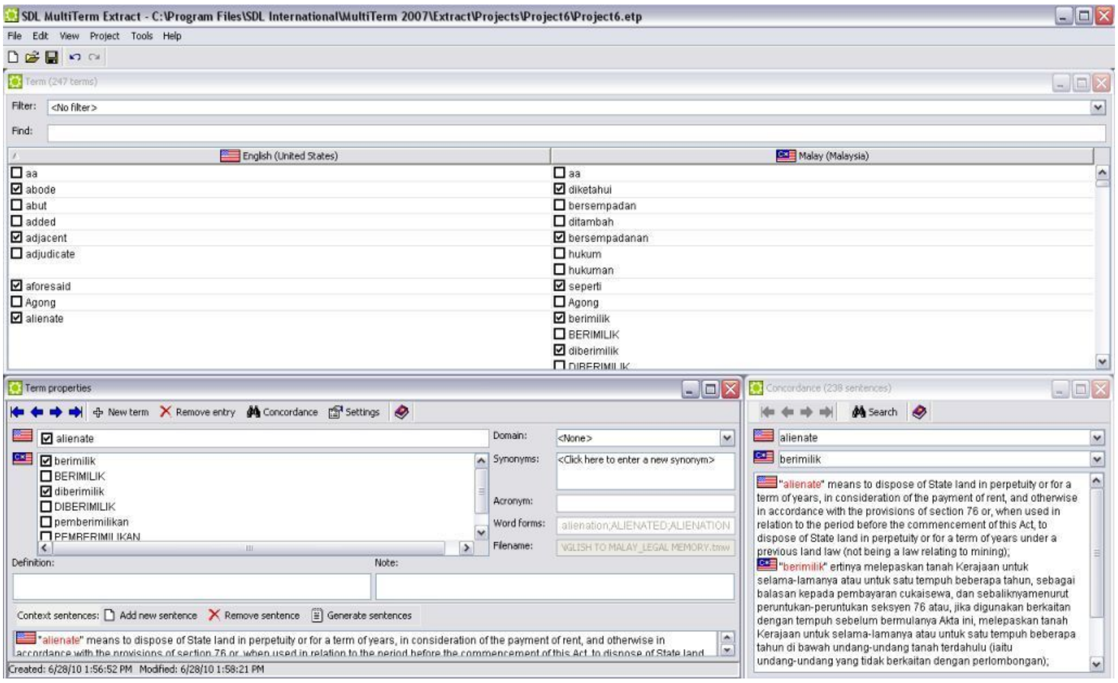
However, since our goal was to create a legal termbase, we needed to separate the legal terminologies from general terms. This option is available by drawing on ‘Excluded Terms’ window where we can upload a file of general terms to be excluded from the termbase.

In order to prepare the ‘Exclude File’, we first had to separate general terms from legal terms. To do so, we created a wordlist from the English sub-corpus using WordList feature of WordSmith 5 tool which displays a list of all the words or word-clusters in a text, in alphabetical or frequency order. After creating the wordlist from the English sub corpus, we manually went through the list, selected all the general terms, and saved them in a separate file named *Excluded Terms*. Returning to the ‘Excluded Terms’ window, we uploaded the *Excluded Terms* file.

We then started the operation by clicking ‘finish’ button and confirming the operation in SDL *MultiTerm Extract Confirm* dialog box.

10.1 Terminology validation and export

By clicking on **Yes** button in SDL *MultiTerm Extract Confirm*, the **Term Extraction** dialog box is displayed with a progress bar indicating the completion of the extraction process. Once the progress bar has reached 100%, we must click on **OK** button to view the extracted terms in SDL MultiTerm Extract and validate them. SDL MultiTerm Extract also allows us to add more terms to the term candidate list by manually extracting them from the TM.



SDL MultiTerm Extract - C:\Program Files\SDL International\MultiTerm 2007\Extract\Projects\Project6\Project6.etp

Filter: <No filter>

Find:

English (United States)	Malay (Malaysia)
□ aa	□ aa
☑ abode	☑ diketahui
□ abut	□ bersempadan
□ added	□ ditambah
☑ adjacent	☑ bersempadanan
□ adjudicate	□ hukum
	□ hukuman
☑ aforesaid	☑ seperti
□ Agong	□ Agong
☑ alienate	☑ berimilik
	□ BERIMILIK
	☑ diberimilik
	□ DIBERIMILIK

☑ alienate

☑ berimilik
□ BERIMILIK
☑ diberimilik
□ DIBERIMILIK
□ pemberimilikan
□ PEMBERIMILIKAN

Domain: <None>

Synonyms: <Click here to enter a new synonym>

Acronym:

Word forms: alienation;ALIENATED;ALIENATION

Filename: NGLISH TO MALAY_LEGAL MEMORY.tmw

Definition:

Note:

"alienate" means to dispose of State land in perpetuity or for a term of years, in consideration of the payment of rent, and otherwise in accordance with the provisions of section 76 or, when used in relation to the period before the commencement of this Act, to dispose of State land

Created: 6/28/10 1:56:52 PM Modified: 6/28/10 1:58:21 PM

alienate

berimilik

"alienate" means to dispose of State land in perpetuity or for a term of years, in consideration of the payment of rent, and otherwise in accordance with the provisions of section 76 or, when used in relation to the period before the commencement of this Act, to dispose of State land in perpetuity or for a term of years under a previous land law (not being a law relating to mining);

"berimilik" ertinya melepaskan tanah Kerajaan untuk selama-lamanya atau untuk satu tempuh beberapa tahun, sebagai balasan kepada pembayaran cukaisewa, dan sebaliknyamenurut peruntukan-peruntukan seksyen 76 atau, jika digunakan berkaitan dengan tempuh sebelum bermulanya Akta ini, melepaskan tanah Kerajaan untuk selama-lamanya atau untuk satu tempuh beberapa tahun di bawah undang-undang tanah terdahulu (iaitu undang-undang yang tidak berkaitan dengan perlombongan);

Figure 2 MultiTerm Extract interface

The SDL MultiTerm Extract interface is divided into 3 parts (Figure 2):

- **Term window** which contains a list of all the terms that SDL MultiTerm Extract has extracted in the current project, allowing us to validate or invalidate each term and its candidate translations.
- **Term properties window** which has three main areas: *Term and translation information area* in which we can validate the source language term and its translation, enter category, synonyms, and antonym for the term, and see information about the word forms, file name, and the date and time the term was created or modified; *Definition and Note Boxes* area in which we can enter the definition of the term and add comments and notes about the term or its translation; *Context Information* area in which we can add any number of new sentences showing the term in context or generate sentences containing the term from the TM
- **Concordance window** which displays the extracted terms in context

After validating the terms and adding additional data for each term, the second stage of building the term base is complete. For the purpose of this project, we exported the terms in two formats: Project Termbase and Tab-deliminated Text formats. Tab-deliminated Text format is the actual glossary while Project Termbase format is used to run the Legal termsbase accessible through both MultiTerm software and Trados WorkBench. Translators can increase their translation speed, translation quality, accuracy, and terminology consistency through the application of these two products.

11 Final remarks

The present project which was completed during 3 academic semesters produced valuable products. Our English-Malay parallel corpus can be used both for translation teaching and translation evaluation. It can further be applied in studies on Translational Malay language in the legal genre. The translation memory and the termbase are also of great value for both translation students and professional translators at the Translation Center of the school. The statistics drawn from this project on the other hand would help us to have an appropriate estimate of the budget, length, and human resources needed for the development of the Malay National Corpus.

References

Baker, M. (1993). Corpus Linguistics and Translation Studies: Implications and Applications. In Baker, M, Francis, F. & Tognini-Bonelli, E. (Eds.) *Text and Technology: In honor of John Sinclair*. Amsterdam and Philadelphia: John Benjamins, pp. 233-250.

Baker, M. (2000). Towards a methodology for investigating the style of a literary translator. *Target* 12: 241-266.

Bernardini, S. (1997). A Trainee Translator's Perspective on Corpora. Paper presented at *Corpus use and learning to translate*. Bertinoro. Retrieved April 6, 2008, from http://www.sslmit.unibo.it/introduz.htm.

Bowker, L. (1998). Using specialized monolingual native language corpora as a translation resource: A pilot study. *Meta* 43 (4): 631-651.

Bowker, L. (2000). A Corpus-Based Approach to Evaluating Student Translations. *The Translator* 6(2): 183-210

Bowker, L. (2002). *Computer-Aided Translation Technology: A Practical Introduction.* Canada: University of Ottawa Press.

Gavioli, L. & Zanettin, F. (1997). Comparable corpora and translation: a pedagogic perspective, Paper presented at the first international conference on *Corpus Use and Learning to Translate*. Bertinoro. Retrieved April 6, 2008, from http://www.sslmit.unibo.it/introduz.htm.

Kenny, D. (2001). *Lexis and Creativity in Translation: A corpus-based study*. Manchester: St Jerome Publishing Company.

Laviosa, S. (2003). Corpora and the translator. In Somers, H (Ed.) *Computers and Translation: A translator's guide.* Amsterdam and Philadelphia: John Benjamins. pp.105–117.

Machniewski, M. (2006). Analysing and teaching translation through corpora: lexical convention and lexical use. In *Poznań Studies in Contemporary Linguistics 41: 237-255*. Retrieved May 29, 2008, from http://ifa.amu.edu.pl/psicl/files/PSiCL_41_Machniewski.pdf

McEnery, T. & Wilson, A. (2001). *Corpus Linguistics*: *An Introduction* (2nd edition). Edinburgh: Edinburgh University Press.

Meyer, C. F. (2002). *English Corpus Linguistics: an introduction*. Cambridge: Cambridge University Press.

Olohan, M. (2004). *Introducing Corpora in Translation Studies*. London New York: Rutledge.

Pearson, J. (1998). *Terms in Context*. Amsterdam / Philadelphia: John Benjamin Publishing Company.

Schmied, J. (2002). A translation corpus as a resource for translators: The case of English and German prepositions. In Maia, B. (Ed.) *Translators as Service Providers*. Porto: Universidad de Porto, pp. 251-269.

Somers, H. (2003). *Computers and Translation: A translator's guide*. Amsterdam and Philadelphia: John Benjamins.

Teubert, W. (2004). Language and Corpus Linguistics. In Halliday, M. A. K., Wolfgang, T., Colin, Y. & Cermakova, A. *Lexicology and Corpus Linguistics: An introduction* .London: MPG Book Ltd, pp. 73-113.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam/Philadelphia: John Benjamin Publishing Company.

Zanettin, F. (2002). Corpora for translation practice. In Yuste-Rodrigo E. (Ed.) *Language Resources for Translation Work and Research*, LREC 2002 Workshop Proceedings, Las Palmas de Gran Canaria, pp. 10-14.