LOW-COST MULTILINGUAL LEXICON CONSTRUCTION FOR UNDER-RESOURCED LANGUAGES

LIM LIAN TZE

DOCTOR OF PHILOSOPHY

MULTIMEDIA UNIVERSITY

FEBRUARY 2013

LOW-COST MULTILINGUAL LEXICON CONSTRUCTION FOR UNDER-RESOURCED LANGUAGES

BY

LIM LIAN TZE

B.Sc. (Hons), University of Warwick, United Kingdom

M.Sc., Universiti Sains Malaysia, Malaysia

THESIS SUBMITTED IN FULFILMENT OF THE

REQUIREMENT FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

(by Research)

in the

Faculty of Computing and Informatics

MULTIMEDIA UNIVERSITY
MALAYSIA

February 2013

The copyright of this thesis belongs to the author under the terms of the Copyright Act 1987 as qualified by Regulation 4(1) of the Multimedia University Intellectual Property Regulations. Due acknowledgement shall always be made of the use of any material contained in, or derived from, this thesis.

© Lim Lian Tze, 2013
All rights reserved

DECLARATION

I hereby declare that the work has been done by myself and no portion of the work contained in this Thesis has been submitted in support of any application for any other degree or qualification on this or any other university or institution of learning.

_____

Lim Lian Tze

ACKNOWLEDGEMENTS

This thesis, or indeed this entire research, although wholly my own, would not have been possible without the wonderful help, wise guidance and tremendous kindness of many people.

To my supervisors: thank you for your continuous guidance, thoughtful advice, the (hopefully occasional) prodding in helping me mould what was initially a mess of shapeless, directionless ramblings and rantings into some semblance of coherent research.

Dr Tang Enya Kong has always kept me firmly focused on computational lexicography as my main research direction, always showing me new perspectives, but not letting me stray overmuch to tumble over the proverbial cliff.

Dr Bali Ranaivo-Malançon is simultaneously a spirit-lifting cheer-leader and meticulous inquisitor, showing me that the research journey is not as treacherous as some make it out to be – if only one know what the true path can be like. Your gracious invitation to me as the keynote speaker to MALINDO was especially important for me to realise the importance of working with under-resourced languages.

Tremendous thanks to Dr Soon Lay-Ki and Dr Lim Tek Yong, who took me under their wings, and provided some much needed objective perspectives – especially from related but distinctly different domains – to ensure that my writings and results are coherent and clear in the latter part of my research. Thank you for your careful scrutiny, all-round checks and overall shepherding.

To my collaborators: thank you for sharing your time, resources and expertise, without which many of the work described in this thesis would not have been completed. I thank Suhaila Saee and Panceras Talita, from Universiti Malaysia Sarawak, for sharing your resources on Iban–English dictionaries, as well as helping to prepare the Iban test data. Same goes to Vee Satayamas from Katsetsart University who pointed me at Yaitron. I thank also Jonathan a.k. Sidi, Jennifer Wilfred, Doris a.k. Francis Harris, Robert Jupit, Wong Li Pei, Tan Tien Ping, Gan Keng Hoon and Saravadee Sae Tan for their efforts in evaluating the results.

To the postgraduate affairs managers: especially Ms Raja Nurul Atikah at the Faculty of Computing and Informatics, Mr Kamal Eby Shah Sabtu and Mr Faizul Kamari at the Institute of Postgraduate Studies – thank you for your meticulous organisation and shepherding of the various administrative procedures, from my registration right up till my thesis submission. Thank you for patiently responding to my queries about various issues all this while.

To my parents: thank you for your unbounded love, your unconditional trust, and for believing in all my choices. You have always taught my brother and I that there is nothing to be ashamed in loving and pursuing knowledge and all that we love. I hope (and think) you are reasonably proud of us. You have made us who we are today.

To my husband: thank you for putting up with my pursuits, tribulations and tempests, and for believing in my aspirations. Years ago, when that someone told me I should leave research and pursuit of knowledge to others, due to his misguided perception that I would be a failure because of my ethnicity and gender, you were the only one who stood up for me immediately.

To my daughter: thank you for all the tears, tantrums and sleepless nights, which helped me keep all this 'research stuff' in perspective.

To fellow travellers on the graduate school journey: we've pretty much kept each other sane on this insane journey by going bonkers on each other once in a while – well, no one's really tumbled off a precipice yet, I think. Here's to all of us: my dear brother Mook Tzeng, Sara, Gan, Chong Chai, Suhaila, Nur Hana, Nur Hussein.

To all the detractors, doubters, nay-sayers, nazgûl, dementors: Friedrich Nietzsche said 'That which does not kill us makes us stronger'. So here, I acknowledge your part in making me who I am now, at the end of my Ph.D. journey.

Thank you all.

To my beloved parents, Lim Yoo Kuang and Gan Choon.

ABSTRACT

Since compiling multilingual lexicons manually from scratch is a time-consuming and labour-intensive undertaking, there have been many efforts to create them via automatic means. Most of these attempts require as input lexical resources with rich content (e.g. semantic networks, domain codes, semantic categories) or large corpora. Such material is often unavailable and difficult to construct for under-resourced languages.

The objective of this research is therefore to propose a flexible framework for constructing multilingual lexicons using low-cost input and means, such that under-resourced languages can be rapidly connected to richer, more dominant languages. The main research contributions are: i) A multilingual lexicon design based on a 'shallow' model of translational equivalence. ii) A multilingual lexicon construction methodologythat requires only simple bilingual dictionaries as input, thereby alleviating the problem of resource scarcity. iii) A method for extracting translation context knowledge from a bilingual comparable corpus using latent semantic indexing (LSI). iv) A flexible annotation schema, SSTC+Lexicon (SSTC+L), for aligning lexicon entries to their occurrences in texts.

A prototype multilingual lexicon, Lexicon+TX, containing six member languages i.e. English, Chinese, Malay, French, Thai and Iban (the last of which is an under-resourced language) has been constructed using only simple dictionaries, most of which are freely available for research or under open-source licences. An accompanying context-dependent lexical lookup module has also been implemented using English and Malay Wikipedia articles as training data. The lookup module works on all Lexicon+TX member languages, including Iban.

From the evaluation, the modified OTIC filtering mechanism was found to achieve best F_1 scores of 0.725 and 0.660 for 500 Malay–Chinese translation pairings and 500 Iban–Malay translation pairings respectively. 91.2 % of 500 random multilingual entries from Lexicon+TX require minimal or no human correction. Human

volunteers who evaluated translation pairings (against which results of the modified OTIC procedure were later checked) were able to work through the data quickly, with many of them finishing 500 pairs within 2–4 hours. Meanwhile, the trained context-dependent lexical lookup module was tested on 80 English, Malay, Chinese and Iban sentences containing ambiguous words. The lookup module had a precision score of 0.650 (compared to 0.550 for baseline strategy of always selecting the most frequent translation), and a mean reciprocal rank score of 0.810 (compared to 0.771 for baseline).

The results have shown that by using simple input data and minimum human linguistics expertise, it is possible to connect under-resourced languages to more dominant, richer-resourced languages via a multilingual lexicon with highly satisfactory results in a relatively short time. This paves the important first step for developing more NLP resources and processing tools for these under-resourced languages, thus helping more communities gain access to information that may previously have been unintelligible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION AND MOTIVATION

1.1 Multilingualism and Content Access

The Internet has broken down geographical barriers to information access, where users from any location can retrieve information hosted remotely. However, this information may not necessarily be in a language that the user understands. English accounts for only 55.1 % of all website contents (W3Techs, 2012), but the content providers (or volunteers) are not always prepared (or able) to translate the contents to other languages, especially the less frequent ones.

Machine translation (MT) systems are computer programs that automatically translate natural language text from a source language (SL) to a target language (TL). MT is difficult not only because each language differs from the next (even those from the same family) both structurally and lexically, but also because natural language is itself inherently ambiguous — again, both structurally and lexically — and always evolving. As such, MT has received much bad press due to unrealistic public expectations that MT systems should produce publishable-quality, no-further-improvements-required translations at the press of a button.

The real value of current MT technology is only apparent when its usage context is viewed correctly. Hovy (1999) and Hutchins (1999) identified three usage scenarios of MT where human end-users are concerned:

Dissemination Producing a translation 'draft' to be manually post-edited to publishable quality.

Assimilation 'Gisting' or multilingual content access, i.e. aiding users to find out essential contents of a document. Lower quality is expected and acceptable.

Interchange/Communication Immediate translation to convey basic contents of messages in multi-turn dialogue, such as telephone conversations and chats.

Hutchins (1999) further listed **information access** as a usage context, where MT is integrated into other computer systems.

A translated text may satisfy assimilation and content-access needs if it contains fairly accurate translated words, even if the output is not syntactically well-formed. Take, for example, the following Welsh input text and its translation output by an online Welsh–English MT system at `http://www.cymraeg.org.uk` (Forcada, 2009):

Input Cafodd gyrrwr a fethodd brawf anadl cyn ymosod ar blismon a gyrru i ffwrdd ar gyflymder o 100 m.y.a. ei garcharu am 27 mis.

Output Driver got and failed *brawf breath before attack on *blismon and drive to a way on a speed of 100 *m.the.and. imprison him for 27 months.

Even though the English translation contains errors, a human reader is still able to gauge the rough meaning of the input Welsh text, relying on the output of the lexical lookup module of the MT system, which uses a bilingual or multilingual lexicon.

In the case of polysemous words (words with multiple meanings) in the input text, the lookup module should be able to select (or prefer) a translation word that best reflects its meaning based on the context. The same is true for information access purposes, particularly cross-lingual information retrieval (IR) applications. A user who specifies search keywords in language L_1 would be able to get results in other languages $L_2, \ldots, L_n$ if the keywords are translated via an embedded MT module or looked up from a multilingual lexicon. Here, the keywords must also be translated correctly so that relevant cross-lingual results can be retrieved.

1.2 Multilingual Lexicons for Lexical Look-up and Translation

Multilingual lexicons are important resources for computer applications and systems dealing with information and text, notably in the fields of natural language processing (NLP), cross-lingual IR and text mining. They are also indispensable reading aids to help human users understand the gist of a text written in a foreign language. Multilingual lexicons list translation equivalents of words, or rather lexical items (LIs), from the vocabularies of different languages. An LI is a unit of the vocabulary of a language such as a word, phrase or term as listed in a dictionary. It usually has a pronounceable or graphic form, fulfils a grammatical role in a sentence, and carries semantic meaning (Hartmann & Stork, 1972, p. 128).

When reading a text in a foreign language, human readers may use a multilingual lexicon to look up translation equivalents of LIs in their own native language to aid their understanding or content-scanning ('gisting') purposes. NLP applications and cross-lingual IR systems also need to access translation equivalents of LIs in different languages that reflect the meanings in the original input text. Translation selection is the process of selecting the most appropriate translation word from a set of TL words corresponding to a SL word, reflecting its sense in a particular context. This task is related to word sense disambiguation (WSD), the problem of identifying which sense of a word is used in a sentence.[1] Translation selection, or any task that involves ranking lexical lookup results depending on the context, will require a multilingual lexicon (or a bilingual one, at the very least).

Note that for the purposes of this research, all translation equivalents in a multilingual lexicon entry should reflect a common meaning or concept. Some online services providing 'multilingual look-up' (e.g. LingvoSoft[2]) display the results by separate language pairs, as shown in Figure 1.1. Instead, what we are interested in

[1] although there are recent opinions that the translation selection task may have more practical benefits than WSD (McCarthy, 2011)

[2] `http://www.lingvozone.com/lingvosoft-online-english-multilanguage-dictionary/`

Figure 1.1: LingvoSoft multilingual look-up results are displayed by separate language pairs, without sorting into sets of common meanings.

English	Chinese	Malay	French
factory	工厂	loji	fabrique
plant		kilang	manufacture
			usine

English	Chinese	Malay	French
plant	植物	tumbuhan	végétal
vegetation		tanaman	végétation
		tumbuh-tumbuhan	

Figure 1.2: Multilingual lexicon entries, with translation equivalents for a common meaning grouped together.

are sense-distinguished entries in the form shown in Figure 1.2, in which translation equivalents for a common meaning are grouped together.

However, since compiling multilingual lexicons manually from scratch is a time-consuming and labour-intensive undertaking, it would be much more feasible

to devise designs and methodologies for creating them automatically from existing resources.

1.3 The Case for Under-Resourced Languages

There have been many multilingual lexicon construction projects (Vossen, 1997; Boitet et al., 2002; Cardeñosa, Gelbukh, & Tovar, 2005; Sammer & Soderland, 2007; Pease, Fellbaum, & Vossen, 2008; Mausam et al., 2009). Most of these attempts require input lexical resources with rich content fields or large corpora. Unfortunately, not all languages have equal amounts of digital resources for developing language technologies.

Berment (2004) categorised human languages into three categories, based on their digital 'readiness' or presence in cyberspace and software tools:

- **τ- or 'tau'-languages:** totally-resourced languages, from French *très bien dotées*,
- **μ- or 'mu'-languages:** medium-resourced languages, from French *moyennement dotées*, and
- **π- or 'pi'-languages:** under-resourced languages, from French *peu dotées*.

In the NLP community, the terms *π-languages*, *less-equipped languages* and *under-resourced languages* are now commonly used to refer to languages with little or no computerised resources for NLP development (Boitet, 2007).

Some languages — like English, French, German and Japanese — have very rich resources, with many language processing tools and resources available, such as lexicons with semantic links, parser tools, and full-fledged MT and text mining systems. Other medium- or under-resourced languages, such as Malay, Swahili, Burmese and Iban, however, may not have as many resources (nor as rich). It is therefore even more important that these languages should be connected to the richer and more dominant

languages via a multilingual lexicon, such that communities speaking these languages may have easier access to information written in the more dominant languages. New bilingual dictionaries between the under-resourced languages and other languages can also then be extracted from the multilingual lexicon for more efficient lookup or processing. Such work is also important for language preservation purposes, especially for endangered languages (Rymer, 2012, see also the Endangered Languages Project at `http://www.endangeredlanguages.com/` and the Enduring Voices Project at `http://travel.nationalgeographic.com/travel/enduring-voices/`).

To counter this shortness of existing resources, the Wiktionary (`http://www.wiktionary.org/`) project takes a crowd-sourcing approach, in which volunteers contribute translation equivalents in various languages over the Internet. While there are huge amount of entries for dominant and rich-resourced languages (419 509 LI entries for English; 213 203 for French; 236 026 for Spanish), the coverage is still poor for medium- and under-resourced languages (6990 LI entries for Vietnamese; 3256 for Arabic; 729 for Afrikaans; 418 entries for Malay) (Wiktionary, 2012). Once an entry exists in Wiktionary, though, the number of its translation equivalents is likely to increase very quickly.

One approach to building multilingual lexicons with under-resourced languages is to first develop the pre-requisite lexical resources and corpora for the under-resourced languages. The multilingual lexicon construction methodologies from the projects cited earlier are then applied. However, this process would likely take a long time. Such an approach would be very expensive from the point of view of human expertise, efforts, time and data-richness. It may well be feasible to look for other means for constructing a multilingual lexicon, preferably using low-cost methods (with respect to expertise, efforts time and data-richness), so that they are applicable to under-resourced languages. Understandably, the use of a rapid method and simple input data may well entail that the accuracy and coverage of the automatically-generated multilingual lexicon could be compromised. Nevertheless, the decision to adopt this course may be justified by the principle of ‘satisficing’ (= ‘satisfy’ + ‘suffice’), i.e.

> ‘to select the first alternative that is “good enough”, because the costs in time and effort are too great to optimize’ (Simon, 1947),

especially for under-resourced languages.

1.4 Research Overview

This thesis proposes a framework for constructing multilingual lexicons. It concerns the design of multilingual lexicons and their data acquisition, as well as their application in practical settings, with particular attention to the constraints of under-resourced languages. This section gives an overview of the research reported in this thesis, by presenting the problem statement, research objectives, research questions, and research contributions of the proposed framework.

1.4.1 Problem Statement

The following problem statement summarises the problem to be addressed:

> *How can a multilingual lexicon be designed and constructed rapidly using low-cost means, especially with the resource constraints faced by under-resourced languages?*

1.4.2 Research Questions

The research questions to be addressed are listed below:

- How should a multilingual lexicon be designed to handle certain multilingual linguistic phenomena?
- How should a multilingual lexicon be structured to allow lay-persons to help verify its contents?
- How can a multilingual lexicon be compiled from simple data, so that it is viable for under-resourced languages?

1.4.3 Research Objectives

The research objectives are summarised below:

RO1. To design the architecture of a multilingual lexicon that facilitates rapid construction.

RO2. To design an algorithm and work flow for constructing a multilingual lexicon using low cost methods, suitable for under-resourced language.

RO3. To demonstrate potential applications of the multilingual lexicon via a content-dependent lexical lookup module.

1.4.4 Proposed Framework

A summarised overview of the proposed framework is shown in Figure 1.3.

In this research, a flexible framework for constructing multilingual lexicons using low-cost means is proposed. The framework includes guidelines for the multilingual lexicon design, as well as the lexicon data acquisition process.

The multilingual lexicon is designed to accommodate linguistic phenomena like diversification, lexical gaps and multi-word expressions (MWEs). It is structured such that human evaluators with minimum linguistic expertise may participate in the project. Data acquisition requires only simple bilingual translation lists, which are easily available for little or no charge, or may be compiled with relative ease.

Once constructed, the multilingual lexicon may be a source from which new bilingual dictionaries can be extracted, especially for less common language pairs. The constructed multilingual lexicon also has applications as an intelligent reading aid by providing context-dependent lexical lookup features. This is facilitated by extracting translation context knowledge from a bilingual comparable corpus of medium-resourced language pairs. The lexical lookup tool is applicable to input texts written in any

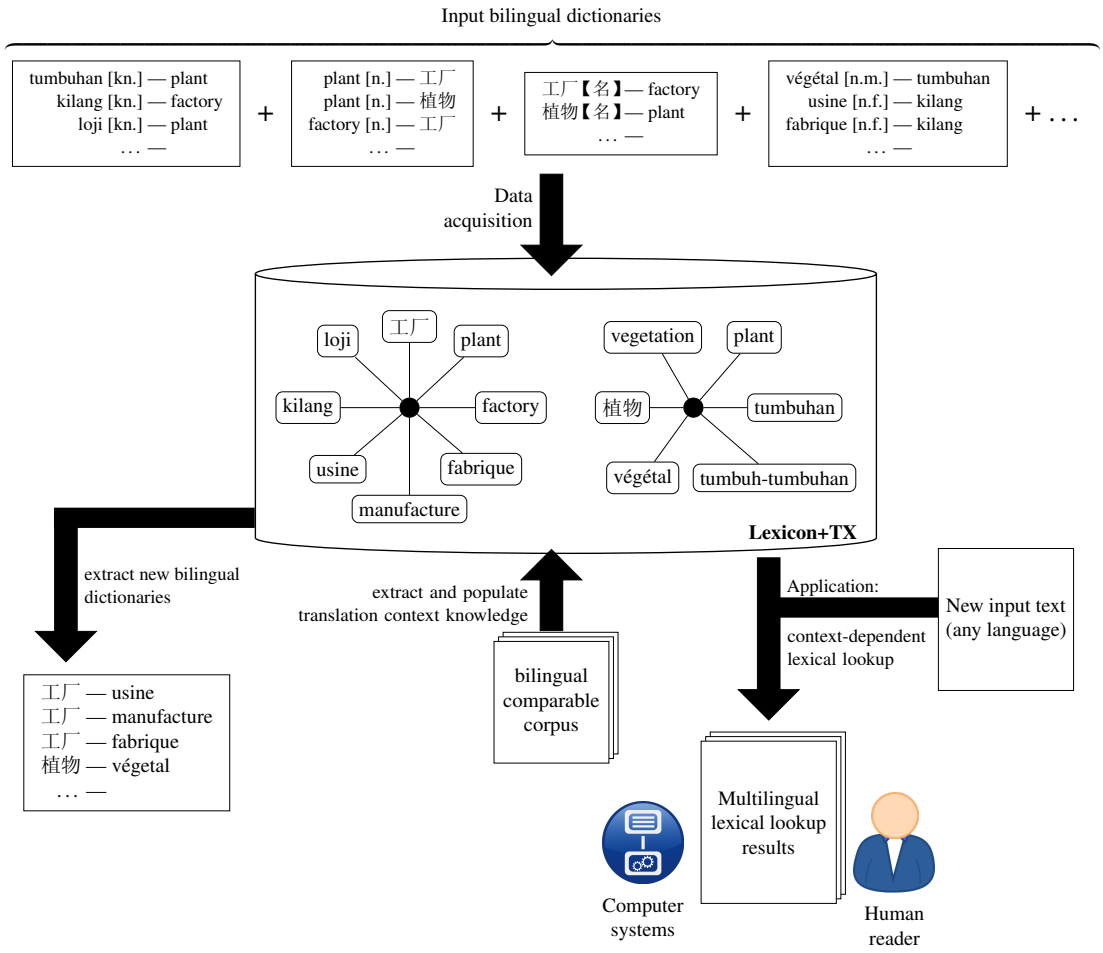


Figure 1.3: Overview of Proposed Framework

language in the multilingual lexicon. The lexical lookup results, packaged in a suitable computer-tractable schema, may also be consumed by other NLP systems.

The proposed framework has two main characteristics:

- Flexibility:
  - Design-wise, the multilingual lexicon has a simple structure, requiring only translation equivalents to be listed. This allows a usable multilingual lexicon to be rapidly constructed. At the same time, the lexicon design also has mechanisms to accommodate different levels and types

of information, such as morphological, syntactic, and richer semantics, which can be added to the lexicon at a later stage.

  - Data-wise, there is very little restriction on the type of input data required to populate the multilingual lexicon. The proposed lexicon data acquisition methodology requires only simple bilingual translation lists to populate the multilingual lexicon: this is especially applicable to under-resourced languages. Rich-resourced languages can always be added to the multilingual lexicon using more sophisticated approaches and lexical resources available if so preferred (see section 2.4).

- Low cost:
  - Data-wise, the proposed lexicon data acquisition methodology requires only simple bilingual translation lists, as mentioned earlier. These are more readily available free-of-charge, or more easily compilable from scratch, compared to other lexical resources like thesauri, wordnets or large-scale corpora. This is an important advantage for under-resourced languages.
  - Expertise-wise, no linguistics expertise is expected of human volunteers to verify or edit the multilingual lexicon contents. Again, this is important for under-resourced languages, where it is already difficult to source for volunteers who speak the languages. It would be impractical to expect them to be knowledgeable about the academic, linguistic aspects of the languages, too.

1.4.5 Research Contributions

The main contribution arising from this research is a framework for rapidly constructing multilingual lexicons from simple inputs. The lexicon can then be used for generating new bilingual dictionaries, and for context-dependent lexical lookup. Here are the contributions in more detail, with the research objective addressed by each contribution highlighted in parentheses:

RC1. A multilingual lexicon design based on a 'shallow' model of translational equivalence, so that volunteers without specialised linguistics background may be recruited to help improve and validate the multilingual lexicon (RO1; section 3.1).

RC2. A multilingual lexicon construction methodology that takes simple bilingual dictionaries as input, thereby alleviating the problem of resource scarcity (RO2; section 3.2).

RC3. A method for extracting translation context knowledge from a bilingual comparable corpus, such that the multilingual lexicon may provide context-dependent lexical lookup functions for other languages, including under-resourced ones (RO3; sections 4.1 and 4.2).

RC4. A flexible annotation schema for aligning canonical lexicon entries with their occurrences in a text (and its translation), capable of handling discontiguous forms of MWEs and lexical gaps (RO3; section 4.3).

The lexicon has mechanisms for handling lexicographic issues such as polysemy, lexical gaps and MWEs (section 2.2). A prototype multilingual lexicon, Lexicon+TX, containing English, Malay, Chinese, French, Thai and Iban, with an accompanying context-dependent lexical lookup module (RO2, RO3; Chapter 5). As far as the author is aware, this is the first time that Iban, an under-resourced Bornean language, is connected to French, Thai and Chinese.

1.4.6 Thesis Organisation

The rest of this thesis is organised as follows. Research objectives and contributions addressed by each chapter is given in parentheses.

Chapter 2 describes issues related to the design of multilingual lexicons from computational and linguistic aspects, with a review of the architectural design of recent multilingual lexicon projects. (RO1)

Chapter 3 first presents the design of a 'shallow' multilingual lexicon, based on the

requirements concluded at the end of Chapter 2. It then proposes a methodology for constructing a multilingual lexicon using easy-to-acquire bilingual dictionary data, which is especially suitable for under-resourced languages. (RO1, RO2; RC1, RC2)

Chapter 4 shows how a context-dependent lexical lookup module can be built, using translation context knowledge extracted from a bilingual comparable corpus. The lookup module is not restricted to the languages of the corpus only: it can also be applied on input texts of any member languages of the multilingual lexicon. This chapter also proposes a new annotation schema, suitable for 'packaging' the lexical lookup results for further use in other NLP systems. (RO3; RC3, RC4)

Chapter 5 presents and discusses the implementation results, which yielded Lexicon+TX a prototype multilingual lexicon containing English, Chinese, Malay, French, Thai and Iban, as well as an accompanying context-dependent lexical lookup module. (RO2, RO3)

Chapter 6 sums up the work reported in this thesis, before closing with a brief rundown of possible future extensions and improvements.

1.5 Summary and Conclusion

Multilingual lexicons are important resources for human readers and NLP systems, yet their creation involving under-resourced languages are often hindered by both the lack of resources and the limited pool of human volunteers who are also skilled in linguistics aspects.

This chapter has laid out the objectives of the research to be undertaken, i.e. to propose a flexible framework, encompassing a design and methodology, of how such multilingual lexicons can be constructed using low cost means. An overview of the research carried out, as well as contributions arising from it, has also been sketched and will be dealt with in detail in the following chapters.

CHAPTER 2

RESEARCH BACKGROUND AND LITERATURE REVIEW

Multilingual lexicons are lexical databases that list lexical items (LIs) from different languages that are translational equivalents conveying a common meaning, and are important resources for NLP applications and human users alike. However, the design of multilingual lexicons must take some linguistic and practical issues into consideration.

This chapter will describe issues related to the design of multilingual lexicons from computational and linguistic aspects, with a review of the architectural design of recent multilingual lexicon projects. Ease of data acquisition must also be considered while planning the lexicon design, especially for under-resourced languages. From this discussion, a set of principles will be derived for designing and constructing a multilingual lexicon with minimum cost, both in terms of expertise requirement of human contributors, as well as input data resources.

It would be desirable for an electronic multilingual lexicon to support context-dependent lookup, i.e. possible translation words are listed in order of relevance to the text being read. To this end, the multilingual lexicon needs to be enriched with extra semantic information. This chapter also briefly reviews the types of training data typically used in WSD and translation selection tasks, which have similar goals, i.e. selecting the most relevant lexical meaning (and respectively lexical translation) of a word in context.

2.1 Computational Architectures of Bilingual and Multilingual Lexicons

Multilingual lexical look-up functions may be provided by a single multilingual lexicon, or by a collection of bilingual lexicons. Nevertheless, a single multilingual lexicon is easier to maintain, as the following discussion will show.

The simplest scheme of a bilingual lexicon is a "flat" list mapping language L_1 lexical items to one or more possible translations in language L_2, sometimes not even making any distinctions between different senses (Figure 2.1). Such lists are actually unidirectional, thus two such lexicons are required to provide look-ups in both directions between a language pair.

POS	English	Malay
N	bank	bank; tabung; tebing; beting; tambak; permatang
V	bank	menyimpan wang; menimbun; terbang mengereng

Figure 2.1: Simple English–Malay bilingual lexicon without sense distinctions

While this bilingual scheme is easy to maintain for a single language pair (requiring two uni-directional bilingual lexicons), the number of inter-lexicon links to maintain grows quickly to $O(n^2)$ in a system involving n languages, as shown in Figure 2.2. Adding a new language requires $O(n)$ new links to link the new language to each of the already existing languages.

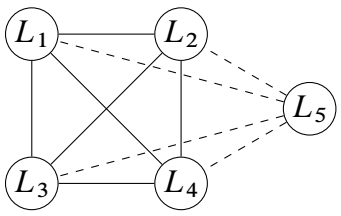


Figure 2.2: Adding a new language in bilingual lexicons setting

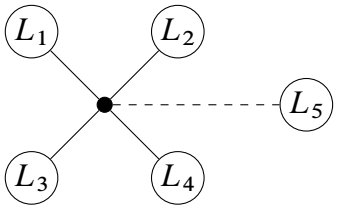


Figure 2.3: Adding a new language in multilingual lexicon setting

On the other hand, one possible scheme for a *multilingual* lexicon, shown in Figure 2.3, requires the maintenance of only $O(n)$ links to a *pivot axis*. Entries for a new language would only need to be linked to the axis, and translation equivalence between the new language and the existing ones would be established via the axis. This is especially beneficial to introduce more language pairs, especially from and to under-resourced languages, into a MT system. At a glance, the axis is similar to an interlingua.

Despite the various objections to the existence of a universal language (mainly from researchers in linguistics and psychology, e.g. Hurford, 1990; Christiansen & Chater, 2008), such a mechanism presents a feasible solution if treated as a computational mechanism rather than for explaining fundamental linguistic issues (Zajac, 1996).

Due to linguistic phenomena and differences across languages, however, merging bilingual lexicons into a single resource is non-trivial. A lexical resource that aims at providing multilingual translation equivalents must be well-designed to address these issues. Some related problems will be described in the next section.

2.2 Issues in Multilingual Lexicography

Creating a multilingual lexicon can have certain difficulties, due to various linguistics and multilingual issues. Hutchins (2007) named two main issues related to bilingual lexical differences, while the treatment of MWEs in each language is also important. These three issues are described briefly below.

In the following discussion, we denote a lexical item (LI) (which may comprise multiple words) with guillemets, e.g. «tree», «science fiction»; while gloss or explanatory phrases are marked with single quotes, e.g. 'chop finely', 'young horse'. The meaning of a non-English LI is given in parentheses, e.g. Malay «buta tuli» (blindly). For LIs in languages which do not use the Latin script, the pronunciations of which are hence not immediately obvious, the phonetic transliteration may also be given in parentheses, along with the meaning in English, e.g. Chinese «蝴蝶» (*húdie*; butterfly) and Japanese «祭り» (*matsuri*; festival).

Also, for brevity's sake, we may sometimes annotate the language of an LI by its 3-letter ISO 639-3 code (see Appendix A) in lower case letters, e.g. «membuta tuli»$_{\mathtt{msa}}$; or the part of speech (POS) (see Appendix B) in upper case letters, e.g. «minute»$_{\mathrm{N}}$.

2.2.1 Lexical Ambiguity

The first issue concerns bilingual lexical ambiguity, or the existence of multiple equivalents in the TL. This could be due to ambiguity in the SL, for example English «glass» ↦ «gelas» (a receptable for fluids) and «glass» ↦ «kaca» (a clear, hard but brittle material made from sand) in Malay.

There are two types of lexical ambiguity: homonymy and polysemy. Homonyms are LIs having the same spelling but different meanings and origins, e.g. «bat» (a nocturnal, flying mammal) and «bat» (a club for hitting the ball in sports). On the other hand, a polyseme is an LIs with different, but related senses, e.g. «man» can mean the human species; or an adult male of the human species. Conventional paper dictionaries usually enter senses of polysemes under the same headword entry, while homonyms are placed in separate headword entries. However, there is usually no difference in treatment of homonyms and polysemes in electronic lexicons or lexical databases.

In other cases, a single meaning of an LI may have translations in the TL that are more specific, such as the Spanish «dedo» ↦ «finger» and «dedo» ↦ «toe» in English as it does not distinguish between appendages on the hand or foot. This phenomenon is known as diversification, and as neutrification in the opposite direction.

2.2.2 Lexical Gaps

The second issue mentioned by Hutchins (2007) is that of lexical gaps, when a concept is not lexicalised in a particular language. This is sometimes due to cultural differences: indeed many words pertaining to culinary or clothing apparels in a specific culture do not have equivalents in other languages, like «cottage», «vodka», «batik», «粽子» (*zòngzi*; Chinese glutinous rice dumpling), «きもの» (*kimono*). Romanised or transliterated forms of such LIs are usually used in the translated text, and often find their way into the TL's vocabulary. In most other cases, a gloss-like expression or a paraphrase is used to translate the SL lexical item. For example, the English noun «fortnight» is translated as a noun phrase 'dua minggu' (two weeks) in Malay. Where

the TL equivalents are not LIs, an electronic lexicon or lexical database may store a gloss text instead, or devise a more comprehensive system for representing lexical gaps.

Occasionally an SL lexical item and one from the TL may be very near synonyms, yet have subtle underlying differences, resulting in near-miss lexical gaps in both languages. Consider Chinese «跳飞机» (*tiào fēijī*) and Indonesian «merantau»: while both describe a situation where a person works in a foreign country without intentions to reside permanently, the former has a negative connotation while the latter does not. Such subtle differences often confuse the TL-speaking user and annoy the SL-speaker. This is perhaps unavoidable, as a human professional translator may have no better strategy but to offer the same translation.

2.2.3 Multiple-word Expressions

A third issue is when either of the mapped expressions (both SL and TL) contain multiple words, they may not necessarily be contiguous. Multi-word expressions (MWEs) should be considered LIs in their own right as they have distinct meanings from their constituent words. For example, if a lexicon does not list the Malay MWE «menjolok mata» (unsightly, gaudy, provocative), a human user or a MT system would interpret the phrase by its constituent words («menjolok» and «mata») by their literal meanings, i.e. 'poke eye'.

MWEs, which include idioms, compound nominals, verb-particle constructions and light verbs, exhibit a wide range of syntactic flexibility (see Sag, Baldwin, Bond, Copestake, & Flickinger, 2002, for an interesting classification). Some MWEs are rigid and frozen, e.g. the idiom «kick the bucket» cannot have variations '*the bucket was kicked' nor '*kick the little bucket'. Other MWEs allow some amount of flexibility, e.g. «throw *someone* to the lions» allows any noun phrase to replace *someone*; and the construction 'earn a meagre living' from «earn a living» is valid. There are also MWEs that are highly flexible. For example, «spill the beans» can be made into a passive voice construction 'the beans are spilt'.

In a conventional bilingual dictionary, translation pairs involving MWEs are listed using human-recognisable place-holders. For example:

- English: earn a living
  Malay: mencari nafkhah
- English: throw *somebody* to the lions
  Chinese: 丢下某人不管
- English: get *one*'s knife into *somebody*
  Malay: berniat jahat terhadap *seseorang*

Nevertheless, such linear sequences may be inadequate in a MT setting, especially when the syntactic tree structures need to be manipulated to translate syntactically flexible MWEs correctly.

2.3 Review of Multilingual Lexicon Designs

A selection of different multilingual lexicon design approaches will be reviewed and discussed in this section. Possible approaches may roughly be classified as 'shallow' or 'deep', depending on whether a formal interlingua system is used for describing underlying lexical semantics.

2.3.1 'Shallow' Multilingual Lexicons

'Shallow' multilingual lexicons typically use a language-independent axis or pivot mechanism for linking LIs from different languages conveying a meaning or concept. Some lexicons allow links among the axes and pivots, as well as additional syntactico-semantic information to be associated to the axes or indiviual LIs.

2.3.1 (a) Wordnet-based Projects

Princeton WordNet (Miller, Beckwith, Fellbaum, Gross, & Miller, 1990) is a lexical database for the English language. It organises LIs conveying the same meaning

into synonym sets or *synsets*, e.g. (car, auto, automobile, machine, motorcar). Various types of links connect the synsets to indicate their semantic relationships to each other, e.g. hypernym (is-a), holonym (part-of), entailment and others, thus forming a lexical semantic network.

The Princeton WordNet is available without charge. Its semantic network is a valuable resource for NLP work, and the many other lexical resource projects that build upon or link to it has made the English language a very rich-resourced one. Some examples include syntactic–semantic relations for verbs in VerbNet (Shi & Mihalcea, 2005; Kipper, Korhonen, Ryant, & Palmer, 2008); case semantics and semantic roles in FrameNet (Fontenelle, 2003; Shi & Mihalcea, 2005); subject field labels (Magnini & Cavaglià, 2000); and ontology class labels from the Suggested Upper Merged Ontology (SUMO) (Niles & Pease, 2001, 2003). These two factors of easy availability and data richness have led to WordNet's being widely used in many NLP applications. Many wordnet systems have also been developed for other languages and aligned to the original Princeton WordNet, in order to leverage the rich data available. This has given rise to several multilingual wordnet-based lexical databases.

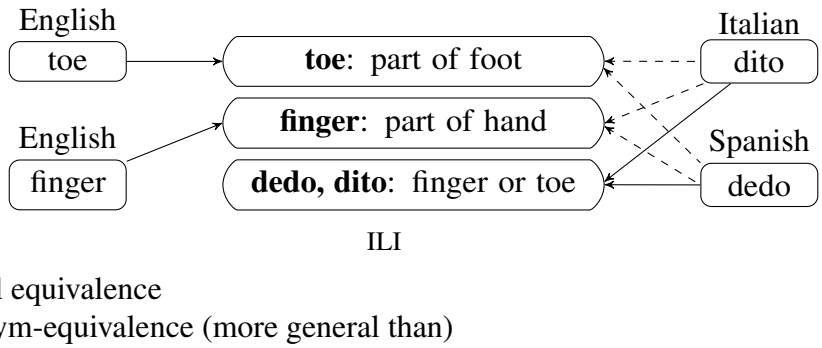


Figure 2.4: EuroWordNet's Unstructured ILI (adapted from Vossen, 1997)

EuroWordNet (Vossen, 1997, 2004) uses a language-independent Inter-Lingual Index (ILI) to link synonymous lexical senses in different languages, using English for convenient naming of the ILI records. Recall the earlier example on Spanish «dedo» (and also Italian «dito») having more specific translations in English «toe» and

«finger». English «toe» and «finger» are linked to the respective ILI records using normal equivalence relations, while «dito» and «dedo» are linked as equivalence to a separate ILI record. «dito» and «dedo» are further linked to **toe** and **finger** ILI records using hyponym-equivalence (more general than) relations. Note that the ILI records are not structured in any way. Such use of hyponym-equivalence and respectively hypernym-equivalence (more specific than) can handle diversification and neutrification. However, as shown in Figure 2.4, EuroWordNet's ILI design would cause an explosion of links to maintain when records similar to **dedo, dito** are created. In addition, we are not aware of any provisions for non-contiguous MWEs in EuroWordNet.

Learning from the experiences from the EuroWordNet project, later wordnet-based multilingual lexicon projects, including BalkaNet (Tufiş, Cristea, & Stamou, 2004), HowNet (Dong & Dong, 2006), the Global WordNet Grid (Pease et al., 2008), Universal Multilingual WordNet (de Melo & Weikum, 2009) and Open Multilingual WordNet (Bond & Paik, 2012) used a *structured* ILI, as opposed to the unstructured one in EuroWordNet. The English Princeton WordNet hypernymy (is-a relation) hierarchy is taken as the initial ILI. For lexical gaps in English or diversifications, new ILI records are inserted at appropriate places in the ILI hyernymy hierarchy. The non-English LIs are then connected to the new ILI record *only*, as its hypernymy relations to the English LIs are already captured via the ILI hypernymy hiearchy.

It may be argued that these multilingual wordnets should be categorised as a 'deep' lexicon, since the structured ILI forms a semantic network. However, this semantic network does not systematically decompose lexical meanings into finer semantic elements (c.f. lexicons reviewed in 2.3.2). Multilingual wordnets are therefore considered as forming a'shallow' multilingual lexicon, with perhaps some leaning towards a 'deep' approach.

Since such a multilingual wordnet scheme uses the English Princeton WordNet as its main 'hub', they may suffer from a frequent critique against the Princeton

WordNet: its sense distinctions are often overly fine. For example, the following three senses of «school»$_N$ are considered distinct in Princeton WordNet:

- an educational institution; “the school was founded in 1900”
- a building where young people receive education; “the school was built in 1932”; “he walked to school every morning”
- an educational institution’s faculty and students; “the school keeps parents informed”; “the whole school turned out for the game”

Variances in syntactic valency or transitivity are also considered as distinct senses, e.g. «break»$_V$:

- become separated into pieces or fragments; “The figurine broke”
- destroy the integrity of; usually by force; cause to separate into pieces or fragments; “He broke the glass plate”; “She broke the match”

Such fine sense distinctions may cause human lexicographers and evaluators working with the wordnets much confusion when contributing translation equivalents. Depending on the goal, some NLP applications may even suffer from such fine sense granularity, which entails a higher number of senses.

2.3.1 (b) Papillon

The Papillon multilingual dictionary project (Boitet et al., 2002) uses a volume of interlingual *axies* to link translation equivalents from different languages. As Papillon’s axies may have relations among themselves, contrary to EuroWordNet’s ILI records, the problem of ‘link explosion’ can be avoided. This is illustrated in Figure 2.5, where the ‘grain’ sense of «rice» and «riz» are linked to an axie that is further linked to two other axies. «米» and «beras» (respectively «御飯» and «nasi») can then be specified as equivalent to each other, and is more specific than «rice» and «riz». Jalabert and Lafourcade (2002) proposed a method for generating glosses for lexical gaps in

Papillon's framework. However, we are unaware of any provisions for non-contiguous MWE equivalents in Papillon.

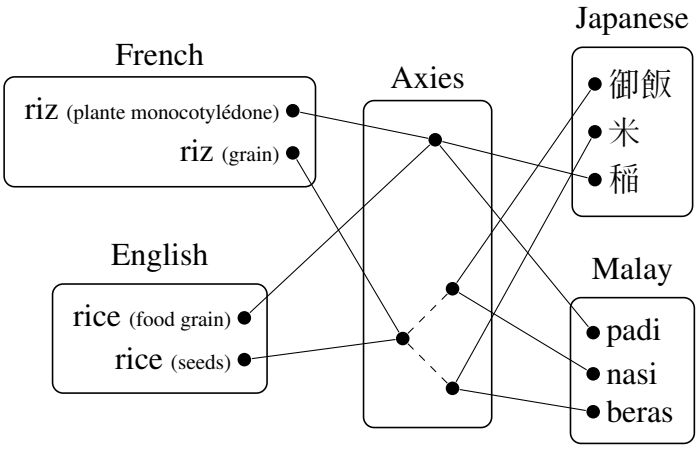


Figure 2.5: Papillon's interlingual axies (adapted from Boitet et al., 2002)

2.3.1 (c) PIVAX

PIVAX (Nguyen et al., 2007) is a lexical database for the creation, maintenance and management of lexical resources of heterogeneous MT systems. As an acknowledgement of the different organisation principles and proprietary information of various MT systems, only the most basic language-specific lexical information is made compulsory in PIVAX. The organisation is similar to that of Papillon's (section 2.3.1 (b)). Interlingual *axie* pivots are connected to language-specific synonymous *axemes*, which are in turn connected to synonymous *lexies* in lexicons of respective MT systems.

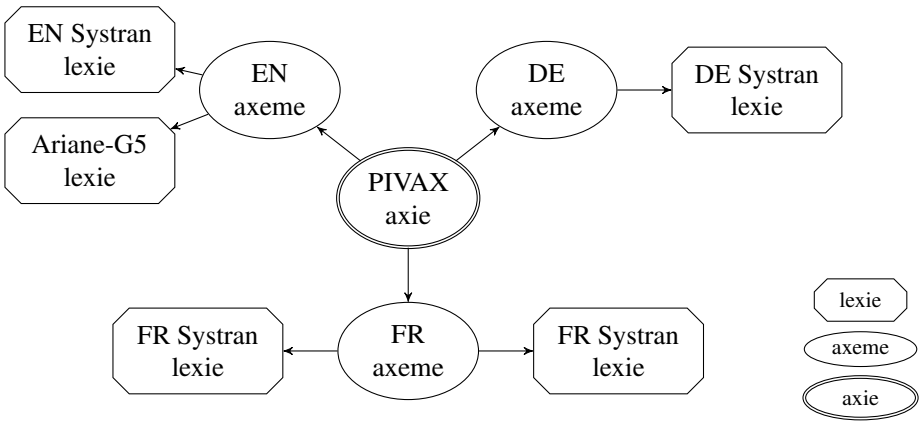


Figure 2.6: Organisation of volumes in PIVAX (adapted from Nguyen et al., 2007)

2.3.1 (d) Lexical Markup Framework

The Lexical Markup Framework (LMF) (ISO24613, 2008; Francopoulo et al., 2009) was introduced as an ISO standard for lexical resource management and provides mechanisms for various aspects of lexicography related to NLP, including morphology, syntax, semantics and multilingualism. The `Sense Axis` in LMF (Figure 2.7) is similar in nature to Papillon's axie. LMF borrowed the idea and term 'axie' from Papillon but changed it to 'axis' to respect English orthography (Francopoulo et al., 2009).

LMF also has a `Transfer Axis` for specifying multilingual translation equivalents with selectional restriction tests (Figure 2.8). For example, English «develop» is translated to Italian «construire» and Spanish «construir» if the second syntactic argument is a building; otherwise it is translated to the more general Spanish «desarrollar». On the other hand, there are rather comprehensive mechanisms in LMF for specifying MWEs (Figure 2.9) and their possibly discontiguous and decomposable constructions, i.e. via its phrase tree structure. Fixed and variable elements can also be specified.

Summarily speaking, the LMF, though lacking any real data, provides a comprehensive framework for defining computational lexicons, covering almost all aspects of linguistics. Its mechanism for handling MWEs is especially attractive.

2.3.1 (e) PanLexicon

PanLexicon (Sammer & Soderland, 2007) organises its entries using translation sets, which are 'a multilingual extension of a WordNet synset' and contain 'one or more LIs in each k languages that all represent the same word sense' (Figure 2.10). Each LI is also accompanied by a usage illustration to indicate its meaning.

PanLexicon is constructed by mining possible translation equivalents from bilingual comparable corpora, extracting topic signatures as contextual data at the same time. This acquisition approach is fast and efficient, but cannot extract MWEs and their translation equivalents at present. Also, it could be difficult to obtain even bilingual

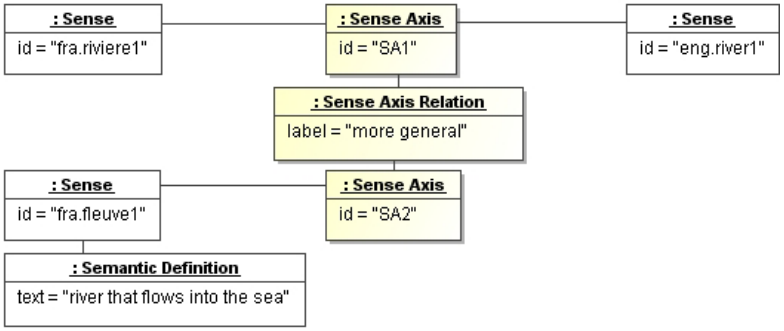


Figure 2.7: `Sense Axis` **in LMF (from ISO24613, 2008)**

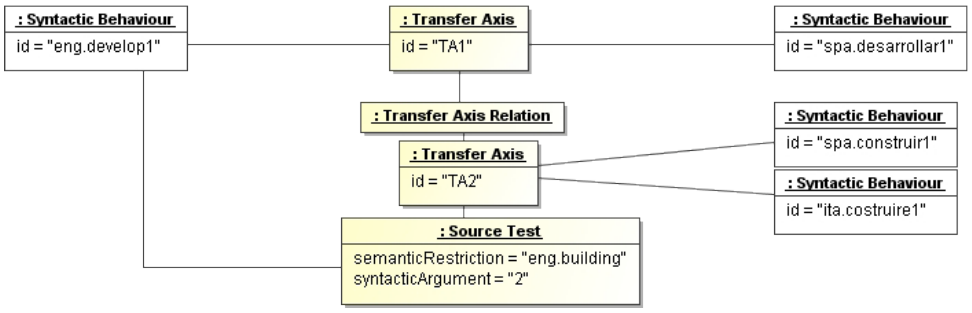


Figure 2.8: `Transfer Axis` **in LMF (from ISO24613, 2008)**

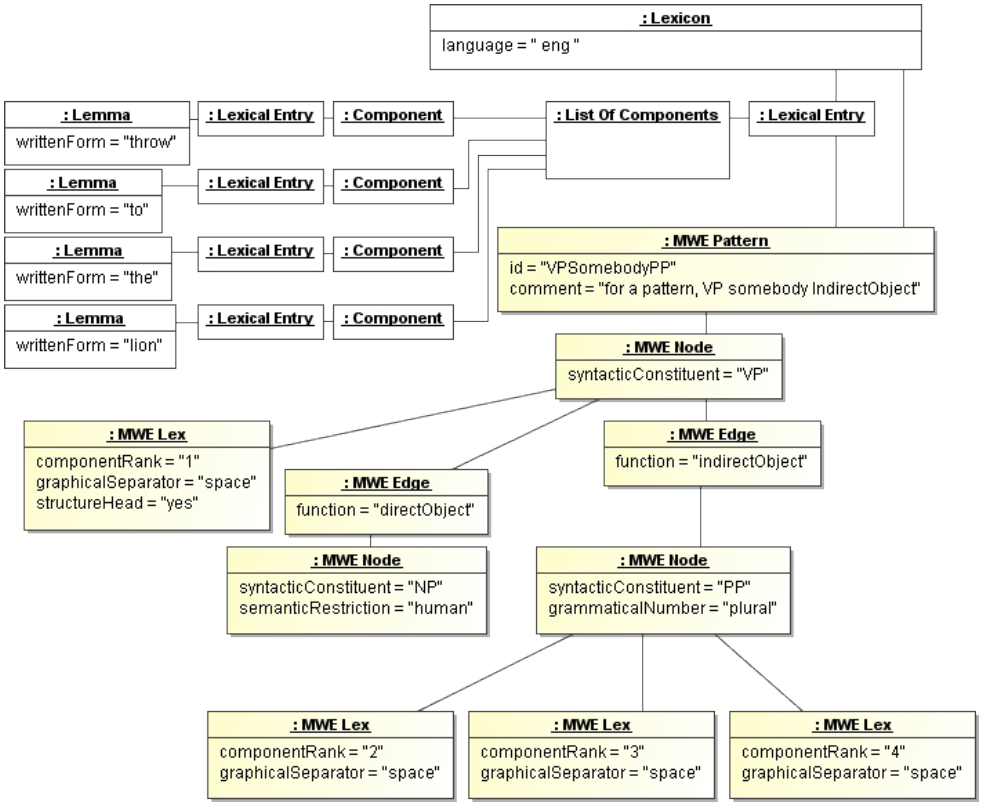


Figure 2.9: `MWE` **classes in LMF (from ISO24613, 2008)**

English		Spanish		Chinese	
plant	aluminium smelting **plant** that employs about 930 workers	**planta**	materiales nucleares de las **plantas** de energía para fabricar armas atómics	厂	工人到厂里来,就是来干活的
				厂房	该厂有8间厂房、5间仓库
factory	food warehouses, an insecticide plant and a fertilizers **factory**	**fábrica**	trabajadores de una **fábrica** privada estaban fundiendo pedazos de aluminio	工厂	生产车间作为工厂的“特区”

Figure 2.10: Example translation set from PanLexicon for the concept ‘industrial plant’ (from Sammer & Soderland, 2007)

comparable corpora for under-resourced languages, especially those spoken by minority ethnic groups.

2.3.2 ‘Deep’ Multilingual Lexicons

‘Deep’ multilingual lexicons seek to represent the semantics underlying lexical meanings with a formal interlingua system. A number of different formalisms have been proposed, some of which are reviewed here.

2.3.2 (a) SIMuLLDA

In the multilingual lexicon projects and frameworks reviewed so far, a pivot-like mechanism is used for linking translation equivalents from different languages. As such, when there is a lexical gap in a particular language L, a translation can only be generated for L by translating the *gloss text*, if available, of a LI in another language. SIMuLLDA (Janssen, 2003, 2004) takes a different approach by using a taxonomic lattice of concepts or definitional attributes as the interlingua, based on Formal Concept Analysis principles. The treatment of MWEs was not mentioned.

In the example on LIs related to *horse* in Figure 2.11, there is a lexical gap in French for English «colt». From the lattice of concepts and definitional attributes, «colt» $\equiv$ COLT $=$ FOAL $+$ **male**. There *is* a French equivalent for FOAL: «poulain». A French translation can therefore be systematically generated, i.e. «poulain mâle».

	horse	male	female	adult	young
HORSE	×				
STALLION	×	×		×	
MARE	×		×	×	
FOAL	×				×
FILLY	×		×		×
COLT	×	×			×

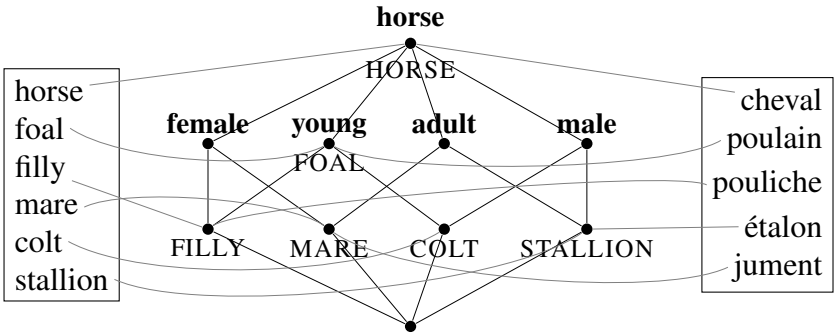


Figure 2.11: SIMuLLDA's lattice of concepts and definitional attributes based on Formal Concept Analysis (adapted from Janssen, 2003)

However, SIMuLLDA's taxonomic considerations do not always agree with lexicographic practices. Translation equivalence cannot be established among many accepted translation pairs if strict logical principles are applied, or would be problematic if it is attempted: see Janssen's (2003) elaboration on French «rivière», «fleuve» and English «river», «stream».

2.3.2 (b) Universal Networking Language

The Universal Networking Language (UNL) (UNL Center, 2004; Cardeñosa et al., 2005) is intended to be a true formal interlingua language, and its vocabulary is made up of language-independent Universal Words (UWs). Natural language expressions or sentences are 'encoded' as compound UWs or UNL hyper-graphs, whose nodes are made up of UWs. These interlingual expressions are then 'decoded' into TLs for translation. The decoder modules are delegated to development partners responsible for each TL.

A basic UW is represented by an English expression as the headword. If the English headword is ambiguous, restrictions are introduced to accompany the headword:

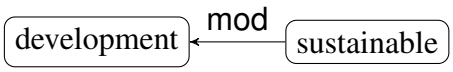
- `state(icl>country)`
- `state(icl>express(agt>thing,gol>person,obj>thing))`
- `state(...)`

and translated into respective TLs by decoder modules.

As for the handling of MWEs, the UNL project avoids multi-word headwords in UWs as much as possible. The rationale is if any free word combination can be made an UW, development partners may not have a matching UWs in their own dictionaries (Bugoslavsky, 2005). Therefore, compositional MWEs are modelled as combination of multiple UWs wherever possible. Following are some examples taken from (Bugoslavsky, 2005):

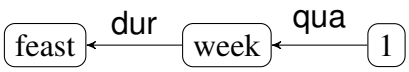
- «sustainable development»

  `mod(development,sustainable)`

- «week-long feast»

  `dur(feast,week)` `qua(week,1)`

Non-compositional MWEs such as «look for» can either be modelled as a multi-word headword:

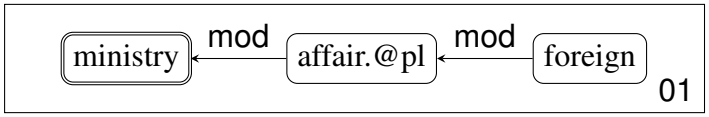
`look for(icl>do,agt>thing,obj>thing)`

or as a specific meaning of the ‘main’ word:

```
look(icl>search>do,agt>thing,obj>thing).
```

The current treatment of compositional MWEs expressing a single concept, for example «Ministry of Foreign Affairs», is to apply scoping to the hypergraph:

```
mod:01(ministry.@entry, affair.@pl)
mod:01(affair.@pl,foreign)
```



An alternative treatment mentioned in (Bugoslavsky, 2005) is to allow UWs to have internal structure:

```
mod(ministry,affair.@pl)&mod(affair.@pl,foreign)
```

Such an approach captures, in a more natural way, both the compositional nature and the single concept it expresses. However, this proposal is not yet implemented in UNL as it requires considerable modifications to the UNL specification and software.

2.3.2 (c) Lexical Knowledge Base of Near-Synonyms

The Lexical Knowledge Base of Near-Synonyms (LKB of NS) (Edmonds & Hirst, 2002; Inkpen & Hirst, 2006) pays explicit attention to multilingual near-synonyms. It uses a formal ontology to model real-world concepts, to which LIs from different languages are mapped on a coarse-grained basis to reflect the core denotational meaning. Edmonds and Hirst (2002) proposed a sub-ontology for lexical choice distinctions between near-synonyms for describing the peripheral concepts, used for distinguishing the near-synonyms in a fine-grained manner.

The example for ERROR nouns in Figure 2.12 shows that «blunder» is associated with a high level of `Blameworthiness` and `Stupidity`, as well as `Pejorative` towards

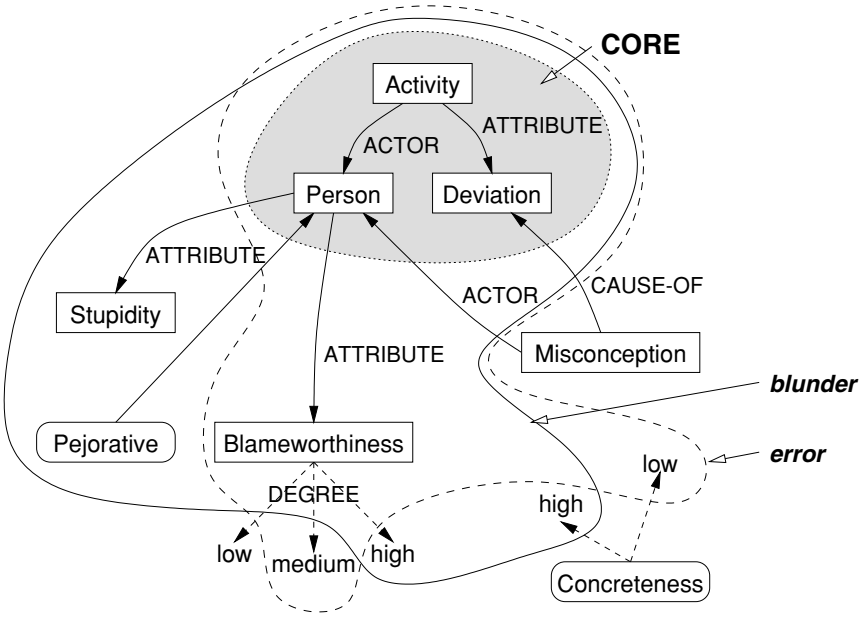


Figure 2.12: The core denotation and some peripheral concepts of cluster of ERROR nouns, i.e. «blunder» and «error» (from Edmonds & Hirst, 2002)

the actor. On the other hand, «error» is more neutral. Cross-lingual near-synonym groups can also be modeled in this two-tier approach, e.g. the coarse-grain near-synonym group for *forest* might contain «forest»$_{eng}$, «woods»$_{eng}$, «copse»$_{eng}$, «Wald»$_{deu}$ (smaller and more urban area of trees than «forest»$_{eng}$), «Gehölz»$_{deu}$ («copse»$_{eng}$ and "smaller" part of «woods»$_{eng}$).

The LKB is intended to be a full-fledged formal ontology to be used in language understanding applications, and therefore requires rather rigorous constructions. There is no mention of the treatment of MWEs.

2.3.3 Discussion

Table 2.1 briefly summarises the comparison between 'shallow' and 'deep' multilingual lexicon types.

To recap briefly, 'deep' multilingual lexicons use a formal interlingua system to represent lexical meanings and decompose semantic concepts, while 'shallow' multilingual lexicons use language-independent axes only as a convenience mechanism for

Table 2.1: Comparison of 'Shallow' and 'Deep' Multilingual Lexicons

Aspect	'Shallow' Approach	'Deep' Approach
Examples	Wordnet systems, Papillon, PIVAX, LMF, PanLexicon	SIM*u*LLDA, UNL, LKB of NS
Principle	Groups mutilingual translation equivalents declaratively with convenience pivot- or axis-like mechanisms	Proposes interlingual formalisms for representing lexical meanings and concepts
Expertise	Easier for lay-persons to edit	Requires certain linguistic and semantic expertise to edit
Applications	May be faster and practical to implement for MT systems	Suitable for language understanding systems

linking multilingual translation equivalents. As a result, 'deep' lexicons often have a more systematic and formal method for generating translations in cases of lexical gaps (c.f. SIM*u*LLDA and UNL). 'Deep' multilingual lexicons are also more suitable for language understanding systems which require rich semantic data to function.

However, this would also mean contributors must be sufficiently knowledgeable in linguistics and the underlying interlingua system to maintain, enrich and improve 'deep' multilingual lexicons effectively. Indeed, the UNL project expects volunteer contributors to have some background in descriptive linguistics, and requires them to complete an online course[1] in order to gain a Certificate of Language Engineering Aptitude, before they are allowed to contribute to UNL dictionaries. Similarly, the LKB's sub-ontology for lexical choice distinctions, while being an elegant solution framework for near-synonyms, would require human contributors to have a good understanding of the system's controlled vocabulary and structure. However, it can perhaps be approximated by usage labels in conventional dictionaries (c.f. section 3.1.2 (c)): see Janssen, Verkuyl, and Jansen (2003) for a discussion.

'Shallow' multilingual lexicons, on the other hand, let polyglot contributors simply list translation (near-)equivalents without having to deal deeply with linguistic

[1] `http://goo.gl/1eU7b`

or semantic properties. With lower requirements on source input data and on personnel expertise, 'shallow' multilingual lexicons may thus be constructed, checked and deployed in some NLP systems in a shorter time, especially systems which do not require deep semantic processing. The necessary compromise, however, is that such 'shallow' lexicons often lack the richer semantic information and structures that are required for deep language understanding applications. Nevertheless, richer semantics data can be added onto 'shallow' semantic lexicons as extra layers at a later stage, by linking to external resources, as has been done in the case of wordnet projects by Magnini and Cavaglià (2000); Niles and Pease (2001); Fontenelle (2003); Niles and Pease (2003); Shi and Mihalcea (2005); Kipper et al. (2008).

Both types of multilingual lexicon are useful and important in NLP, although it is always useful to consider the context and purpose of the lexicon before choosing one approach over the other. For the case of under-resourced languages, one important consideration is the availability of speakers of the language to help check and verify the contents. It is often difficult enough to source speakers of under-resourced languages. The pool of eligible volunteers will be greatly reduced if they are expected to have backgrounds in linguistics and semantics just for adding translation equivalents to the lexicon. Nevertheless, they may gain expertise as they progress through their involvement with the lexicon development, and could help transition the lexicon to a 'deep' one when the data coverage and content reach a more mature level in future.

For reasons of practicality, it may be more efficient to adopt a 'shallow' approach to develop a multilingual lexicon. Polyglot speakers of under-resourced languages should be allowed to contribute or verify translation equivalent entries into a simple but structured framework, without having to delve deeply into the linguistic and semantics details. This would increase the pool of qualified volunteers and help speed up the verification process. A usable multilingual lexicon can then be obtained more quickly for lexical lookup and content-scanning purposes, and also as a first step towards a lexical resource for NLP. Once a 'draft' multilingual lexicon (with shallow information)

is in place, it can be further enhanced with richer information, or adopt a 'deeper' design, to support more advanced NLP functionalities such as language understanding.

2.4 Lexicon Data Acquisition Bottleneck

As lexical resources are usually costly to construct by hand from scratch, a 'draft' copy is usually bootstrapped or automatically acquired from existing resources. There has been much work on automatic data acquisition of multilingual lexicons, but these commonly have various requirements on the input resources.

Some multilingual projects align translation equivalents from existing bilingual lexicons, using available lexical resource data. The information required include semantic relations from monolingual wordnets (Verma & Bhattacharyya, 2003; Varga, Yokoyama, & Hashimoto, 2009; Zhong & Ng, 2009); domain or categoric codes, semantic labels from existing dictionaries or lexical databases (Jalabert & Lafourcade, 2002; Lafourcade, 2002; Bond & Ogura, 2008); definition or gloss texts (Janssen, 2003, 2004; Inkpen & Hirst, 2006). Unfortunately, such comprehensive information may not be available in dictionaries of under-resourced languages at all. In the worst case, the sole field present may only be a list of translation equivalents in a TL.

S. Shirai and Yamamoto (2001) and Bond, Ruhaida, Yamazaki, and Ogura (2001) proposed a method for generating a bilingual dictionary for a new language pair, using only bilingual mappings of existing language pairs. Their method may therefore be more suitable for under-resourced languages, and might be extended to produce a multilingual dictionary. If richer resources (e.g. domain codes) become available later, the accuracy can be further improved as demonstrated by Bond and Ogura (2008) and Varga et al. (2009).

Elsewhere, Mausam et al. (2009) used probabilistic translation inference graphs, constructed from *existing* sense-distinguished multilingual Wiktionaries, to compose a massive multilingual dictionary of over 1000 languages.

Other projects attempt to mine translation equivalents from bilingual corpora (Rapp, 1999; Diab & Finch, 2000; Otero, Campos, Ramom, Campos, & Compostela, 2005; Markó, Schulz, & Hahn, 2005; Sammer & Soderland, 2007; Dorow, Laws, Michelbacher, Scheible, & Utt, 2009), which may be more readily available than specialised dictionaries, but still be difficult to obtain for under-resourced languages. Markó et al. (2005) made use of cognate mappings to derive new translation pairs, later validated by processing parallel corpora in the medical domain. However, their approach requires large aligned corpora, although such resources may be more readily available for specific domains such as medicine. Also, the cognate-based approach is not applicable for language pairs that are not closely related. Sammer and Soderland (2007) proposed a method for mining equivalents by learning context word vectors from monolingual corpora for two languages. This avoids the burden of acquiring a parallel corpus, but their particular algorithm can be prone to semantically related but erroneous equivalents, e.g. «shoot» and «bullet». Again, corpora for under-resourced languages may be hard to obtain or prepare in short time, or too small to yield satisfactory results.

In short, existing methods for automatically acquiring multilingual translation equivalents data from existing resources are abundant. However, these are often unsuitable for under-resourced languages, as the type of information or data required (e.g. special labels or codes, semantic networks, gloss texts, corpora of sufficient size) are not readily available. In a worst case scenario, the only available input may be a flat list of bilingual mappings. These are more likely to be made available by digitising existing (simple) paper bilingual dictionaries, or more easily compiled by enlisting the help of speakers of the language.

2.5 Training Resources for Translation Selection

There is a large body of work around WSD and translation selection, a good overview of which is given in (Ide & Véronis, 1998). WSD and translation selection approaches may hence be broadly classified into two categories depending on the type of learning resources used: knowledge- and corpus-based.

Knowledge-based approaches make use of various types of information from existing dictionaries, thesauri, or other lexical resources. The type of knowledge used include definition or gloss text (Lesk, 1986; Banerjee & Pedersen, 2003), subject codes (Wilks & Stevenson, 1998; Magnini, Strapparava, Pezzulo, & Gliozzo, 2001), semantic primitives (Wilks et al., 1993), semantic networks (Wu & Palmer, 1994; Agirre & Rigau, 1996; Lin, 1998; Leacock & Chodorow, 1998; Resnik, 1999; K. Shirai & Yagi, 2004) and others.

Nevertheless, lexical resources of such rich content types are usually available for medium- to rich-resourced languages only, and are costly to build and verify by hand. Knowledge-based approaches also often lack newly coined terms, or new senses of words that have emerged from popular use.

Corpus-based approaches use bilingual corpora as learning resources for translation selection, and are more likely to contain new terms and word senses. Resnik and Yarowsky (1997); Ide, Erjavec, and Tufiş (2002); Ng, Wang, and Chan (2003); Zhong and Ng (2009) used parallel or aligned corpora in their work. As it is not always possible to acquire parallel corpora, comparable corpora, or even independent second-language corpora have also been shown to be suitable for training purposes, either by purely numerical means (Brown, Pietra, Pietra, & Mercer, 1991; Fung & Lo, 1998; Li & Li, 2004) or with the aid of syntactic relations (Dagan & Itai, 1994; Zhou, Ding, & Huang, 2001). Vector-based models, which capture the context of a translation or meaning, has also been used (Schütze, 1998).

Problems with corpus-based approaches include data sparseness, i.e. 'minor' word senses are often dominated by the high occurrence frequency of 'major' senses in the corpora, as well as noisy signals from the training corpus. As a result, hybrid approaches combining knowledge- and corpus-based models have become more widely used (Stevenson & Wilks, 2001; O'Hara, Bruce, Donner, & Wiebe, 2004; Tufiş, Ion, & Ide, 2004). Also, most corpus-based approaches work on bilingual training data for bilingual translation selection tasks. Since multilingual corpora may be more difficult

to obtain, it would be interesting to see if a model trained from bilingual corpora may be applied for multilingual tasks.

2.6 Summary and Conclusion

Tables 2.2, 2.3 and 2.4 summarise the design approaches of existing multilingual lexicons efforts, as well as input data requirements for their automatic construction and translation selection. In summary, the following considerations should be taken into account while designing and constructing a multilingual lexicon, especially when under-resourced languages are to be included:

- Mechanisms for handing lexical ambiguity, lexical gaps and MWEs;
- Low linguistics expertise requirement on volunteers to optimise the pool of available speakers of under-resourced languages;
- Low requirement on input bilingual data resources to avoid data acquisition bottleneck for under-resourced languages.

To address these factors, the review of existing multilingual projects has yielded some interesting inspirations:

- A 'shallow' multilingual lexicon approach, i.e. using a language-independent axies mechanism for linking or grouping together translation equivalents (Boitet et al., 2002; Nguyen et al., 2007; ISO24613, 2008; Francopoulo et al., 2009), may lower the barrier for volunteers without linguistics knowledge to start contributing to the lexicon's development and verification;
- The LMF's extension for modelling MWEs is a flexible and comprehensive one, although the use of a phrase structure tree may not be consistent with some existing NLP applications;
- The LKB of NS approach of handling fine-grained sense distinctions, i.e. using a formal ontological model, may be approximated with dictionary usage labels;

Table 2.2: Summary of multilingual lexicon design approaches. Whether a multilingual lexicon adopts a deep or shallow approach will decide how diversification and lexical gaps are handled.

Work Cited	Approach			MWE support	Notes
	Shallow	Deep	Main structure		
Vossen (1997, 2004)	✓		wordnet		Fine sense granularity; link explosion
Tufiş et al. (2004)	✓		wordnet		Fine sense granularity
Pease et al. (2008)	✓		wordnet		Fine sense granularity
Boitet et al. (2002)	✓		pivot/axis		
Nguyen et al. (2007)	✓		pivot/axis		
ISO24613 (2008); Francopoulo et al. (2009)	✓		pivot/axis	✓	Different levels of information may be added orthogonally
Sammer and Soderland (2007)	✓		pivot/axis		
Janssen (2003, 2004)		✓	concept lattice		Some translation equivalences cannot be established; Requires in-depth semantics knowledge
UNL Center (2004); Cardeñosa et al. (2005)		✓	hypergraph	✓	Requires in-depth linguistics and semantics knowledge
Edmonds and Hirst (2002); Inkpen and Hirst (2006)		✓	ontology		Requires in-depth semantics knowledge
Proposed Work	✓		pivot/axis	✓	Flexibility to add other levels of information

Table 2.3: Summary of input data requirements of multilingual lexicon data acquisition approaches

Cited work	Input data resources required						Feasibility for under-resourced languages*
	translation lists	gloss text	category labels	wordnets	existing sense-distinguished multilingual lexicon	corpora	
Verma and Bhattacharyya (2003)	✓			✓			×
Varga et al. (2009)	✓			✓			×
Jalabert and Lafourcade (2002)	✓		✓				×
Lafourcade (2002)	✓		✓				×
Bond and Ogura (2008)	✓		✓				×
Janssen (2003, 2004)		✓					△
Inkpen and Hirst (2006)	✓	✓					△
S. Shirai and Yamamoto (2001)	✓						○
Bond et al. (2001)	✓						○
Mausam et al. (2009)					✓		×
Rapp (1999)						✓	△
Diab and Finch (2000)						✓	△
Otero et al. (2005)						✓	△
Markó et al. (2005)	✓					✓	△
Sammer and Soderland (2007)	✓					✓	△
Dorow et al. (2009)	✓					✓	△
Proposed Work	✓						○

* ×— Unfeasible; △—May be feasible; ○—Feasible

Table 2.4: Summary of training data sources for translation selection and/or WSD approaches

Cited work	Training data required						Feasibility for under-resourced languages*
	gloss text	category labels	wordnets	monolingual corpora	comparable corpora	aligned or tagged corpus	
Lesk (1986)	✓						△
Banerjee and Pedersen (2003)	✓						△
Wilks and Stevenson (1998)		✓					×
Magnini et al. (2001)		✓					×
Wu and Palmer (1994)			✓				×
Agirre and Rigau (1996)			✓				×
Lin (1998)			✓				×
Leacock and Chodorow (1998)			✓				×
Resnik and Yarowsky (1997)						✓	×
Ide et al. (2002)						✓	×
Ng et al. (2003)						✓	×
Fung and Lo (1998)					✓		△
Li and Li (2004)					✓		△
Dagan and Itai (1994)				✓			△
Zhou et al. (2001)				✓			△
Stevenson and Wilks (2001)	✓	✓		✓			×
O'Hara et al. (2004)	✓		✓	✓			×
Tufiş et al. (2004)			✓			✓	×
Proposed work					✓		○

* ×— Unfeasible; △—May be feasible; ○—Feasible

- Ideally, the data acquisition process should require only very simple input bilingual data, especially in the case of under-resourced languages, in order to automatically produce a ‘first draft’ of a shallow multilingual lexicon quickly.

- To provide context-dependent lookup features, similar to translation selection or WSD features, the multilingual lexicon should be suitably enriched with extra information, the model of which can be preferably learnt from easily acquired data and applicable to under-resourced languages as well.

CHAPTER 3

DESIGN AND CONSTRUCTION OF LEXICON+TX

Lexicon+TX (a lexicon with applications to **T**ranslation and cross(**X**)-lingual lookup) is a multilingual translation lexicon designed to be easy to construct, use and maintain. Its purpose is to connect under-resourced languages to richer-resourced languages by providing translation equivalents from different languages, so that NLP applications and human users can benefit from more language pairs.

The design and construction of Lexicon+TX is driven by two main principles:

1. The lexicon framework should assume minimum linguistics knowledge and expertise on the part of contributors, so that a larger pool of contributors may participate in the construction and maintenance of the lexicon content.
2. It should be possible to automatically generate a first draft or prototype of the lexicon, imposing only minimum requirements on the input lexical data. This is especially important for the inclusion of under-resourced languages.

This chapter will first describe the design of Lexicon+TX which is largely inspired by the LMF (ISO24613, 2008) and the Papillon Multilingual Dictionary (Boitet et al., 2002). We then describe how a prototype of the lexicon can be generated using simple data which are easier to obtain. The lexicon prototype can then be checked and improved by human contributors, thus cutting down the efforts required of the contributors if they were asked to create the entire lexicon contents from scratch.

3.1 Design of Lexicon+TX

Lexicon+TX is a 'shallow' multilingual lexicon. It does not attempt to propose any interlingual framework to describe the underlying semantic components of lexical

meanings. Instead, Lexicon+TX simply lists translation (near-)equivalents of different languages that express the same concept, on a coarse-grained basis.

Lexicon+TX is designed to be easy to construct and use. In particular, its framework does not require a human contributor to have extensive linguistics expertise. The goal is to allow a bilingual or multilingual speaker to simply specify which LIs from different languages denote the same meaning for the lexicon prototype, without having to understand or delve deeply into the semantic details.

The macrostructure (how multilingual entries are organised), and the microstructure (lexical information about each monolingual LI) of Lexicon+TX will be presented in the following subsections. This discussion focuses on the listing of multilingual translation equivalents aspect only. Modelling of other linguistics aspects (e.g. morphological and syntactical) for each individual language are outside the scope of this thesis. Nevertheless, the multilingual aspect of Lexicon+TX is orthogonal to these aspects (as in the LMF). Therefore, extensions for these purposes, such as those from the LMF, may be introduced into an implementation of Lexicon+TX without conflict.

3.1.1 Macrostructure

The macrostructure specifies how lexical entries in Lexicon+TX are organised and related to each other and can be summarised as below:

$\langle$*Lexicon*$\rangle$ ::= $\langle$*translation set*$\rangle$+

$\langle$*translation_set*$\rangle$ ::= ($\langle$*trans_equiv*$\rangle$+, $\langle$*seminfo*$\rangle$?)

$\langle$*trans_equiv*$\rangle$ ::= an entry of a LI or a gloss phrase in a TL; see next section.

$\langle$*seminfo*$\rangle$ ::= data for semantic processing purposes

Entries in Lexicon+TX are organised as multilingual *translation sets*. Each translation set corresponds to a coarse-grained lexical sense or concept, and is accessed

by a language-independent axis node. Translation equivalents expressing the same sense are connected to the axis, similar to the structural scheme used in the multilingual extension of LMF (ISO24613, 2008; Francopoulo et al., 2009) and the Papillon Multilingual Dictionary (Boitet et al., 2002). The scheme makes it easy to add a new language to the lexicon, as new translation equivalents are added to the translation set via the language-independent axis, as opposed to being linked to every other existing language in the lexicon.

Each translation set may be associated with extra semantic information, which may be used for semantic processing purposes (including translation selection). The nature and approach of this semantic information is up to the lexicon designer's choice and needs of specific applications. One possibility is in the form of semantic relations between the axis nodes, i.e. a semantic network similar to a wordnet. Another possible approach requiring only minimal human effort, using distributional information extracted from comparable bilingual corpora, is described in Chapter 4.

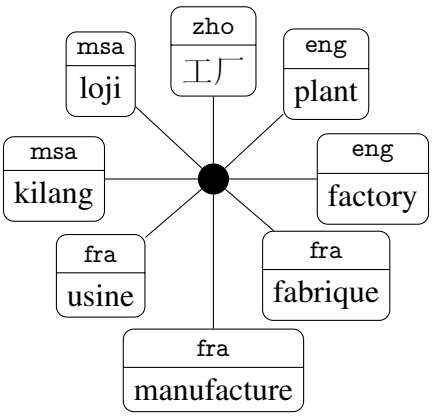
Content-wise, Lexicon+TX's translation sets are similar to Sammer and Soderland's (2007) data structures of the same name: 'a multilingual extension of a WordNet synset (Fellbaum, 1998)' and contains 'one or more LIs in each k languages that all represent the same word sense'. Figure 3.1 shows the conceptual view of two example translation sets: one representing the concept of *industrial plant*, and the other of *plant life*, with lexicalisations or translations from English, Chinese, Malay and French.

The following subsections will give further examples to illustrate how such a language-independent axis framework handles different multilingual issues and lexicography requirements in Lexicon+TX.

3.1.1 (a) Multiple Senses

Similar to other multilingual lexicon projects, an LI with multiple senses will appear in translation sets corresponding to those senses. For example, the English noun «plant» has (amongst others) two senses: one for *industrial plant*, and one for *plant*

English	Chinese	Malay	French
factory plant	工厂	loji kilang	fabrique manufacture usine



English	Chinese	Malay	French
plant vegetation	植物	tumbuhan tumbuh-tumbuhan	végétal

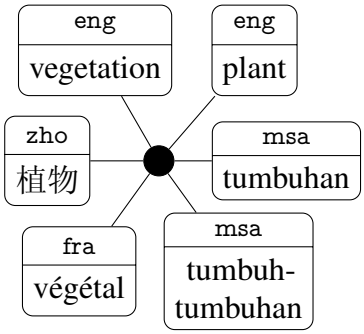


Figure 3.1: Example translation sets for the word senses *industrial plant* and *plant life*, with lexical items from English, Chinese, Malay and French.

life. The LI «plant» therefore appears in those two relevant translation sets, as shown in Figure 3.1.

Lexicon+TX adopts a coarse-grained sense distinction, with TL translation items being a driving principle. As a comparison, WordNet distinguishes between «chicken» the animal and «chicken» the edible meat, and also between «break» the transitive action (‘he broke the glass’) and «break» the intransitive verb (‘the glass broke’). Lexicon+TX discerns only one sense of «chicken» and «break» in these cases, unless they are translated differently in some TL. This would then be regarded as a diversification case (see next subsection).

3.1.1 (b) Diversification

Diversification in Lexicon+TX is handled via diversification links between the language-independent axis nodes, as is done in Papillon and LMF. (This can be

considered a kind of ⟨*seminfo*⟩ mentioned in section 3.1.1.) Figure 3.2 shows an example «rice». English «rice» and French «riz» do not distinguish between *cooked* rice (Malay «nasi» and Chinese «饭») and *uncooked* rice grains (Malay «beras» and Chinese «米»). The axis connecting «rice» and «riz» is therefore *diversified* to two other axes, each representing the concepts *cooked* and *uncooked* rice respectively.

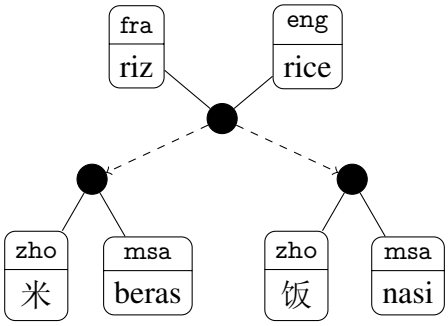


Figure 3.2: Handling diversification of «rice» in Lexicon+TX

3.1.1 (c) Lexical Gaps

A translation set can be created for a concept in Lexicon+TX, even if a lexical gap occurs in a member language, i.e. the concept is not lexicalised in that language. For example, English «foal» (a young horse) has the Chinese LI «驹子» (jūzi) and French LI «poulain» as translations, but can only be translated as the noun phrase 'anak kuda' in Malay. All four items can be included in the translation set, but the entry 'anak kuda' will be marked explicitly as a gloss item, while the other three entries will be marked as LIs.

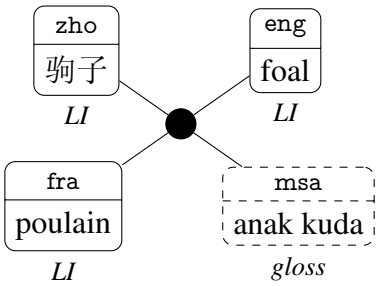


Figure 3.3: Representing lexical gaps with gloss phrases in Lexicon+TX

3.1.2 Microstructure

The microstructure design concerns fields used to document information about each translation form in Lexicon+TX. Only the most essential fields for lexical translation purposes are discussed here for focus. See the LMF specification for modelling of various other linguistic properties, including for morphology and syntactical attributes.

The core microstructure of each translation entry connected to the Lexicon+TX language-independent axis nodes can be summarised thus:

⟨*trans_equiv*⟩	::=	(⟨*language*⟩, ⟨*lemma*⟩|⟨*gloss*⟩, ⟨*label*⟩∗)+
⟨*language*⟩	::=	3-letter ISO 639-3 identifier of a language (Appendix A)
⟨*lemma*⟩	::=	(⟨*string*⟩, ⟨*tree representation*⟩)
⟨*gloss*⟩	::=	(⟨*string*⟩, ⟨*tree representation*⟩)
⟨*label*⟩	::=	various usage labels e.g, subject-field, geographical, etc.

In particular, the modelling of MWEs and gloss phrases using tree representation, annotated using SSTC (Appendix C), is a novel contribution in lexicon design.

3.1.2 (a) Language Identifier

The language of each translation entry is identified by the 3-letter ISO 639-3 code (`http://www.sil.org/iso639-3/`). For example, the code for English is `eng`, and the code for Malay is `msa`. See Appendix A for a list of ISO 639-3 codes used in this thesis.

3.1.2 (b) Form and Tree Structure of Lemma or Gloss

Translation equivalents Lexicon+TX can be either LIs, or gloss phrases in cases of lexical gaps. In addition, Lexicon+TX also accepts MWEs as LIs. It is therefore desirable to record the internal structures of these constructs as trees, to enable MT

systems to produce syntactically correct translations. This is especially helpful in the case of syntactically flexible MWEs with ‘placeholders’.

Any arbitrary tree structure may be used for representing the internal structure of lexical forms. The LMF MWE extension uses phrase structure trees. The functional dependency tree representation is adopted in this thesis. See Appendix C for a description of Structured String-Tree Correspondence (SSTC), a possible representation schema of this string-tree structure, the use of which is a novel element in lexicon design. Both inflected and lemma forms may be recorded in an SSTC. Therefore, if a token manifests in with an affix in an MWE e.g. «berat sama dipikul», both the lemma «pikul» and affixed «dipikul» are recorded in the relevant tree node – see Appendix C for further elaboration.

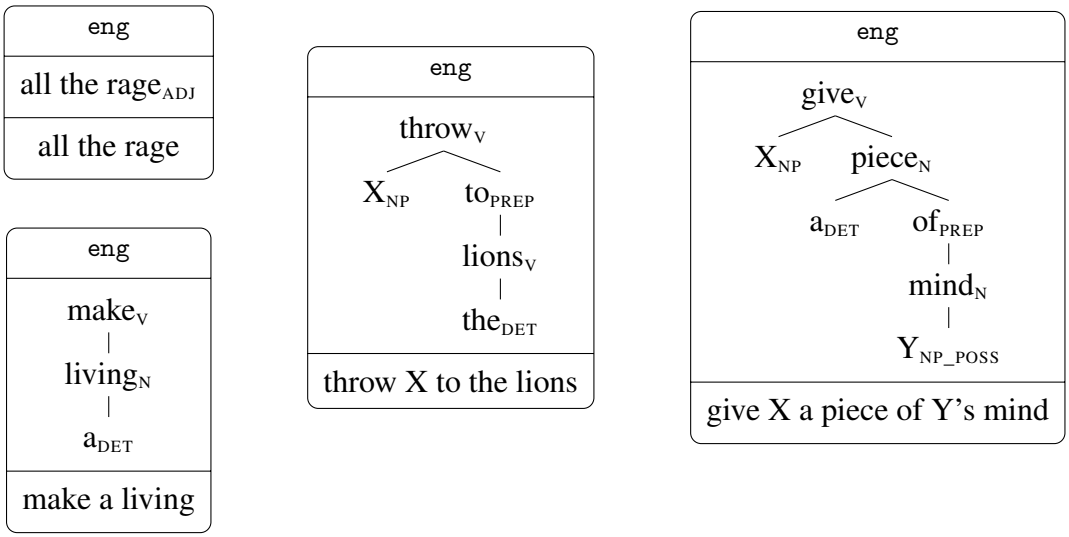


Figure 3.4: Modelling MWEs in Lexicon+TX

Figure 3.4 shows some examples of how MWEs are represented in Lexicon+TX. MWEs that are not deemed discomposable, such as «all the rage» (as well as single word LIs), have a trivial tree with a single node as the internal tree structure. Note the use of ‘placeholders’ in «throw X to the lions» and «give X a piece of Y’s mind».

In practice, such tree representations (which are not shallow) are not present in

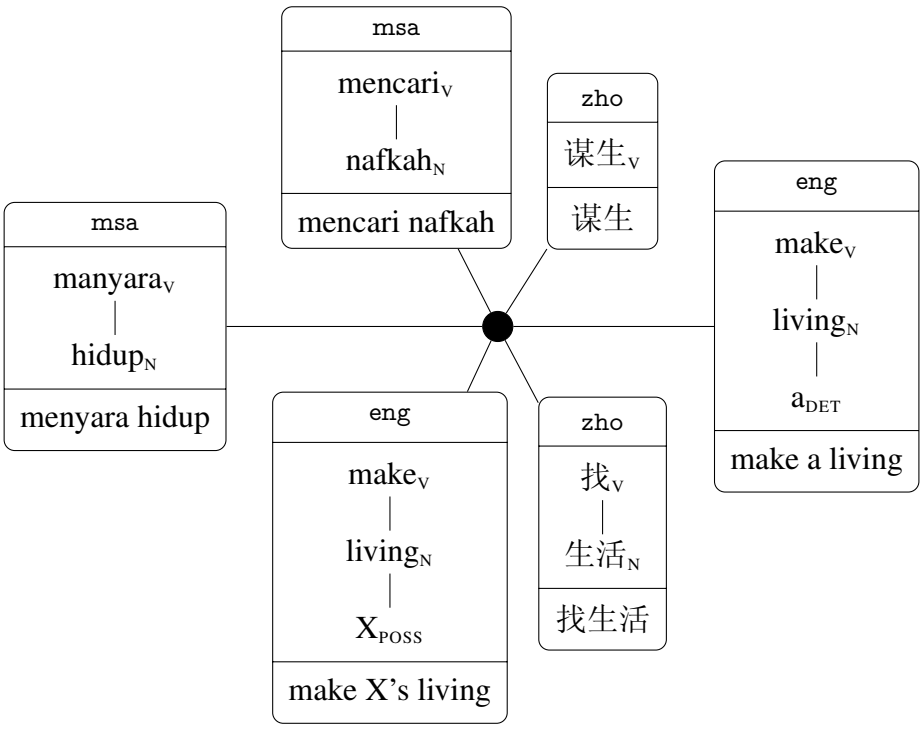


Figure 3.5: A translation set with MWEs as members

bilingual dictionaries, but can be generated automatically using parsers. In addition, the 'placeholders' may be inserted by processing the dictionary entry (a quick search-and-replace may suffice), which is often given in the form of 'throw *somebody* to the lions' or 'give *somebody* a piece of *one's* mind'.

For the sake of illustration, Figure 3.5 shows a translation set containing MWEs members as the LIs. There may also be translation equivalents which are *not* MWEs; in such cases it is also desirable to have the tree representation of the gloss phrases, as mentioned at the beginning of this section. Note that for brevity's sake, the tree representation of a translation entry may sometimes be omitted from figures in this thesis.

3.1.2 (c) Usage Labels

Usage labels may be attached to each translation equivalent in a translation set, to indicate when one translation is preferred to another. Note that usage labels are meant to help distinguish between near-synonyms as opposed to diversification.

Diversification is due to lexicalisations of more specific senses in some languages, while near-synonyms convey the same sense but differ in their context of use.

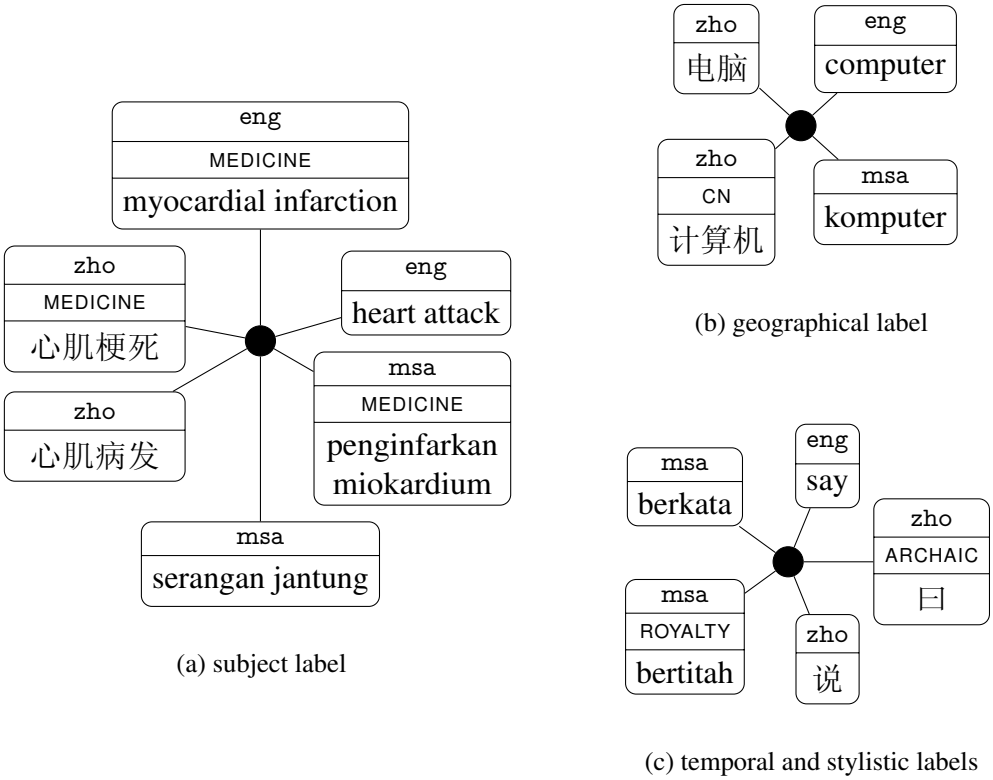


Figure 3.6: Example labels of translation equivalents

Usage labels may pertain to various aspects, as analysed at length by Janssen et al. (2003). Figure 3.6 shows some examples:

subject «myocardial infarction», «心肌梗死» and «penginfarkan miokardium» in Figure 3.6(a) are technical terms used in MEDICINE for «heart attack»;

geographical as shown in in Figure 3.6(b), a «computer» is known as a «计算机» in China (CN). («计算机» is only used to indicate a «calculator» in other Chinese-speaking regions.)

temporal some LIs, e.g. «曰» (yuē) in Figure 3.6(c), are ARCHAIC.

stylistic in Figure 3.6(c), «bertitah» is a form of «say» reserved for ROYALTY in Malay.

Each LI may be associated with multiple usage labels as necessary. A more detailed organisation, such as a hierarchy or even an ontology for usage labels (Edmonds & Hirst, 2002) may be desirable but is not discussed here, as it is out of the scope of this thesis (but see section 6.4.1 (c)).

3.2 Constructing Lexicon+TX with Simple Input Data

Manually populating Lexicon+TX with translation equivalents by human contributors would ensure the highest accuracy, but would also be a very labour- and time-intensive task. A more feasible solution would be to automatically generate a first draft of the lexicon from available data, then asking human contributors to improve upon the draft lexicon.

There is much work on mining translation equivalents from parallel corpus, but the lexical senses obtained are often constrained by the corpus domain, while less-dominant lexical senses are often missed as they occur less frequently in the corpus. In addition, bilingual corpora in under-resourced languages may not be readily available. On the other hand, bilingual dictionaries and terminology lists would have a larger overall coverage of both dominant and minor lexical senses. One frequent complaint against dictionary sources is that they lack proper noun entries, especially names of people, places and organisations, as well as newly coined terms (neologisms) related to new technologies and sub-cultures. These entries and their translations can be obtained easily from terminology bases, or from Wikipedia article titles instead, which are linked to multilingual articles (if available) about the same topic. An example is 'bromance',[1] the English Wikipedia article on which is linked to the Malay Wikipedia article entitled 'cinta antara saudara' and also Chinese '兄弟情'.

The following subsections will describe how multilingual entries for Lexicon+TX can be obtained automatically from two types of easily available sources, namely Wikipedia article titles and bilingual translation lists.

[1] a close but nonsexual relationship between two men.

3.2.1 Using Wikipedia Article Titles

Wikipedia (http://www.wikipedia.org/) is an online free encyclopædia that anyone from the online community can edit. Thanks to this 'crowdsourcing' approach, Wikipedia has over 20,000,000 articles on various topics, including new topics emerging from contemporary technology and sub-culture, in 284 languages.[2] All Wikipedia text contents are licensed under the Creative Commons Attribution-ShareAlike License-BY-SA and the GNU Free Documentation License (GFDL), and can be obtained without charge from http://dumps.wikimedia.org/ or http://en.wikipedia.org/wiki/Special:Export. These factors make Wikipedia articles a desirable data source for various NLP research and development purposes.

```
<page>
  <title>Florence</title>
  <id>11525</id>
  ...
  <revision>
    <text>
    ...
    [[de:Florenz]]    [[ko:피렌체]]
    [[fr:Florence]]    [[it:Firenze]]
    [[ms:Florence]]    [[ru:Флоренция]]
    [[zh:佛罗伦萨]]
    ...
    </text>
  </revision>
</page>
```

```
<page>
  <title>翡冷翠</title>
  <id>113446</id>
  <revision>
    ...
    <text>
      #REDIRECT [[佛罗伦萨]]
    </text>
  </revision>
</page>
```

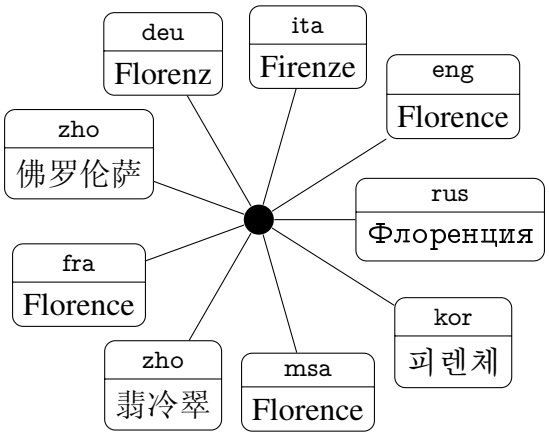


Figure 3.7: Quick extraction of translations of names from Wikipedia article titles

[2] http://s23.org/wikistats/wikipedias_html, visited on 4 February 2013.

Each Wikipedia article is linked to other articles about the same topic title in other languages, if they are available. Spelling alternatives or acronyms of a topic title in the same language are also linked. This provides us with a convenient source with multilingual translations of named persons, organisations, places, events and things, which are easy to extract programmatically. An example is given in Figure 3.7, where translations of the city name of Florence can be extracted quickly by simply parsing the Wikipedia article about the city. Nevertheless, translations in under-resourced languages are still likely not in abundance, as there are few (if any) Wikipedia articles in these languages.

3.2.2 Using Bilingual Translation Lists

Bilingual dictionaries with substantial content for any given language are likely to exist.[3] Given the abundance of bilingual machine-readable dictionaries (MRDs) and lexicons, there have been many efforts at automatically merging these bilingual lexicons into a sense-distinguished multilingual lexicon (Lafourcade, 2002; Janssen, 2003, 2004; Tufiş et al., 2004; Vossen, 2004; Inkpen & Hirst, 2006).

Many of these approaches require the input bilingual MRDs to include certain types of information besides equivalents in the TL, such as gloss or definition text, domain labels or semantic field codes. Unfortunately, bilingual MRDs with such features are not always available, especially for under-resourced language pairs. Moreover, bilingual MRDs vary greatly both in their sense distinction granularities and structural organisation, which add to the difficulty of aligning entries at the sense level. More often than not, the lowest common denominator across bilingual lexicons is just a simple list of mappings from a SL item to one or more TL equivalents.

This research proposes that multilingual translation sets can be bootstrapped from simple lists of bilingual translations, which are easier for native speakers to provide, or extracted from bilingual MRDs. Such low resource requirements (as well as the low-cost method that will be described) are especially suitable for under-resourced

[3]except for highly endangered languages, or those without a writing system

language pairs. This is achieved using a modified version of the one-time inverse consultation (OTIC) procedure proposed by Tanaka, Umemura, and Iwasaki (1998).

3.2.2 (a) One-time Inverse Consultation

Tanaka et al. (1998) first proposed the OTIC procedure to generate a bilingual lexicon for a new language pair L_1–L_3 via an intermediate language L_2, given existing bilingual lexicons for language pairs L_1–L_2, L_2–L_3 and L_3–L_2. Following is an example of a OTIC procedure for linking Japanese words to their Malay translations via English:

- For every Japanese word, look up all English translations ($\mathbb{E}_1$).
- For every English translation, look up its Malay translations ($\mathbb{M}$).
- For every Malay translation, look up its English translations ($\mathbb{E}_2$), and see how many match those in $\mathbb{E}_1$.
- For each $m \in \mathbb{M}$, the more matches between $\mathbb{E}_1$ and $\mathbb{E}_2$, the better m is as a candidate translation of the original Japanese word.

$$\text{score}(m) = 2 \times \frac{|\mathbb{E}_1 \cap \mathbb{E}_2|}{|\mathbb{E}_1| + |\mathbb{E}_2|}$$

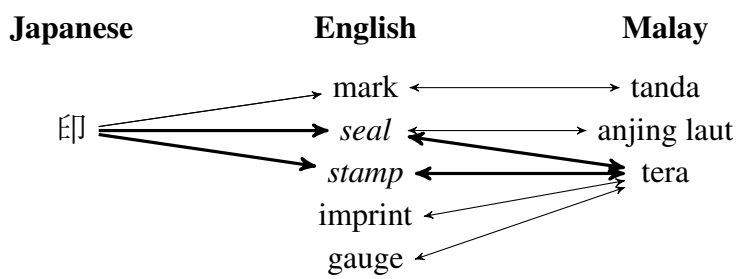


Figure 3.8: Using OTIC, Malay «tera» is determined to be the most likely translation of Japanese «印» as they are linked by the most number of English words in both directions, with score(«tera») $= 2 \times \frac{2}{3+4} = 0.57$. (Diagram from Bond & Ogura, 2008)

A worked example is shown in Figure 3.8. The Japanese word «白» (*shirushi*) has 3 English translations, which in turn yield another three Malay translations. Among them, «tera» has 4 English translation, 2 of which are also present in the earlier set of 3 English translations. The one-time inverse consultation score for «tera» is thus $2 \times \frac{2}{3+4} = 0.57$, and indicates «tera» is the most likely Malay translation for «白».

Bond et al. (2001) extended OTIC by linking through two languages, as well as utilising semantic field codes and classifier information to increase precision, but these extensions may not always be possible as not all lexical resources include these information (nor do all languages use classifiers).

3.2.2 (b) Extension to OTIC

OTIC was originally conceived to produce a list of bilingual translations for a new language pair. As our aim is a multilingual lexicon instead, we modified the OTIC procedure to produce trilingual translation triples and translation sets, as outlined in Algorithm 1.

Algorithm 1 Generating trilingual translation triples from bilingual translation lists

```
GENERATETRIPLES(𝕃_{L1−L2}, 𝕃_{L2−L3}, 𝕃_{L3−L2})
FILTERSETS(T, α, β)
MERGESETS(T)

procedure GENERATETRIPLES(𝕃_{L1−L2}, 𝕃_{L2−L3}, 𝕃_{L3−L2})
    T ← empty set
    for all lexical items w_h ∈ L_1 do
        𝕎_m ← translations of w_h in L_2 (from 𝕃_{L1−L2})
        for all w_m ∈ 𝕎_m do
            𝕎_t ← translations of w_m in L_3 (from 𝕃_{L2−L3})
            for all w_t ∈ 𝕎_t do
                Add translation triple (w_h, w_m, w_t) to T
                𝕎_{m_r} ← translations of w_t in L_2 (from 𝕃_{L3−L2})
```

13: $\text{score}(w_h, w_m, w_t) \leftarrow \sum_{w \in \mathbb{W}_m} \frac{\text{no. of common words in } w_{m_r} \in \mathbb{W}_{m_r} \text{ and } w}{\text{no. of words in } w_{m_r} \in \mathbb{W}_{m_r}}$

14: **end for**

15: $\text{score}(w_h, w_t) \leftarrow 2 \times \frac{\sum_{w \in \mathbb{W}_m} \text{score}(w_h, w, w_t)}{|\mathbb{W}_m| + |\mathbb{W}_{m_r}|}$

16: **end for**

17: **end for**

18: **end procedure**

19: **procedure** FILTERTRIPLES(T, α, β) ▷ T is a set of translation triples (w_h, w_m, w_t) with a score

20: **for all** lexical items $w_h \in L_1$ **do**

21: $X \leftarrow \max_{w_t \in \mathbb{W}_t} \text{score}(w_h, w_t)$

22: **for all** distinct translation pairs (w_h, w_t) **do**

23: **if** $\text{score}(w_h, w_t) \geq \alpha X$ or $(\text{score}(w_h, w_t))^2 \geq \beta X$ **then**

24: Place $w_h \in L_1$, $w_m \in L_2$, $w_t \in L_3$ from all triples $(w_h, w_{...}, w_t)$ in same translation set

25: Record $\text{score}(w_h, w_t)$ and $\text{score}(w_h, w_m, w_t)$

26: **else**

27: Discard all triples $(w_h, w_{...}, w_t)$

28: **end if**

29: **end for**

30: **end for** ▷ The sets are now grouped by (w_h, w_t)

31: **end procedure**

32: **procedure** MERGESETS(T)

33: Merge all translation sets containing triples with same (w_h, w_m)

34: Merge all translation sets containing triples with same (w_m, w_t)

35: **end procedure**

Algorithm 1 allows partial word matches between the 'forward' ($\mathbb{W}_m$) and 'reverse' ($\mathbb{W}_{m_r}$) sets of intermediate language words. For example, if the 'forward' set

(garang, 凶猛) 0.143

(garang, ferocious, 凶猛)
(garang, fierce, 凶猛)

(garang, 激烈) 0.125

(garang, jazzy, 激烈)

~~(garang, 大胆) 0.111~~

~~(garang, bold, 大胆)~~

~~(garang, 黑体) 0.048~~

~~(garang, bold, 黑体)~~

~~(garang, 粗体) 0.048~~

~~(garang, bold, 粗体)~~

⋮

Figure 3.9: Generated translation triples from Algorithm 1

contains «coach» and the reverse set contains «sports coach», the modified OTIC score is $\frac{1}{2} = 0.5$, instead of 0. This would also serve as a likelihood measure for detecting diversification in future improvements of the algorithm. The score computation for (w_h, w_t) is also adjusted accordingly to take into account this substring matching score (line 15), as opposed to the exact matching score in the original OTIC.

We retain the intermediate language words along with the 'head' and 'tail' languages, i.e. the OTIC procedure will output translation *triples* instead of pairs. α and β on line 23 are threshold weights to filter translation triples of sufficiently high scores. Bond et al. (2001) did not discard any translation pairs in their work; they left this task to the lexicographers who preferred to whittle down a large list rather than adding new translations. In our case, however, highly suspect translation triples must be discarded to ensure the merged multilingual entries are sufficiently accurate. Specifically, the problem is when an intermediate language word is polysemous. Erroneous translation triples (w_h, w_m, w_t) may then be generated (with lower scores), where the translation pair (w_h, w_m) does not reflect the same meaning as (w_m, w_t). If such triples are allowed to enter the merging phase, the generated multilingual entries would eventually contain words of different meanings from the various member languages: for example, English «bold», Chinese «黑体» (*hēitǐ*, 'bold typeface') and Malay «garang» ('fierce') might be placed in the same translation set by error.

As an example, consider the (w_h, w_m, w_t) translation triples with non-zero

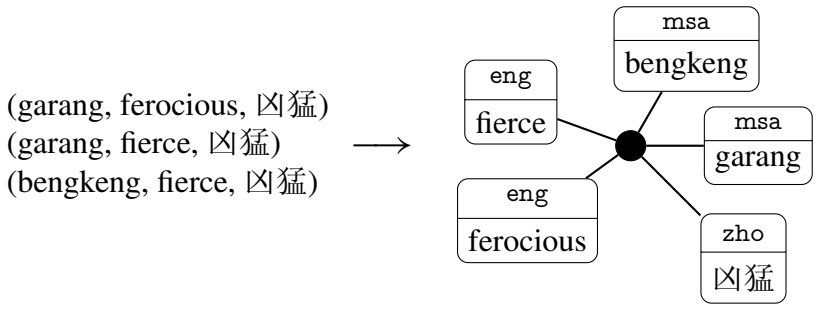


Figure 3.10: Merging translation triples into translation sets

scores generated by OTIC where w_h = «garang», presented in Figure 3.9. The highest score(w_h, w_t) is 0.143. When $\alpha = 0.8$ and $\beta = 0.2$, (w_h, w_t) pairs whose score is less than $\alpha \times 0.143 = 0.1144$, or when squared is less than $\beta \times 0.143 = 0.0286$ will be discarded. Therefore, triples containing (garang, 大胆) (and other pairs of lower scores) will be discarded as its score 0.111 and squared score 0.0123 are lower than both threshold values.

The retained translation triples are then merged into translation sets based on overlapping translation pairs among the languages. An example is shown in Figure 3.10, where the translation triples are merged into one translation set with five members.

3.2.2 (c) Adding More Languages

The algorithm described in the previous section gives us a trilingual translation lexicon for languages $\{L_1, L_2, L_3\}$. Algorithm 2 outlines how a new language L_4, or more generally L_{k+1}, can be added to an existing multilingual lexicon of languages $\{L_1, L_2, \ldots, L_k\}$. We first run OTIC to produce translation triples for L_{k+1} and two other languages already included in the existing lexicon. These new triples are then compared against the existing multilingual translation set entries. If two words in a triple are present in an existing translation set, the third word is added to that translation set as well.

Algorithm 2 Adding L_{k+1} to multilingual lexicon $\mathbb{L}$ of $\{L_1, L_2, \ldots, L_k\}$

1: GENERATETRIPLES($\mathbb{L}_{L_{k+1}-L_m}$, $\mathbb{L}_{L_m-L_n}$, $\mathbb{L}_{L_n-L_m}$) ▷ Or other permutations

```
FILTERSETS(T, α, β)
ADDLANG(T, 𝕃_{L_1,...,L_k})

procedure ADDLANG(T, 𝕃_{L_1,...,L_k})
    repeat
        cnt ← |T|
        for all (w_{L_{k+1}}, w_{L_m}, w_{L_n}) ∈ T do
            if there exists translation sets in 𝕃 that contains both w_{L_m} and w_{L_n} then
                Add w_{L_{k+1}} to all these translation sets
                Delete (w_{L_{k+1}}, w_{L_m}, w_{L_n}) from T
            end if
        end for
        cnt′ ← |T|
    until cnt = cnt′
    MERGESETS(T)
    Add new translation sets to 𝕃_{L_1,...,L_k}
end procedure
```

Figure 3.11 gives such an example: given the English–Chinese–Malay translation set earlier, we prepare translation triples for French–English–Malay. By detecting overlapping English–Malay translation pairs in the translation set and triples, two new French LIs «cruel» and «féroce» are added to the existing translation set.

If there is available resources for generating triples in more languages for matching, then the approach outlined in Bond and Ogura (2008) can be applied, which would also increase the accuracy.

3.2.2 (d) Extracting Bilingual Dictionaries for New Languages

The constructed Lexicon+TX is also a repository from which bilingual dictionaries for new language pairs, especially less common ones, can be quickly extracted.

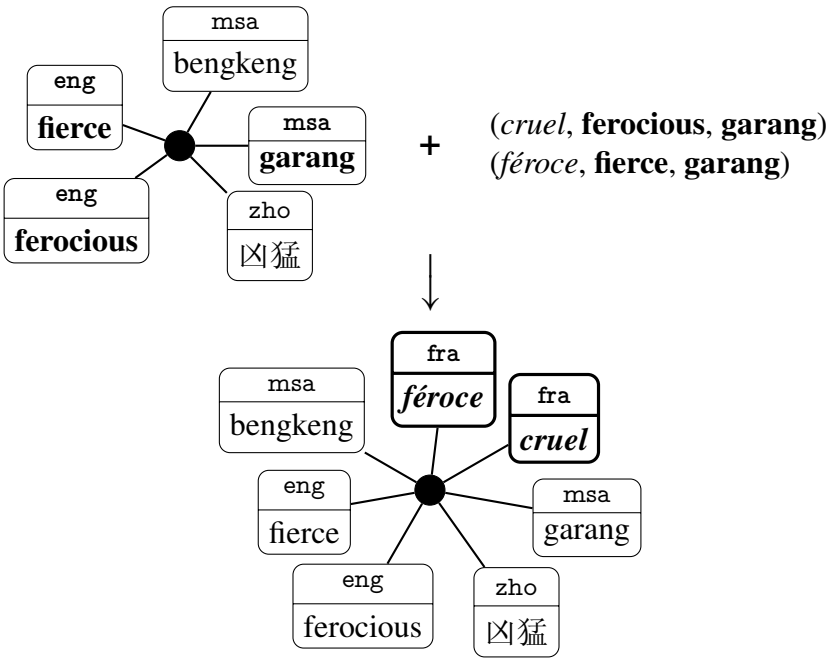


Figure 3.11: Adding French members to existing translation sets

Based on the methods proposed in the previous sections, the work flow for constructing a new multilingual lexicon, or adding new languages to an existing one, for the express purpose of extracting new bilingual dictionaries, is summarised in Figure 3.12.

If the first few languages to be added to the multilingual lexicons are resource-rich, other construction approaches utilising richer lexical resources reviewed in section 2.4 can be used to build the initial multilingual lexicon instead. Under-resourced languages can then be added to this multilingual lexicon following the workflow in Figure 3.12.

3.2.3 Lexicon Maintenance

Once a draft copy of Lexicon+TXhas been created, maintenance is relatively straightforward and would consist of the following main operations, based on a human judge's evaluation of a translation set (see figure 3.9 and sections 5.1.2, 5.1.3):

- merging translation sets;
- deleting entire translation sets;

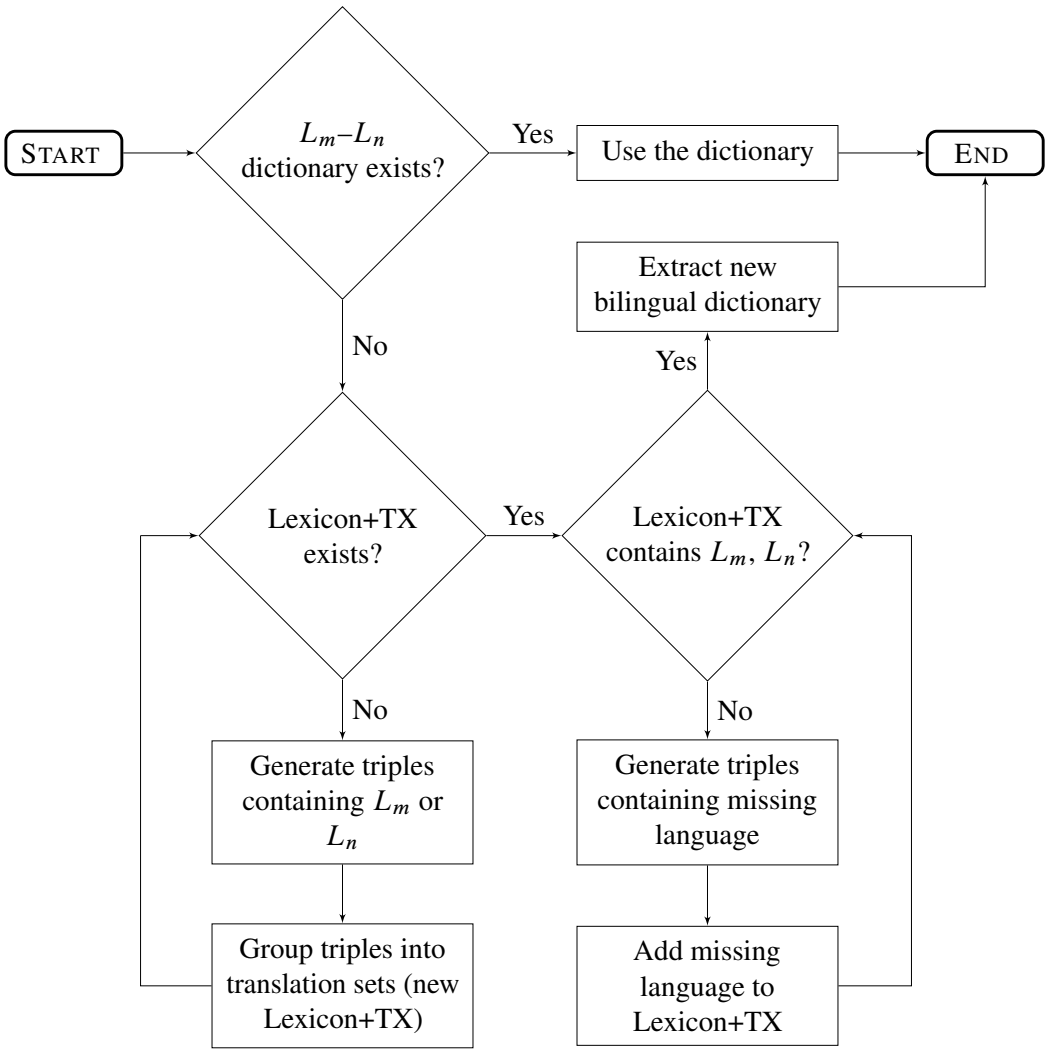


Figure 3.12: Flowchart for creating a new multilingual lexicon (Lexicon+TX) and adding new languages, so that new bilingual dictionaries can be extracted

- deleting a member LI from a translation set;
- adding a member LI to a translation set;
- splitting one translation set into more sets, which may be distinct or connected by diversification links.

When the original input dictionaries are updated, the changes may be propagated to Lexicon+TX. If new entries are added to the original input dictionaries, new translation triples can be generated and added to exiting translation sets. However, there

is currently no good way of propagating deletions of entries and translation equivalence from the input dictionaries to Lexicon+TX.

3.3 Summary and Conclusion

This chapter presented the design of Lexicon+TX, a multilingual lexicon which does not presume linguistic expertise on its human contributors. The structure of translation sets that make up Lexicon+TX is inspired by the LMF, and uses tree structures for handling MWEs as its translation equivalents members. Gloss phrases may also be used in cases of lexical gaps. The design also allows for richer information to be added to the lexicon at a later stage, allowing the initial effort to focus on acquiring multilingual translation equivalents only. This chapter also proposed procedures for automatically generating 'draft' multilingual translation sets from data sources that are easier to obtain, i.e. Wikipedia article titles and bilingual translation lists.

By enforcing the principle of 'minimum requirements' on linguistic expertise and input date richness, the proposed design and construction procedure allows the prototype of a multilingual lexicon, especially for under-resourced languages, to be created quickly and with minimum cost.

The next chapter will demonstrate how the constructed multilingual lexicon, Lexicon+TX, can be used as a reading aid via intelligent word look-up functions.

CHAPTER 4

CONTEXT-DEPENDENT MULTILINGUAL LEXICON LOOK-UP AND TRANSLATION SELECTION

Once Lexicon+TX with member languages $L_1, L_2, \ldots, L_N$ (see Chapter 3) is in place, the next step would be to provide context-dependent lexical lookup functions. Given an input text in language L_i ($1 \leq i \leq N$), the lookup module should return a list of multilingual translation set entries, which would contain $L_1, L_2, \ldots, L_N$ translation equivalents of LIs in the input text, wherever available.

For polysemous LIs in the input text, the lookup module should return translation sets that convey the appropriate meaning in context. This bears some similarity to WSD (which word sense is used in a context) and translation selection (which TL items should be used to translate a SL item). To this end, some kind of model and data for translation knowledge is necessary.

This chapter proposes a relatively low-cost approach to perform context-dependent lexical lookup, based on translation knowledge acquired from a comparable bilingual corpus and transferred into Lexicon+TX (sections 4.1, 4.2). The use of comparable corpus eliminates the need for acquiring or constructing a parallel aligned corpus, which is a time- and labour-intensive effort. Under-resourced language pairs can then leverage the translation knowledge available to richer-resourced languages, for which comparable bilingual corpora are easier to obtain. The lexical lookup procedure will also identify occurrences of MWEs, which may comprise discontiguous strings in the input text.

For consumption of other NLP systems, results from the context-dependent lookup module need to be packaged in a machine-tractable format. A new annotation schema, SSTC+Lexicon (SSTC+L), is proposed for relating lemmas from a lexicon to

their occurrences in an input text (section 4.3). The SSTC+L can handle discontiguous MWE occurrences, as well as annotating translational lexical gaps when used in conjunction with the Synchronous SSTC (S-SSTC) (see also Appendix C).

4.1 Mining Translation Knowledge from Comparable Bilingual Corpora

Corpus-driven translation selection approaches typically derive supporting semantic information from an aligned corpus, in which a text and its translation are aligned at the sentence, phrase and word level. However, aligned corpora can be difficult to obtain for under-resourced language pairs, and are expensive to construct.

On the other hand, documents in a comparable corpus comprise bilingual or multilingual text of a similar nature, and need not even be exact translations of each other. The texts are therefore unaligned except at the document level. Comparable corpora are relatively easier and cheaper to obtain, especially for richer-resourced languages. This section describes a proposed approach for extracting translation knowledge, in the form of translation equivalence contexts, from a bilingual comparable corpus. The extracted data will be used for context-dependent lexical lookup or translation selection on any member language of Lexicon+TX, including under-resourced languages.

4.1.1 Latent Semantic Indexing

Based on the premise of distributional semantics that words that occur in the same contexts tend to have similar meanings (Harris, 1954), various vectorial representations have been designed to model word meanings (Salton, Wong, & Yang, 1975). Typically, a lexical meaning or concept is associated with a numerical vector $V = (v_1, v_2, \ldots, v_n)$, usually constructed based on the context of the word or concept in a corpus. The conceptual similarity between two lexical meanings, associated respectively with vectors U and V, is then the cosine similarity of U and V, or the cosine of the angle between them:[1]

[1] although cosine similarity is used here, other similarities or distances may also be used.

$$\begin{aligned}\mathrm{CSim}(U,V) &= \frac{U \cdot V}{|U| \times |V|} \\ &= \frac{\sum_{i=1}^{n} u_i v_i}{\sqrt{\sum_{i=1}^{n} (u_i)^2} \times \sqrt{\sum_{i=1}^{n} (v_i)^2}}\end{aligned} \tag{4.1}$$

Thus two items are said to be highly related if the angle between their vectors is small i.e. if they have a high CSim (cosine similarity) score.

While any vector model can be used, the latent semantic indexing (LSI) model (Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990) is adopted here, as it is robust in handling synonymy and polysemy (Deerwester et al., 1990). LSI uses singular value decomposition (SVD) to identify latent patterns in the terms and concepts in a text collection, including second-order co-occurrence patterns. What this means is that if «bank» and «economy» do not co-occur in a corpus, but each co-occurs with «finance», the LSI model will still be able to detect a relation between «bank» and «economy».

In LSI, an $m \times n$ term-document matrix M is first constructed, in which each row represents a document in the corpus, and each column position represents a term (word or lexical unit). Each element of the term-document matrix is the number of times a term occurs in a particular document. SVD is then performed on M, which will rewrite M as

$$M = U \Sigma V^T. \tag{4.2}$$

In SVD, the columns of U ($m \times r$ matrix) are m-dimensional vectors and known as the left singular vectors, while the columns of V ($n \times r$ matrix) are n-dimensional vectors and known as the right singular vectors. Σ is a $r \times r$ diagonal matrix. The singular vectors are eigenvectors of $M^T M$ and $M M^T$, while the values on the diagonal of Σ are the square roots of eigenvalues from $M^T M$ or $M M^T$.

When applied in LSI, each left singular vector represents a term, and each right singular vector a document in the corpus. It is also common to take only the first k elements of the term and document vectors in LSI, thus effectively reducing the large dimensions of the original term-document matrix M to k factors.

4.1.2 Translation Context Knowledge Acquisition as a Cross-Lingual LSI Task

In this work, translation context knowledge is modelled as a bag-of-words consisting of the context of a translation equivalence in the corpus. While LSI is usually used in IR systems, this task of translation knowledge acquisition can be recast as a cross-lingual indexing task, following the approach of Dumais, Littman, and Landauer (1997). The proposed approach makes use of a comparable corpus instead of a parallel aligned corpus, i.e. adopting a bag-of-words model. The underlying intuition is that in a comparable English–Malay corpus, a document pair about botany would be more likely to contain «plant»$_{\texttt{eng}}$ and «tumbuhan»$_{\texttt{msa}}$ (as opposed to «kilang»$_{\texttt{msa}}$ for the 'factory' meaning). The words appearing in this document pair would then be an indicative context for the translation equivalence between «plant»$_{\texttt{eng}}$ and «tumbuhan»$_{\texttt{msa}}$.

Given Lexicon+TX, a multilingual lexicon containing translation sets of languages $L_1, L_2, \ldots, L_N$ and a comparable corpus of languages L_i, L_j $(1 \leq i, j \leq N)$, the vector representing the translation knowledge (i.e. latent context information) of each translation set in Lexicon+TX is computed as follows:

1. Each bilingual pair of documents is merged as one single document, with each LI tagged with its respective language code.

2. Pre-process the corpus if necessary, e.g. remove stop words, lemmatise all words, perform word segmentation for languages without word boundaries (Chinese, Thai, etc).

3. Construct a term-document matrix, using the frequency of terms (each made up by a LI and its language tag) in each document. Apply further weighting if necessary.

4. Perform LSI on the term-document matrix. A vector is then obtained for every LI (in both languages) occurring in the comparable corpus.

5. Set the vector associated with each translation set to be the sum of all available vectors of all its member LIs. This sum vector then serves as a "bag-of-context" of all LIs in the translation set.

Note that if LSI was run on a monolingual corpus without sense-tags, the vector for a polysemous term e.g. «bank»$_{\texttt{eng}}$ would contain contexts applying to both the *financial institution* and *river side* meanings. In a bilingual corpus setting such as ours, however, the translation equivalents present serve as a kind of implicit sense-tagging.

As a demonstration, consider the small English–Malay comparable corpus in Table 4.1. A vector is obtained for each LI after running LSI with two factors[2] on the pre-processed corpus, as listed in Table 4.2. (The Java library EJML from `http://code.google.com/p/efficient-java-matrix-library/` was used for this indexing.)

Table 4.1: Small English–Malay bilingual comparable corpus.

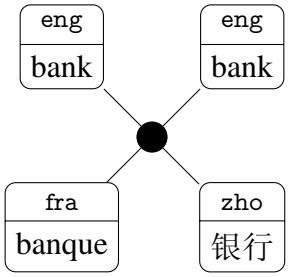
#	English	Malay
1	I deposited my salary with the bank	Saya memasukkan wang gaji saya di bank
2	You should only borrow money from a bank	Pinjam lah wang dari bank sahaja
3	Money lending activities	Aktiviti meminjam wang
4	We lazed by the river bank	Kami berehat di tepi tebing sungai
5	The river bank was soon inundated by the flood water	Tebing sungai dibanjiri air bah
6	We bathed in the cool river water	Kami bermandi-manda di tengah sungai

Now given two translation sets from Lexicon+TX, corresponding to the *financial institution* and *riverside* senses of «bank»$_{\texttt{eng}}$ respectively in Figure 4.1, their respective

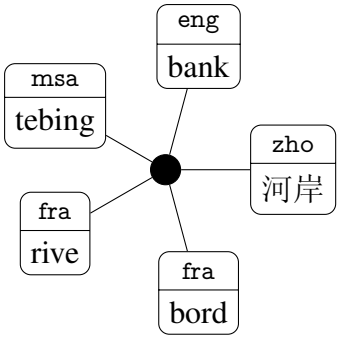
[2]i.e. the number of elements in each term and document vector will be capped at two. This capping is a common practice in LSI.

Table 4.2: Vectors of LIs after running LSI on the small corpus with 2 factors

Lang.	LI	Vector	Lang.	LI	Vector
eng	rest	(0.109, -0.007)	eng	deposit	(0.048, 0.169)
eng	river	(0.397, -0.148)	eng	soon	(0.178, -0.057)
eng	water	(0.288, -0.141)	eng	cool	(0.11, -0.084)
eng	bath	(0.11, -0.084)	eng	salary	(0.048, 0.169)
eng	bank	(0.386, 0.306)	eng	lend	(0.016, 0.112)
eng	inundate	(0.178, -0.057)	eng	flood	(0.178, -0.057)
eng	borrow	(0.051, 0.201)	eng	money	(0.067, 0.314)
msa	gaji	(0.048, 0.169)	msa	wang	(0.115, 0.482)
msa	bermandi-manda	(0.11, -0.084)	msa	tengah	(0.11, -0.084)
msa	memasukkan	(0.048, 0.169)	msa	sungai	(0.397, -0.148)
msa	tebing	(0.287, -0.064)	msa	tepi	(0.109, -0.007)
msa	air	(0.288, -0.141)	msa	berehat	(0.109, -0.007)
msa	sahaja	(0.051, 0.201)	msa	pinjam	(0.067, 0.314)
msa	bah	(0.178, -0.057)	msa	dibanjiri	(0.178, -0.057)
msa	bank	(0.099, 0.37)			



(a) Translation set TS_1 (bank as a financial institution)



(b) Translation set TS_2 (bank as riverside land)

Figure 4.1: Translation sets containing «bank»$_{\texttt{eng}}$

vectors can be computed as

$$V(TS_1) = V(\text{«bank»}_{\texttt{eng}}) + V(\text{«bank»}_{\texttt{msa}}) \quad = (0.484, 0.676) \tag{4.3}$$

$$V(TS_2) = V(\text{«bank»}_{\texttt{eng}}) + V(\text{«tebing»}_{\texttt{msa}}) = (0.672, 0.243). \tag{4.4}$$

The computed vectors for translation sets are added to Lexicon+TX. The next

section shows how these vectors are used for context-dependent multilingual lexical lookup, even when the text to be looked up is not of any languages of the indexed comparable corpus.

4.2 Context-Dependent Multilingual Lexical Lookup

For polysemous LIs, the lookup module should return translation sets that convey the appropriate meaning in context. In addition, the lookup module should also be able to recognise MWEs, which may occur as discontiguous strings in the input text.

The next subsections will first describe how LIs in an input text are matched, including detecting (possibly discontiguous) MWEs. We then present how the retrieved translation sets for each LI are ranked, based on the input context and translation knowledge vectors.

4.2.1 Matching Lexical Items in Input Text

Given an input text, modelled here as a sequence $S = w_1 w_2 \ldots w_n$ in language L, where each w_i is either a word token as delimited by word boundaries (for English, Malay, Italian, etc) or as produced by a word segmentation procedure (for Chinese, Japanese, German, etc), the LI-matching module should return a list of language L open class LIs found in the text S. It should be noted that LIs include MWEs, which may occur as discontiguous string sequences in S.

As an example, given the following sentence:

'He makes a meagre living planting sweet potatoes.'

the LI-matching module should return the list

{«make a living»$_{\text{V}}$, «meagre»$_{\text{A}}$, «plant»$_{\text{V}}$, «sweet potato»$_{\text{N}}$}.

Algorithm 3 returns a list of LIs present in a language L string sequence

$S = w_1 w_2 \ldots w_i \ldots w_n$, where each w_i is a word token as defined previously. The input tokens are POS-tagged and lemmatised (if applicable). A list of candidate LIs are retrieved from the lexicon, each of which contains at least one input lemma. The score of each candidate LI, c, is computed by taking the sum of squared lengths of longest common subsequences of c and the input lemmas that cover c. LIs containing longer continuous subsequences therefore receive a higher score. The algorithm returns the top ranking LIs that covers as many of the input lemmas as possible.

Algorithm 3 Finding list of LIs in string sequence $S = w_1 w_2 \ldots w_i \ldots w_n$

```
for all w_i do
    w'_i ← POS-tagged and lemmatised (if applicable) w_i
end for
InputTokens ← w'_1 w'_2 ... w'_i ... w'_n
Candidates ← all open-class LIs containing at least one w'_i ∈ InputTokens
for all c ∈ Candidates do
    subseqs ← longest common subsequences of c and InputTokens
    Score(c) ← Σ_{s∈subseqs} (length(s))^2
end for
Sort Candidates by descending Score(c) for c ∈ Candidates
repeat
    c ← pop(Candidates)
    if c ⊆ InputTokens, ignoring 'placeholder' elements in c then
        Add c to MatchedLis
        Delete c from InputTokens
    end if
until no more c ∈ Candidates such that c ⊆ InputTokens
Return MatchedLIs
```

Consider the earlier example input:

'He makes a meagre living planting sweet potatoes.'

POS-tagging and lemmatising (for English) gives the input tokens:

he$_{\text{PRON}}$ make$_{\text{V}}$ a$_{\text{DET}}$ meagre$_{\text{A}}$ living$_{\text{N}}$ plant$_{\text{V}}$ sweet$_{\text{A}}$ potato$_{\text{N}}$

Table 4.3 illustrates how Algorithm 3 ranks and selects open class LIs that best cover the input sentence. «Make a living» is successfully matched, even though it occurs discontiguously as *'make a ... living'* (Score $= 2^2 + 1^2 = 5$) in the input sentence. Similarly, «sweet potato» is chosen over «hot potato», «sweet» and «potato».

Table 4.3: Matching LIs in *'He makes a meagre living planting sweet potatoes'*

Candidate LI	Score	Matched	Remaining input tokens
make a living	$2^2 + 1^2 = 5$	**Y**	he ~~make~~ ~~a~~ meagre ~~living~~ plant sweet potato
sweet potato	$2^2 = 4$	**Y**	he meagre plant ~~sweet~~ ~~potato~~
hot potato	$1^2 = 1$	N	he meagre plant
make	$1^2 = 1$	N	he meagre plant
meagre	$1^2 = 1$	**Y**	he ~~meagre~~ plant
living	$1^2 = 1$	N	he plant
plant	$1^2 = 1$	**Y**	he ~~plant~~
sweet	$1^2 = 1$	N	he
potato	$1^2 = 1$	N	he

The matching algorithm will also match MWEs with 'placeholder' elements, typically marked as *'someone'*, *'something'*, *'one's'* and *'oneself'* in dictionaries, using the POS of the input tokens (e.g. an PRON input token matches both PRON and N 'placeholder' elements). Table 4.4 shows the matched LIs in the input

'He's not embarrassed to wash the family's dirty linen in public.'

Table 4.4: Matched LIs in *'He is not embarrassed to wash the famliy's dirty linen in public.'*

Candidate LI	Score	Remaining input tokens
wash one's dirty linen in public	$1^2 + 4^2 = 17$	he is not embarrassed to ~~wash~~ the family ~~dirty~~ ~~linen~~ ~~in~~ ~~public~~
embarrassed	$1^2 = 1$	he is not ~~embarrassed~~ to the family
family	$1^2 = 1$	he is not to the ~~family~~

4.2.2 Ranking Translation Sets in Context

Having determined $LW = \{l_1, l_2, \ldots, l_n\}$, the list of LIs present in a language L input text S, the translation selection module should then return a ranked list of multilingual translation sets for each LI $l_i \in LW$, particularly when l_i is polysemous. Algorithm 4 does this using the translation knowledge vectors computed in section 4.1.

Algorithm 4 Ranking translation sets for a given list of LIs, $LW = \{l_1, l_2, \ldots, l_n\}$

```
1:  V_Q ← zero vector                                  ▷ Compute the input 'query' vector
2:  for all l_i ∈ LW do
3:      if lookup(V(l_i)) ≠ null then
4:          V_Q ← V_Q + V(l_i)
5:      else
6:          TS_{l_i} ← getTransSets(l_i)
7:          for all t ∈ TS_{l_i} do
8:              V_Q ← V_Q + V(t)
9:          end for
10:     end if
11: end for
                                                       ▷ Rank translation sets containing each input LI
12: for all l_i ∈ LW do
13:     TS_{l_i} ← getTransSets(l_i)
14:     for all t ∈ TS_{l_i} do
15:         score(t) = CSim(V(t), V_Q)
16:     end for
17:     Output t ∈ TS_{l_i} by descending score(t)
18: end for
```

Briefly, the algorithm first computes a 'query' vector $V(Q)$ by summing up the translation knowledge vectors of all $l_i \in LW$. If no vector is found for l_i in Lexicon+TX, the sum of vectors associated with all translation sets containing l_i is used instead

(lookup($V(l_i)$) performs this check). For the selection phase, the list of all translation sets containing $l_i \in LW$, is retrieved into TS_{l_i}. The list of translation sets is then sorted in descending order of $\text{CSim}(t, V_Q)$ for all $t \in TS_{l_i}$ (see Equation 4.1).

As a quick demonstration, consider «bank»$_{\texttt{eng}}$, which could mean a financial institution (TS_1 in Figure 4.1(a)) or a riverside area (TS_2 in Figure 4.1(b)), in the running example with the small corpus in Tables 4.1 and 4.2. Recall also that the translation knowledge vectors for translation sets TS_1 and TS_2 were given in Equations (4.3) and (4.4) respectively:

$$V(TS_1) = (0.484, 0.676) \qquad \text{(from 4.3)}$$

$$V(TS_2) = (0.672, 0.243) \qquad \text{(from 4.4)}$$

Given the English input *'The bank lent me the capital'*, the algorithm computes:

$$\begin{aligned} V_Q &= V(\text{«bank»}_{\texttt{eng}}) + V(\text{«lend»}_{\texttt{eng}}) + V(\text{«capital»}_{\texttt{eng}}) \\ &= (0.402, 0.419) \\ \text{CSim}(V(TS_1), V_Q) &= 0.990 \\ \text{CSim}(V(TS_2), V_Q) &= 0.896. \end{aligned}$$

The algorithm therefore prefers TS_1 («bank»$_{\texttt{eng}}$ as a financial institution) over TS_2 for this particular input sentence. In other words, «bank»$_{\texttt{msa}}$, «银行»$_{\texttt{zho}}$ and «banque»$_{\texttt{fra}}$ are selected as the more likely translation equivalents in the respective TLs. Note that although «bank»$_{\texttt{eng}}$ does not co-occur with either «lend» or «capital» in the corpus (Table 4.1), the LSI-generated vectors are able to capture the latent relationship between them.

Conversely, given another input sentence *'He bathed near the bank'*, the algorithm computes:

$$
\begin{aligned}
V_Q &= V(\text{«bath»}_{\texttt{eng}}) + V(\text{«bank»}_{\texttt{eng}}) \\
&= (0.495, 0.222) \\
\text{CSim}(V(TS_1), V_Q) &= 0.864 \\
\text{CSim}(V(TS_2), V_Q) &= 0.997.
\end{aligned}
$$

This time, the algorithm selects TS_2 («bank»$_{\texttt{eng}}$ as riverside land) as the preferred translation set, thereby outputting «tebing»$_{\texttt{msa}}$, «河岸»$_{\texttt{zho}}$ and «bord»$_{\texttt{fra}}$ as the more likely translation equivalents. Again, notice that «bank»$_{\texttt{eng}}$ and «bath»$_{\texttt{eng}}$ do not co-occur in the bilingual comparable corpus.

4.3 Annotating Text with Links to Multilingual Lexicon Entries

For NLP applications, it would be desirable to have an annotation schema that can relate LIs in a text to lemma entries in a lexicon, particularly in cases where the LIs may manifest as discontiguous strings (i.e. syntactically flexible MWEs). In addition, the annotation schema should also be able to handle translational equivalence given a parallel text and a multilingual lexicon, where lexical gaps may cause an LI to be translated as a phrasal construction. The following sections will describe annotation schemas suitable for these purposes, including one which is newly proposed.

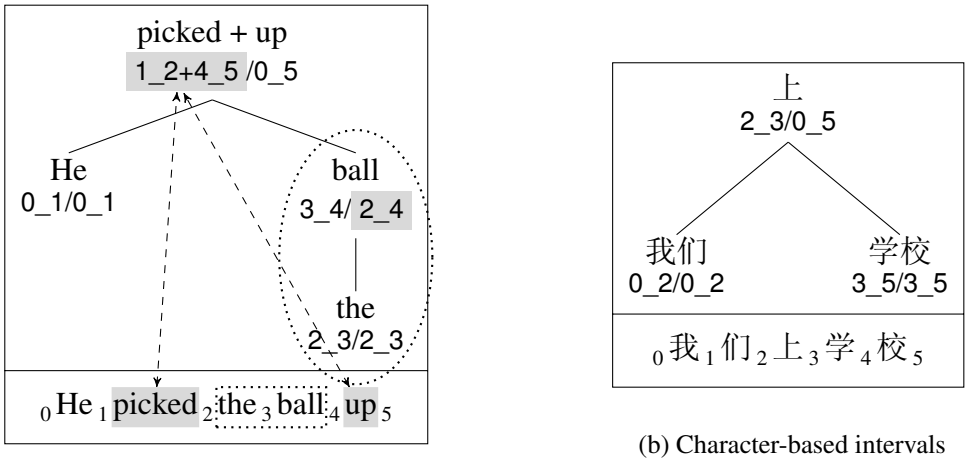
4.3.1 Structured String-Tree Correspondence

The Structured String-Tree Correspondence (SSTC) (Boitet & Zaharin, 1988) is an annotation schema for declaratively specifying multi-level correspondences between a string and its tree representation structure of arbitrary choice.

An SSTC comprises a string *st*, its tree representation structure *tr*, and the correspondences between them, *co*. (The formal definition is given in Appendix C.)

Substrings of S are identified by *intervals*, which serve as mechanisms for specifying the correspondences between *st* and *tr* on two levels:

- lexical level, i.e. between (possibly discontiguous) substrings of *st* and tree nodes of *tr*, using SNODE intervals; and
- phrase level, i.e. between (possibly discontiguous) substrings of *st* and (possibly incomplete) subtrees of *tr*, using STREE intervals.



(a) Word boundary-based intervals

(b) Character-based intervals

Figure 4.2: SSTCs with word boundary- and character-based intervals

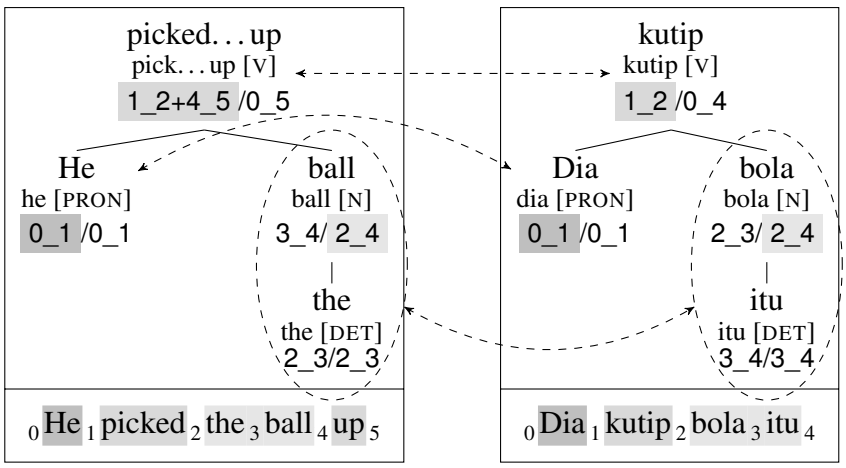
Intervals may be word boundary-based or character-based, depending on the writing or script system in use. For example, text in languages using the Latin script, such as English, might use a word boundary-based interval scheme, so in Figure 4.2(a), the interval 0_1 would indicate the substring 'he'; while 2_4 and 1_2+4_5 indicate 'the ball' and 'picked... up' respectively. Note how the former STREE interval relates the phrase 'the ball' to a subtree, and how the latter SNODE interval specifies the discontiguous substring ('picked... up') and relates it to a single node in the dependency tree structure.

On the other hand, when using a script without word boundaries (such as Chinese) or agglutinative languages (such as German), a character-based interval

scheme is used instead. An example is shown in Figure 4.2(b), where the substring '学校' is indicated by the interval 3_5.

The SSTC is a highly flexible structure, such that non-standard language phenomena, such as non-projectivity and ellipsis, can be captured declaratively. Its extension, the Synchronous SSTC (S-SSTC) schema (Al-Adhaileh, Tang, & Zaharin, 2002), consists of a pair of SSTCs. (The formal definition is given in Appendix C.) Figure 4.3 shows how S-SSTC can be used for annotating translation examples. The S-SSTC retains and extends the multi-level annotation flexibility, which is robust enough to declaratively describe complex and irregular correspondence phenomena, such as crossed dependencies and inverted dominance. See Appendix C for a full description of the SSTC and S-SSTC.

Due to such flexibility, both annotation schemas have applications in diverse NLP applications including MT (Al-Adhaileh et al., 2002; Boitet, Zaharin, & Tang, 2011), question answering (Song, Cheah, Tang, & Ranaivo-Malançon, 2008), speech synthesis (Sabrina, Rosni, & Tang, 2011) and recognition (Hong, Tan, & Tang, 2012).



SNODE correspondences		STREE correspondences	
(0_1, 0_1)	(1_2+4_5, 1_2)	(0_5, 0_5)	(0_1, 0_1)
(3_4, 2_3)	(2_3, 3_4)	(2_4, 2_4)	(2_3, 3_4)

Figure 4.3: An English–Malay translation example as an S-SSTC

4.3.2 SSTC+Lexicon

This section presents SSTC+Lexicon (SSTC+L), a proposed extension of the SSTC, for linking (possibly discontiguous) substrings in a text to corresponding items in an external repository, e.g. LI entries in a lexicon.

Formally, an SSTC+L is a tuple $(S, L, t_{S,L})$ where

- S is an SSTC,
- L is an external repository of items (e.g. a lexicon),
- $t_{S,L}$ is the set of correspondences between S and L.

The correspondence links $t_{S,L}$ between the SSTC S and the repository (or lexicon) L can be encoded by recording (X, w) where

- X is a sequence of SNODE or STREE intervals $\in co$ from S,
- w is the identifying key of item $w \in L$,
- w corresponds to the (possibly discontinuous) substring and (possibly incomplete) subtree from S indicated by X.

As a basic example, in the English sentence *'He made a meagre living planting sweet potatoes'* shown in Figure 4.4, the substring 'planting' (interval `5_6`) corresponds to the lexicon LI entry «plant»$_{\text{V}}$, while 'sweet potatoes' (interval `6_8`) corresponds the multi-word LI «sweet potato»$_{\text{N}}$. Note also if L is a multilingual lexicon (such as Lexicon+TX), substrings in the text then correspond to the multilingual translation sets, using the English LIs as access identifiers.

The SSTC+L schema is able to handle the annotation of syntactically flexible MWEs (section 2.2.3) and translational lexical gaps (section 2.2.2), as the following subsections demonstrate.

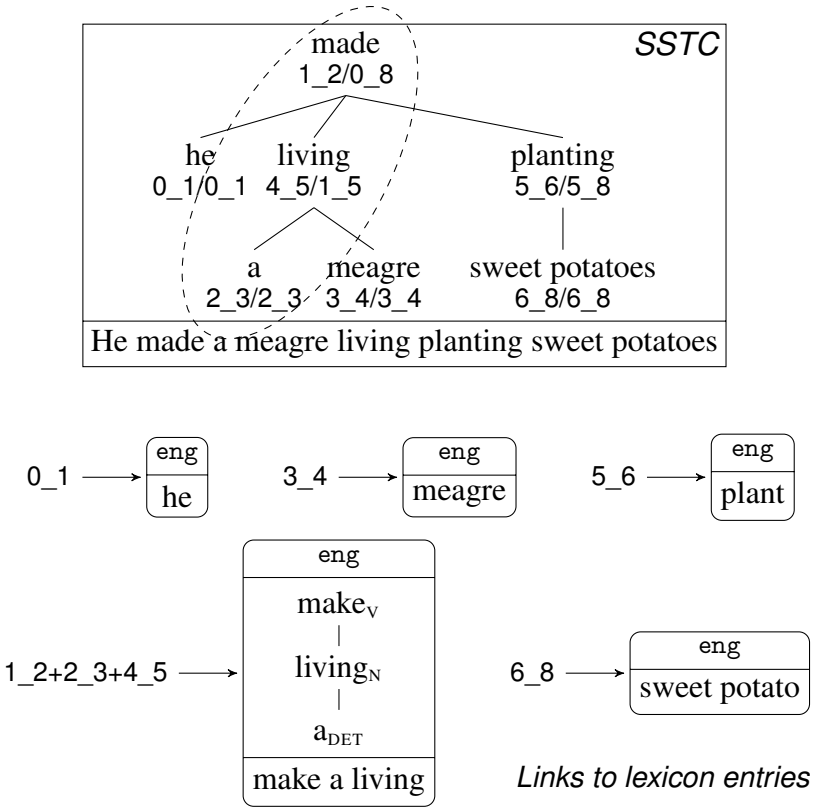


Figure 4.4: An SSTC+L relating LI occurrences in *'He made a meagre living planting sweet potatoes'* to lexicon entries

4.3.3 Discontiguous and Syntactically-Flexible MWEs

As described in section 2.2.3, MWEs exhibit a wide range of syntactic flexibility (Sag et al., 2002). This presents some problems when annotating their occurrences in corpora, so that they may be properly consumed by NLP systems. This section shows how the SSTC+L can be used to handle such MWEs.

Figure 4.4 contains an example of an occurrence of a syntactically flexible MWE, where the English saying «make a living» occurs in the sentence 'He made a meagre living…' as a discontiguous string segment. Modelling the text as an SSTC+L, the dependency tree captures «meagre» as syntactically modifying «living», while the SNODE interval 1_2+2_3+4_5 links the discontiguous string 'made a … living' to the lexicon entry for LI «make a living». The SSTC+L therefore successfully captures «make a living» as an LI (by identifying and relating it to a lexicon entry using SNODE

intervals), as well as a flexible MWE construction, where an adjective is allowed to modify one of its elements.

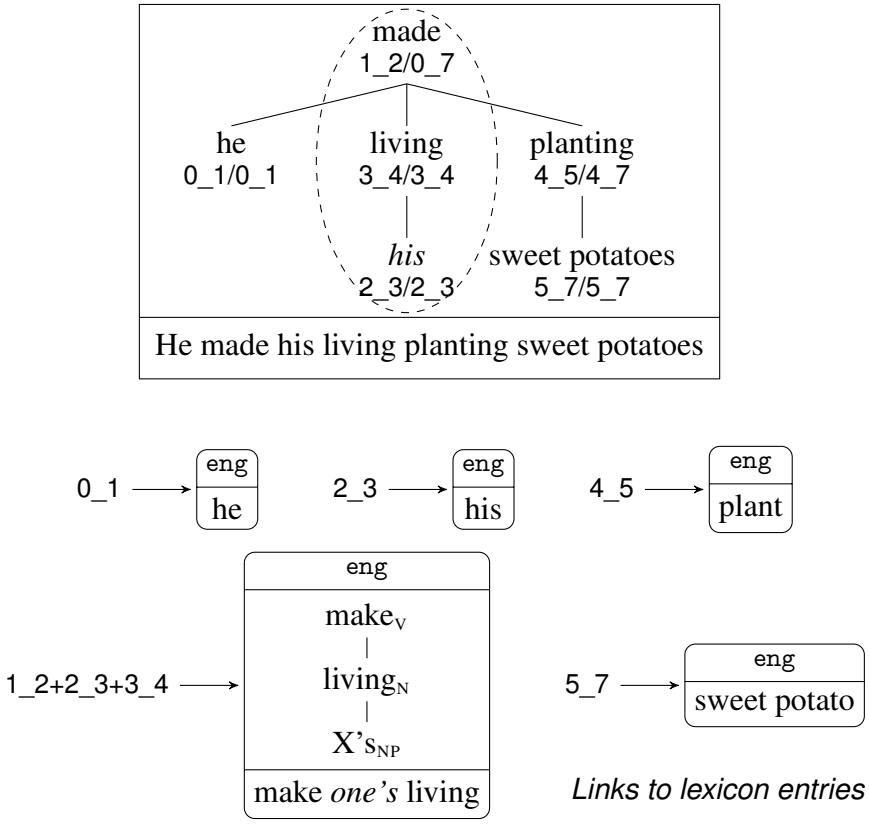


Figure 4.5: An SSTC+L containing an MWE with a 'placeholder'

Figure 4.5 demonstrates a similar scenario, but one which involves an MWE with a 'placeholder' variable, i.e. «make *one's* living». The SNODE interval mechanism again plays its role in relating the lexicon's LI entries to their occurrences in the text, whose syntactic and dependency structure is captured accurately by the tree structure in the SSTC.

Finally, the SSTC+L schema is especially useful for relating MWEs of high syntactic flexibility, e.g. those that can be passivised, to their canonical lemma form in a lexicon. An example is shown in Figure 4.6, where the passive construction 'the beans are spilt' corresponds to the LI «spill the beans».

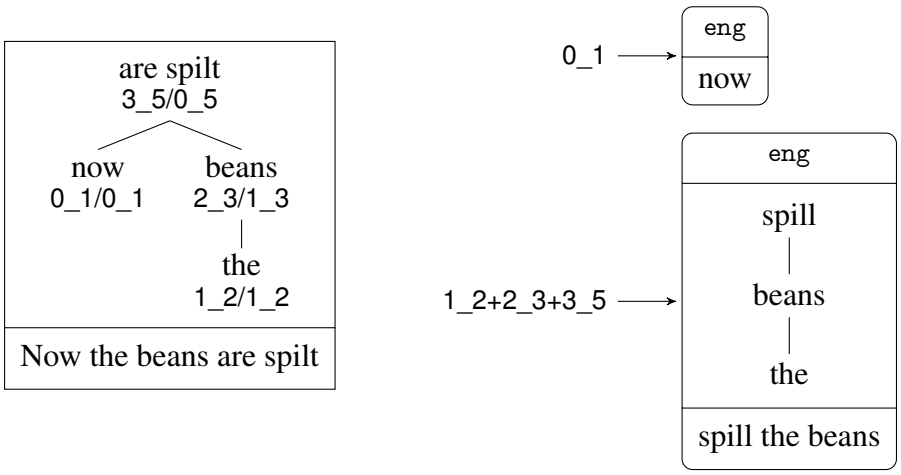
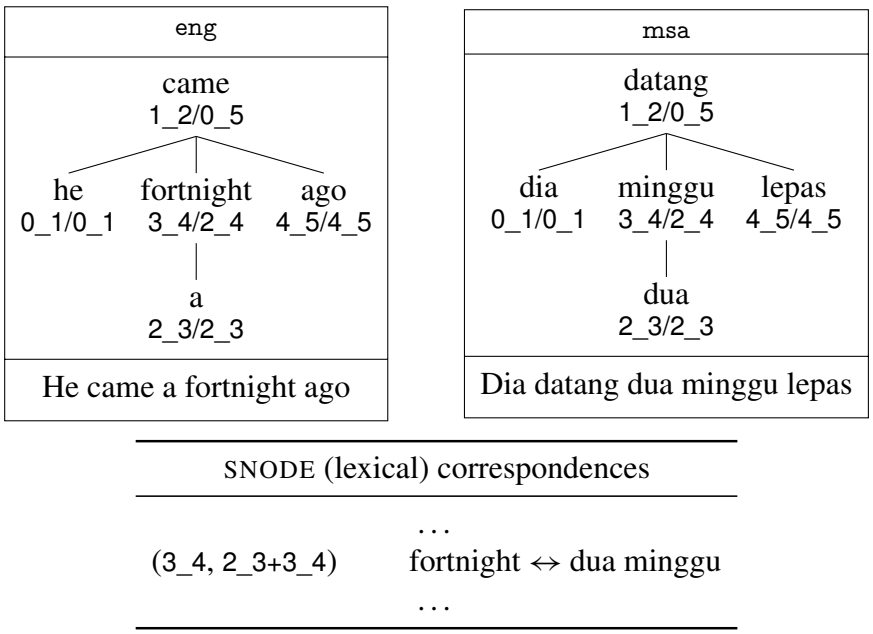


Figure 4.6: An SSTC+L relating a passivised MWE to its canonical lexicon entry

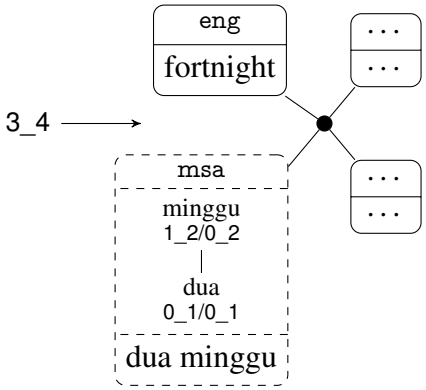
4.3.4 Annotating Lexical Gaps in Translation Examples

Lexical gaps occur when an LI in a source language (SL) is not lexicalised in a target language (TL), and therefore have to be translated as a gloss-like phrase (sections 2.2.2 and 3.1.1 (c)). In Figure 4.7(a), the S-SSTC captures that English 'fortnight' is translated to 'dua minggu' in Malay via the SNODE correspondence (3_4, 2_3+3_4). However, from the Malay monolingual, lexical point of view, there is no way to tell if 'dua minggu' here is a valid Malay LI (as a MWE), or a phrasal construction for translating 'fortnight' because of a lexical gap in Malay.

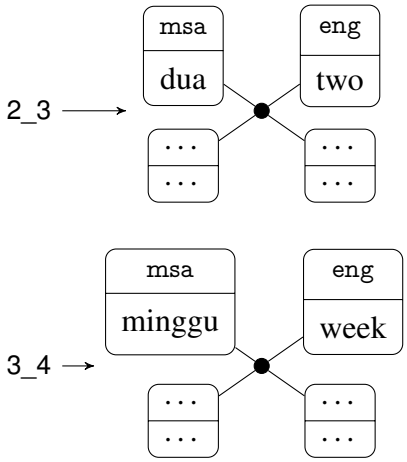
This can be remedied by adding SSTC+L structures to our annotation collection. For the English segment, the SSTC+L (Figure 4.7(b)) contains a link from the SNODE interval 3_4 to the multilingual translation set containing the LI «fortnight»$_{\texttt{eng}}$, which also contain 'dua minggu'$_{\texttt{msa}}$ as a translation equivalent member in the form of a gloss-like phrasal construction. On the other hand, from the Malay segment (Figure 4.7(c)), «dua»$_{\texttt{msa}}$ and «minggu»$_{\texttt{msa}}$ are considered as distinct LIs. Therefore the SSTC+L for the Malay segment contains two separate entries for the SNODE intervals 2_3 and 3_4, to translation sets containing «dua»$_{\texttt{msa}}$ and «minggu»$_{\texttt{msa}}$ respectively. Thus by using the S-SSTC and SSTC+L annotation schemas in tandem, translation phenomena between two text segments can be captured declaratively, while maintaining the lexicality of each language.



(a) English–Malay S-SSTC relates 'fortnight' to 'dua minggu' as translation equivalents, but does not indicate if both are LIs in their respective languages



(b) SSTC+L for English segment



(c) SSTC+L for Malay segment

Figure 4.7: Annotating lexical gaps

4.4 Summary and Conclusion

This chapter has described how, given a coarse-grained, 'shallow' multilingual lexicon, a context-dependent multilingual lexical look-up module can be built, one which benefits even under-resourced languages. Comparable bilingual corpora, which are more readily available than aligned parallel corpora, are used to extract distributional information about the context of translation equivalents. This information, in the form of

numerical vectors, acts as a form of translation context knowledge for the multilingual translation sets in Lexicon+TX. Under-resourced language member LIs in the translation sets therefore benefit from the richer-resourced languages (i.e. those of the comparable corpus), which would otherwise lack any usable data to support translation selection. This translation context knowledge is then used to perform context-dependent lexical lookup on new input texts.

A new annotation schema, the SSTC+L, was also proposed for marking up LI occurrences in natural language text, together with the links to their canonical lemma entries in a given lexicon. Examples have been given to demonstrate how the SSTC+L is capable of handling syntactically flexible MWEs, as well as annotating translational lexical gaps effectively when used in tandem with the S-SSTC.

CHAPTER 5

IMPLEMENTATION RESULTS AND DISCUSSION

This chapter presents our implementation and experimental results based on the design and algorithms described in Chapters 3 and 4. Specifically, a prototype of Lexicon+TX comprising six languages (English, Malay, Chinese, French, Iban and Thai) has been constructed from six bilingual and one trilingual dictionaries. Translation sets in Lexicon+TX were then enriched with vectors obtained by running LSI on an English–Malay comparable corpus, extracted from Wikipedia articles. These data were then used to implement a context-dependent multilingual dictionary lookup tool.

The inclusion of Iban, an under-resourced ethnic Bornean language with 600 000 speakers (Ethnologue, 2012), demonstrates the suitability of the methodologies proposed in previous chapters for under-resourced languages. 91.2 % of 500 random multilingual entries in Lexicon+TX require minimal or no human correction. Lexicon+TX was enriched with translation context knowledge extracted from bilingual comparable corpus (e.g. Wikipedia articles), so that it may provide context-dependent lexical lookup purposes. The ranked multilingual translation sets returned by the lookup module in the evaluation achieved a precision score of 0.650 and a mean reciprocal rank score of 0.810.

Four experiments, including the two results mentioned above, were conducted to evaluate different aspects of the proposed framework, as shown in Figure 5.1 and Table 5.1 and described further in the following sections. Note that since the work described here involved multilingual lexicons and under-resourced languages, benchmark test data was not available for all evaluations. Instead, the results obtained are compared to those achieved in state-of-the-art related work.

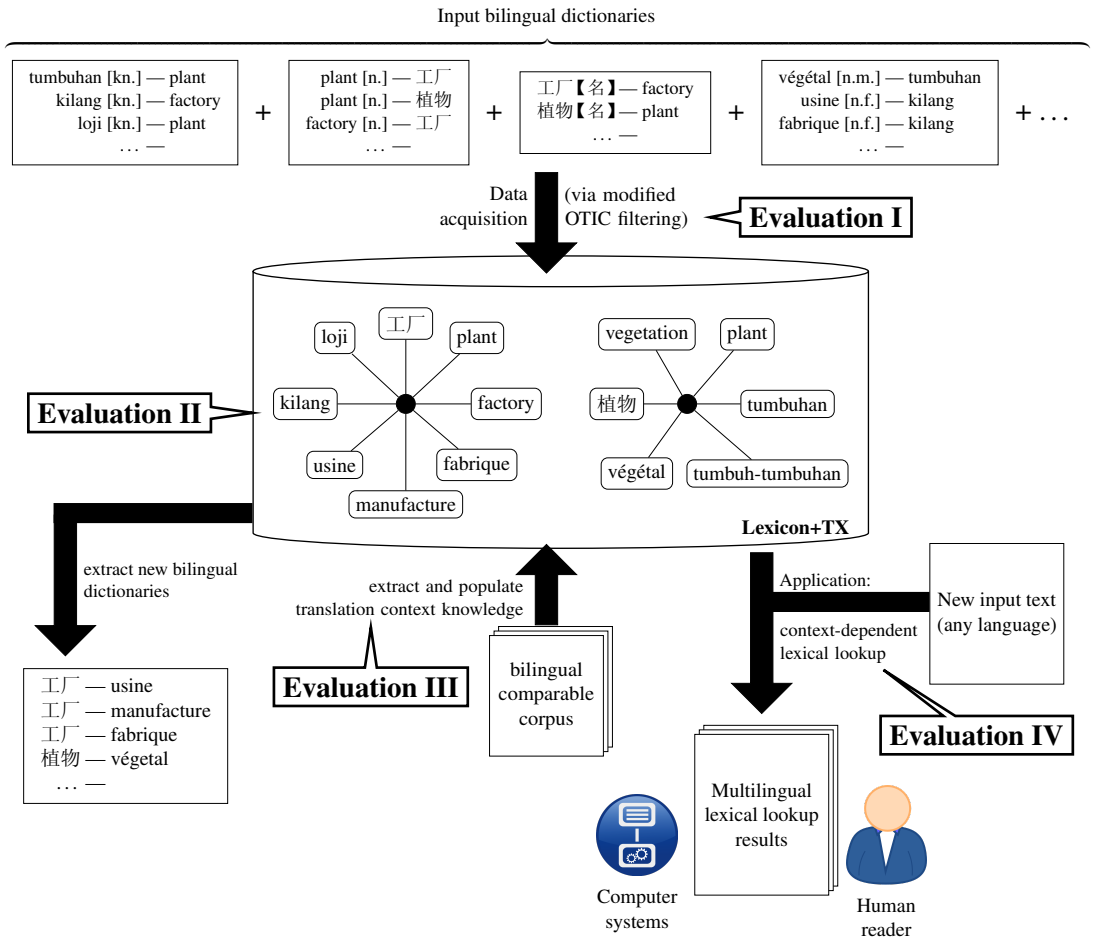


Figure 5.1: Evaluations on proposed framework

Table 5.1: Evaluations on proposed framework

Eval.	Description	Benchmark	Notes
I	Modified OTIC filtering	—	Compared to related work
II	Merged translation sets	—	Compared to related work
III	Translation context knowledge	WordSim-353	Word similarity score test
IV	Context-dependent lexical lookup	—	Compared to related work

5.1 Lexicon+TX Construction using Bilingual Dictionaries

The multilingual lexicon construction methodology proposed in Chapter 3 has been implemented in Java. A detailed manual for using the implemented Java tools, is attached in Appendix D. Specifically, the manual contains a step-by-step account of how Thai can be added to an existing Lexicon+TX which already contains English, Malay, Chinese and French; and how a new Thai–French bilingual dictionary can be

extracted at the end. Tools and instructions for running modified OTIC have been made available at `https://bitbucket.org/liantze/lexicontx`.

The following dictionaries were used as input, choosing open-source and free options wherever possible:

- SiSTeC-EMDict: Part of the SiSTeC-EBMT machine translation system (Boitet et al., 2011). 94 604 Malay items; 82 342 English items. Used here as a Malay–English dictionary.
- Kamus Inggeris-Melayu Dewan (KIMD) (Johns, 2000): 37 618 English items, 56 368 Malay items.
- XDict:[1] Open source. 177 799 English items, 194 571 Chinese items.
- CC-CEDICT:[2] Open source. 93 847 Chinese items, 107 228 English items.
- FeM:[3] Available for research. 28 288 French items, 23 148 English items, 41 519 Malay items.
- Handy Reference Dictionary of Iban and English (HRIE) (Sutlive & Sutlive, 1992): 9825 Iban items, 14 201 English items.
- Yaitron:[4] Open source, 32 347 Thai items, 22 660 English items.

Translation triples were generated (Table 5.2) and later aggregated using the modified OTIC procedure, to gradually build up Lexicon+TX that eventually comprise English, Malay, Chinese, French, Iban and Thai member LIs.

As an implementation detail, even though POS information is not a compulsory requirement of the OTIC process, POS filtering was applied during the alignment

[1] `http://packages.debian.org/sid/text/dict-xdict`

[2] `http://cc-cedict.org/wiki/start`

[3] `http://www-clips.imag.fr/cgi-bin/geta/fem/fem.pl`

[4] `https://github.com/veer66/Yaitron`

Table 5.2: Generated translation triples for expanding Lexicon+TX

Triples	Input dictionaries	New language added
Malay–English–Chinese	SiSTeC-EMDict, Xdict, CC-CEDICT	Malay, Chinese, English
French–English–Malay	FeM	French
Iban–English–Malay	HRIE, KIMD, SiSTeC-EMDict	Iban
Thai–English–Chinese	Yaitron, Xdict, CC-CEDICT	Thai

of translation pairs from input dictionaries, as this helped to eliminate many trivial alignment errors. However, it was found that many common LIs that can be both nouns and adjectives (e.g. «red») are only listed as nouns in some dictionaries, and only as adjectives in others. This caused many translation pairs to fail to match. To remedy this, adjectives were allowed to match nouns (and *vice versa*) during the modified OTIC process.

MWEs were run through a parser (Stanford parser), which generated the SSTC-annotated tree representation for the microstructure. The parser includes morphological information in its output, which is included in the SSTC annotations.

5.1.1 Lexicon+TX Prototype

Lexicon+TX was implemented as a MySQL relational database, the simplified schema of which is shown in Figure 5.2. It is populated with translation sets generated from the modified OTIC process, using input dictionaries listed in the previous section. All open source input dictionaries, together with tools for running modified OTIC to build Lexicon+TX, are available at `https://bitbucket.org/liantze/lexicontx`.

Figure 5.3 shows an example generated translation set, and Table 5.3 shows the number of new target languages that LIs in each source language are connected to. Note that since the Iban–English dictionary contained fewer entries compared to other input dictionaries, the number of LIs connected to all five other languages are therefore limited. Currently Lexicon+TX contains about 46 000 English MWEs.

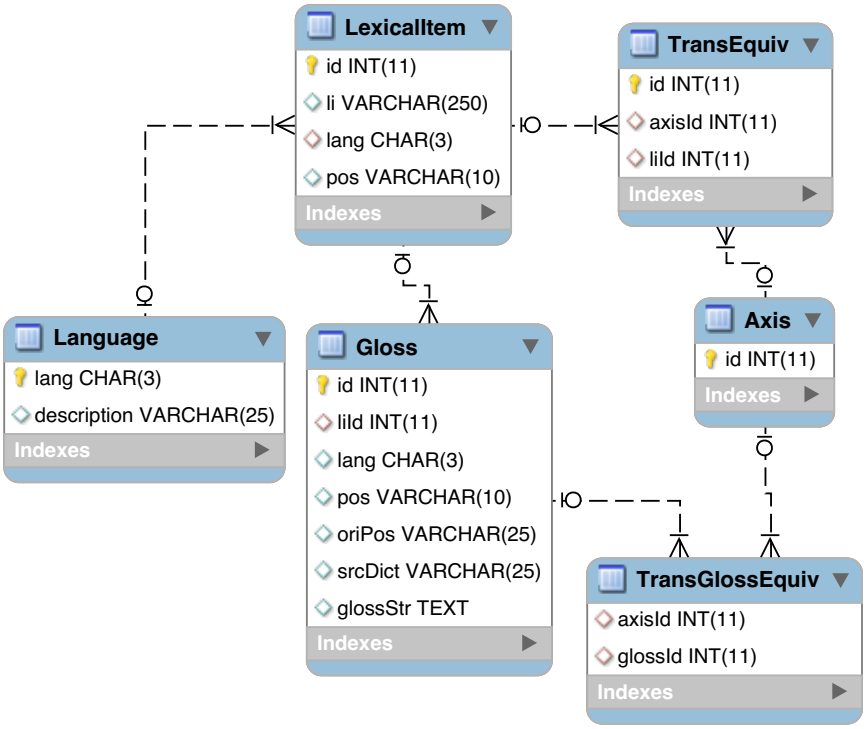


Figure 5.2: Simplified schema of Lexicon+TX relational database

#8795

English
- rainbow (LI#240974[N])

Bahasa Melayu
- pelangi (LI#55687[N])

中文
- 彩虹 (LI#331638[N])

français
- arc-en-ciel (LI#405617[N])

Bahasa Iban
- anakraja (LI#455757[N])
- emperaja (LI#457867[N])

ไทย
- รุ้ง (LI#529562[N])
- รุ้งกินน้ำ (LI#529563[N])
- สายรุ้ง (LI#532641[N])
- อินทรธนู (LI#536019[N])

Figure 5.3: Example generated translation set containing 6 languages

Bilingual dictionaries between any pairings of the above-mentioned six languages are now available from Lexicon+TX. In particular, Iban, an under-resourced language with 600 000 speakers in Borneo[5] is now connected to French, Thai and Chinese with relatively minimal effort and cost (albeit with precision trade-offs), all of which are rare language pairings. Currently, advanced lexical resources, such as

[5] http://www.ethnologue.com/show_language.asp?code=iba

Table 5.3: Number of Lexicon+TX LIs connected to other languages

Source Language	No. of LIs with translations in multiple languages			
	≥ 2 langs.	≥ 3 langs.	≥ 4 langs.	5 langs.
English	24 371	11 244	7696	3912
Chinese	13 226	9023	6044	2774
Malay	35 640	14 987	9919	5053
French	17 063	7383	5609	3363
Iban	5629	5101	4294	3580
Thai	14 687	13 037	10 883	6587

Table 5.4: Lexicon+TX type and token coverage of 500 English and Malay Wikipedia articles

Language	Total tokens	Token coverage (%)	Total types	Type coverage (%)
English	892 224	804 184 (90.1)	70 238	31 630 (45.0)
Malay	206 682	156 105 (75.5)	33 650	12 689 (37.7)

wordnet systems and domain code labels, or even a well-sized corpus for the Iban language are still lacking or in development (Yeo, Suhaila, & Wilfred, 2008). Therefore, many of the reviewed multilingual lexicon construction methods in Chapter 2 cannot be used. Using the proposed low cost method, however, Iban is now successfully connected to five other languages using just an Iban–English and two other simple bilingual dictionaries. Note that since the Iban–English dictionary contained a smaller number of entries compared to other input dictionaries, the number of LIs connected to all five other languages are therefore limited.

To gauge the coverage of Lexicon+TX, 500 English articles and 500 Malay articles were downloaded from Wikipedia. The total number of lemmatised tokens and types in each language were then counted, as well as the coverage of Lexicon+TX entries. The results are summarised in Table 5.4. In addition, Lexicon+TX contains 5078 (92.9 %) of the 5464 most frequent English lemmas in the British National Corpus (Kilgariff, 1996).

5.1.2 Evaluation I: Evaluating OTIC Filtering

As there is no widely accepted method for evaluating generated multilingual lexicons (Varga et al., 2009), two common metrics in IR and NLP tagging tasks, precision and recall, are used here.

For evaluation purposes, 500 random Malay–Chinese and Iban–Malay translation pairs generated from OTIC (before filtering) were extracted. There were graded by human evaluators as *accept*, *reject* or *unsure*. The gold standard was then obtained by taking the majority vote to reach an *accept* or *reject* verdict for each translation pairing. An *accept* verdict is assumed in case of a tie.

OTIC filtering was then run with varying threshold parameters. The precision, recall and harmonic mean (F_1) scores of the OTIC filtering, as compared to the gold standard, are computed as:

$$\text{Precision} = \frac{tp}{tp + fp} \tag{5.1}$$

$$\text{Recall} = \frac{tp}{tp + fn} \tag{5.2}$$

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5.3}$$

$$\text{where } tp = \text{true positive}, \qquad fp = \text{false positive},$$

$$tn = \text{true negative}, \qquad fn = \text{false negative}.$$

Note that because the level of overlap between dictionaries depends on the sets of dictionaries used, the precision and recall can vary for different language pairs and input dictionaries. Detailed results from the OTIC filtering decisions using different threshold parameter values, as well as the human decisions leading to the gold standard, can be found in Appendix E. The best precision and F_1 score achieved are shown in Table 5.5, with the corresponding precision and recall in parentheses.

While a higher precision score from higher filter threshold parameters is undoubtedly desirable, this would also mean a lower recall as more translation triples (and

Table 5.5: Best precision and F_1 scores achieved by OTIC in filtering Malay–Chinese and Iban–Malay translation pairs

Translation pairs	Best precision (recall)	Best F_1 (precision/recall)
Malay–Chinese	0.770 (0.380)	0.725 (0.636/0.843)
Iban–Malay	0.565 (0.354)	0.660 (0.492/1.000)

Table 5.6: Precision comparison with related work

Cited work	Precision	Resources used
Proposed method	0.77	Translation lists
Sammer and Soderland (2007)	0.73	Translation lists, monolingual corpora
Varga et al. (2009)	0.79	Translation lists, WordNet

hence equivalence links) are rejected. Some trade-off between precision and recall is therefore required in determining the filter threshold parameters, to ensure the multilingual translation sets are sufficiently accurate and contain a reasonable number of LIs, as indicated by that F_1 score. The threshold parameters that yield the best F_1 scores for each language pair is used to generate the final translation triples and translation sets.

Table 5.6 compares the results achieved by the proposed method with two related works on aligning translation pairs while maintaining the senses. Sammer and Soderland (2007) generated English–Spanish–Chinese sets using bilingual dictionaries and monolingual corpora. (Their method is considered low cost as monolingual corpora are more readily available.) Varga et al. (2009) generated Japanese–English–Hungarian sets using bilingual dictionaries and the English WordNet, which may not be applicable for under-resourced languages due to the WordNet requirement. Note again that the numbers reported in this table may not be suitable for comparative evaluation due to differences in the experiment methodology and language differences. Rather, the performance of the two related work are cited here to provide a context. As the table shows, the proposed method performed quite favourably, especially in view of the richness of resource types used in each work.

An unexpected outcome from this exercise was the relatively short time the

evaluators took for grading the translation pairs. Although they were initially asked to evaluate only 100 pairs each, most of the evaluators took about 2–4 hours to return their decisions for all 500 pairs. The evaluators' decisions can be used to purge erroneous translation triples from Lexicon+TX immediately (see Figure 3.9).

5.1.3 Evaluation II: Evaluating Translation Sets

500 translation sets were randomly extracted from Lexicon+TX and manually evaluated. Due to limitations of the evaluator's linguistics capabilities, only the English, Chinese and Malay members of each translation set are considered. Each translation set is given a score of 0 to 3 depending on the amount of work required to improve it. The summarised results are shown in Table 5.7, while the evaluated translation sets are attached in Appendix F. As the table shows, 91.2 % of the translation sets (score ≥ 2) require minimal or no correction, with an overall average score of 2.57.

Table 5.7: Satisfaction score of 500 randomly selected translation sets

Score	Description	No. of sets	(%)
3	No further work needed	365	73.0
2	Minor correction: delete errant LIs	91	18.2
1	Major correction: regroup into multiple translation sets	10	2.0
0	Bad: unintelligible translation set, discard	34	6.8
	Total	500	100.0

Table 5.8: Comparison of precision of merged translation sets with related work

Cited work	Precision	Resources used
Proposed method	0.73	Translation lists
Sammer and Soderland (2007)	0.20	Translation lists, monolingual corpora
Mausam et al. (2009)	0.90	Pre-existing sense-distinguished multilingual lexicons

Table 5.8 compares the precision of the translation sets from Lexicon+TX to multilingual lexicons generated by other related work. Here, the precision metric only counts translation sets in which all member LIs indicate the same meaning, i.e. entries

with score 3 in Table 5.7. The low precision score of Sammer and Soderland's (2007) method is mainly due to many semantically related words that are not synonyms being included in the same translation set, e.g. «bullet» and «shot». The multilingual lexicon produced by Mausam et al. (2009) graph-walking algorithm has a very high precision, but their work actually involved merging *pre-existing* sense-distinguished multilingual lexicons (crowd-sourced Wiktionaries), in which the presence and coverage of under-resourced languages are not guaranteed (see section 1.3, p. 6). In contrast, the proposed modified OTIC method attempts to build a sense-distinguished multilingual lexicon from unaligned bilingual dictionaries, including for under-resourced languages.

5.1.4 Discussion

Almost all errors were due to the presence of a polysemous 'mid' language LI in the translation set, which may cause a 'tail' language LI to be connected to the 'head' language LI erroneously during the OTIC process. While errors can be reduced (i.e. raising the precision) by increasing the OTIC filtering threshold parameters, this would also entail a lower recall as more translation triples (and hence equivalence links) are rejected. Some trade-off between precision and recall is therefore required in determining the filter thresholds to ensure the multilingual dictionary is sufficiently accurate and contains a reasonable number of LIs. In addition, as the amount of overlap differs for different sets of input dictionaries (due to coverage, translations given in the input dictionaries etc), it is essential to determine the optimum filtering threshold parameters when running OTIC for different language pairs and triples (as had been done for Malay–Chinese and Iban–Malay in Table 5.5).

The OTIC procedure relies on the input bilingual dictionaries to have a sufficient number of overlaps between the LIs and glosses of the same language. The choice of input dictionaries is therefore important to ensure there is as much overlap as possible. For example, one may want to confirm if American spelling or British spelling is adopted for English entries in all chosen dictionaries; or Chinese dictionary published in Singapore and one published in China may contain very different LIs for the same meaning. (For example, a computer is more likely to be listed as «电脑» in a Singapore-

published dictionary, and as «计算机» in a China-published one.) In such scenarios, many translation pairs would fail to match, resulting in a multilingual lexicon with few translation equivalence links. This, as a consequence, leads to a drawback of the proposed method: the number of acquired translation equivalence links are constrained by the degree of overlap between the input dictionaries (see also Table 5.3).

Overall, the results are highly satisfactory, considering the simplicity of the input data required: see Tables 5.6 and 5.8. Specifically, the proposed modified OTIC procedure provides a fast, cheap and effective way for generating a first draft of a multilingual lexicon, which will then be improved by human evaluators. The method requires only simple bilingual translation lists as input data, and is therefore suitable for under-resourced languages (e.g. Iban). The construction process is fast, taking a little under 30 minutes to add a new language on MacBook Pro with a 2.3 GHz processor and 4 GB RAM. In addition, the 'shallow' model of the multilingual translation equivalence of Lexicon+TX means no linguistics expertise requirement is imposed on potential human evaluators.

One drawback of the proposed method is that the number of acquired translation equivalence links are constrained by the degree of overlap between the input dictionaries. Table 5.3 shows that as the number of TLs increases, the number of LIs having translations in all other TLs decreases. In addition, since the Iban–English dictionary contains far less entries than the other input dictionaries, the number of LIs with translations in all 5 TLs are limited.

Currently, diversification nodes and links need to be manually created and updated in Lexicon+TX. Methods for automatically acquiring diversification links is left as future work (section 6.4.1 (b)).

5.2 Context-Dependent Lexical Lookup using Translation Context Knowledge

Translation context knowledge is considered as a bag-of-words model, and is acquired for Lexicon+TX translation sets by running LSI on a bilingual comparable

corpus constructed from Wikipedia articles, as outlined in section 4.1. To rank translation sets for each LI in an input sentence, the cosine similarity between the translation set vectors and the 'query vector' of the input sentence is computed (section 4.2.2).

5.2.1 Corpus Preparation and Indexing

Wikipedia articles are freely available under a Creative Commons license, thus providing a source of bilingual comparable corpus. Malay Wikipedia articles[6] and their corresponding English articles of the same topics[7] were first downloaded.

To form the bilingual corpus, each Malay Wikipedia article is concatenated with its corresponding English Wikipedia article as one document. Words in the English articles are lemmatised, and stop words in both English and Malay articles are discarded. Malay morphological affixes, such as 'di-' and '-nya', are also discarded. Multiple words constituting the anchor text of a URL are grouped as a single term. For example, in the following snippet:

> ... life on the Malay archipelago dates back ...

«Malay archipelago» is regarded as a single item, instead of two separate items «Malay» and «archipelago».

The term-document matrix constructed from this corpus contains 62 993 documents and 67 499 terms, including both English and Malay items. The term-document matrix is weighted by term frequencey–inverse document frequency (TF-IDF), then processed by LSI using the Gensim Python library.[8]. The indexing process, using 1000 factors, took about 45 minutes on the same MacBook Pro mentioned previously.

[6] from `http://dumps.wikimedia.org/mswiki/`, retrieved 1 August 2011

[7] via the interface at `http://en.wikipedia.org/wiki/Special:Export`, on 3 August 2011

[8] `http://radimrehurek.com/gensim/` Gensim is used here instead of EJML, the Java library mentioned in section 4.1.2, because Gensim's numerical iterative approach could better handle large matrices using constant memory footprint.

A vector was thus obtained for each English and Malay LI appearing in the term-document matrix.These vectors were then used to populate the translation context knowledge vectors for each Lexicon+TX translation set, as described in section 4.1.

5.2.2 Evaluation III: Vector Similarity Score Evaluation

As a preliminary evaluation of the vectors obtained from LSI, a conceptual similarity experiment was conducted. The WordSim-353 benchmark (Finkelstein et al., 2002) contains human-assigned similarity scores of 353 pairs of English words, and has been widely used to evaluate lexical similarity measures.

In this experiment, the similarity score for a pair of words is defined to be the cosine similarity of their vectors (Equation 4.1) from the LSI indexing. The cosine similarity for the 353 pairs of words from WordSim-353 was computed with different numbers of LSI factors (Appendix G). The Spearman's ρ correlation coefficient with the benchmark scores was then computed and summarised in Table 5.9 and Figure 5.4. The highest Spearman's ρ correlation with the WordSim-353 benchmark achieved by our vector cosine similarity score is 0.629 using 600 factors. To give some context, a comparison to Agirre et al.'s (2009) state-of-the-art work on computing word similarity scores using corpus- or distributional-based approaches is given in Table 5.10.

Table 5.9: Correlation of LSI vector cosine similarity with WordSim-353 benchmark

No. of Factors	Spearman's ρ Correlation
300	0.612
400	0.618
500	0.625
600	0.629
700	0.616
800	0.604

In their work, Agirre et al. (2009) reported 0.66–0.69 Spearman's ρ correlation with WordSim-353, running a distributional-based algorithm on 4 billion Web

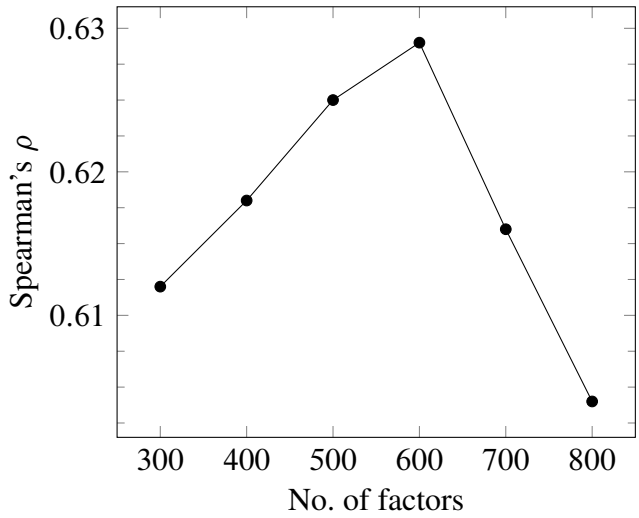


Figure 5.4: Correlation of LSI vector cosine similarity with WordSim-353 benchmark

Table 5.10: Comparison of Spearman's ρ correlation with WordSim-353 benchmark to related work

Cited work	Best Spearman's ρ	Document size	No. of processors	Time
Proposed method	0.63	62 993 documents, 67 499 words	1	45 minutes
Agirre et al. (2009)	0.69	4 billion documents, 1.5 Terawords	2000	15 minutes

documents containing 1.6 Terawords, using 2000 processing cores in 15 minutes. In comparison, the results obtained by the proposed method fare favourably, especially considering the simple input data and processing resources used.

5.2.3 Evaluation IV: Context-Dependent Lexical Lookup

A simple context-dependent lexical lookup tool, LEXICALSELECTOR, was developed following the design outlined in section 4.2. Figure 5.5 shows the top ranked translation sets output by LEXICALSELECTOR for the input English sentence 'The plant has its own generator for electricity'. Here, the 'factory' meaning of «plant» was ranked higher than the 'vegetation' meaning (not shown in the figure).

```
plant [1_2]
zho:  工厂,
tha:  ร.ง., โรงงาน,
fra:  fabrique, manufacture, usine,
msa:  loji, kilang,


own [4_5]
zho:  自己, 自我,
tha:  ถือกรรมสิทธิ์, ถือครอง, อัตตา,
msa:  diri, kendiri, sendiri,


generator [5_6]
zho:  发电机,
tha:  เครื่องกำเนิดไฟฟ้า,
msa:  penjana, generator, janakuasa,


electricity [7_8]
zho:  电力,
tha:  ไฟฟ้า,
fra:  électricité,
msa:  kuasa elektrik,
```

Figure 5.5: Top translation sets selected by LEXICALSELECTOR for 'The plant has its own generator for electricity.'

LEXICALSELECTOR can also detect MWEs in all member languages, including those occurring in text as discontiguous text segments, e.g. in 'He makes a meagre living planting sweet potatoes' (Figure 5.6). The SNODE intervals (in square brackets) are useful in indicating the position of the MWEs occurrences, i.e. using the SSTC+L annotation schema (see section 4.3).

For the evaluation, 80 input sentences containing LIs with translation ambiguities were randomly selected from the Internet (English, Malay and Chinese) and contributed by a native speaker (Iban). The test words are:

- English «plant» (vegetation or factory),
- English «bank» (financial institution or riverside land),

```
make a living [1_2, 2_3, 4_5]
zho:  谋生,
tha:  ทำมาหากิน, ประกอบอาชีพ,
msa:  cari makan, cari rezeki,


meager [3_4]
zho:  微薄,
fra:  maigre,
msa:  tidak mencukupi, kurang, sedikit,


plant [5_6]
zho:  栽培,
fra:  planter,
msa:  menanam,
iba:  tanam, ngujak, tambak,


sweet potato [6_7, 7_8]
zho:  山芋, 白薯, 甘薯,
tha:  มันเทศ,
msa:  ubi jalar, ubi keledek,
iba:  abuk,
```

Figure 5.6: Top translation sets selected by LEXICALSELECTOR for 'He makes a meagre living planting sweet potatoes.'

- Malay «kabinet» (governmental Cabinet or household furniture),
- Malay «mangga» (mango or padlock),
- Chinese «谷» (*gù*, valley or grain) and
- Iban «emperaja» (rainbow or lover).

Each test sentence was first POS-tagged. In addition, the English test sentences were lemmatised, and the Chinese input sentences segmented,[9] so that LIs and their associated vectors can be retrieved from Lexicon+TX.

The ranking strategy, `wiki-lsi`, first computes a 'query vector' by taking the vectorial sum of all LIs in each test sentence. The list of translation sets containing

[9]with the Stanford Chinese Word Segmenter tool (`http://nlp.stanford.edu/software/segmenter.shtml`)

the ambiguous LI is then sorted in descending order of the cosine similarity between the query vector and vectors of the translation sets (see section 4.2.2). The baseline strategy, base-freq), is to always select the translation set whose members occur most frequently in the bilingual Wikipedia corpus.

As a comparison, the English, Chinese and Malay test sentences were fed to Google Translate,[10] which is trained on 'parallel texts such as Arabic and English into the computer, using United Nations and European Union documents as key sources' (Tanner, 2007). (Google Translate does not support Iban currently.) The highest rank of the correct translation for the test words in English/Chinese/Malay are used to evaluate goog-tr. Ranks of the correct translation set output by wiki-lsi, goog-tr and base-freq strategies are attached in Appendix H and summarised in Table 5.11.

Table 5.11: Precision and MRR scores of context-dependent lexical lookup

Strategy	Incl. English & Iban		W/o English & Iban	
	Precision	MRR	Precision	MRR
wiki-lsi	0.650	0.810	0.690	0.845
base-freq	0.550	0.771	0.524	0.762
goog-tr	0.797	0.812	0.690	0.708

The first evaluation metric is by taking the precision of the first translation set returned by our lookup module, i.e. whether the top ranked translation set contains the correct translation of the ambiguous item. The precision metric is important for applications like MT and WSD, where only the top-ranked meaning or translation is considered. For this metric, wiki-lsi scored 0.650 when all 80 input sentences are tested, while the base-freq baseline scored 0.550. goog-tr has the highest precision at 0.797. Since English is an official language for both United Nations and European Union documents (United Nations, n.d.), goog-tr has a huge amount of training data for English and therefore performs very well on the English inputs (see Appendix H). However, if only the Chinese and Malay inputs — which has less presence on the

[10] http://translate.google.com on 3 October 2012

Internet and 'less rich' than English — were tested (since goog-tr cannot accept Iban inputs), wiki-lsi and goog-tr actually performs equally well at 0.690 precision.

The results may also be evaluated similar to a document retrieval task, i.e. as a ranked lexical lookup list for human consumption. This can then be measured by the mean reciprocal rank (MRR), i.e. the average of the reciprocal ranks of the correct translation set for each input in the set of test sentences, T:

$$\text{MRR} = \frac{1}{|T|} \sum_{i=1}^{|T|} \frac{1}{\text{rank}_i} \tag{5.4}$$

The MRR is a better metric for describing the lexical lookup process by a human reader while browsing a text. In our evaluation, the MRR score of wiki-lsi is 0.810, while base-freq scored 0.771. wiki-lsi even outperforms goog-tr when only the Chinese and Malay test sentences are considered for the MRR metric, as goog-tr did not present the correct translation in its list of alternative translation candidates for some test sentences. This suggests that the LSI-backed translation context knowledge vectors would be helpful in building an intelligent reading aid.

5.2.4 Discussion

In the context-dependent lexical lookup experiment, wiki-lsi performed better than base-freq for both the precision and the MRR metrics. wiki-lsi even outperforms goog-tr for less rich language test sentences (Malay, Chinese). While wiki-lsi is not yet sufficiently accurate to be used directly in a MT system, it is helpful in producing a list of ranked multilingual translation sets depending on the input context, as part of an intelligent reading aid. Specifically, the lookup module would have benefited if syntactic information (e.g. syntactic relations and parse trees) was incorporated during the training and testing phase. This would require more time in parsing the training corpus, as well as assuming that syntactic analysis tools are available to process test sentences of all languages, including the under-resourced ones.

Notice that even though the translation context knowledge vectors were extracted

from an English–Malay corpus, the same vectors can be applied on Chinese and Iban input sentences as well. This is especially significant for Iban, which otherwise lacks resources from which a lookup or disambiguation tool can be trained. Translation context knowledge vectors mined via LSI from a bilingual comparable corpus, therefore offers a fast, low cost and efficient fallback strategy for acquiring multilingual translation equivalence context information.

One drawback of the proposed approach is that the translation context knowledge may fail if lexical ambiguity is shared between the languages of the bilingual corpora. For example, if «cabinet»$_{\texttt{eng}}$ and «kabinet»$_{\texttt{msa}}$ were the only English and Malay members of the translation sets meaning 'government' and 'furniture' to appear in the bilingual corpus, the extracted vectors would not be able to differentiate between these two meanings.

In the meantime, translation knowledge vectors, mined via LSI from a bilingual comparable corpus, may offer a fast, low cost and efficient fallback strategy for acquiring multilingual translation equivalence context information.

5.3 Summary and Conclusion

A prototype Lexicon+TX comprising six member languages has been automatically constructed from seven bilingual dictionaries via the modified OTIC procedure. The OTIC filtering achieved F_1 scores in the range of 0.65–0.72 for Iban–Malay and Malay–Chinese. 91.2 % of 500 randomly chosen entries require minimal or no human correction. Due to OTIC's low requirement on input data, as well as the low expectation of linguistic expertise on human contributors due to Lexicon+TX's 'shallow' model, the proposed design and work flow for creating a first draft of a multilingual lexicon is especially suitable for under-resourced languages.

Lexicon+TX was also enriched with translation context knowledge, in the form of vectors resulting from running LSI on a bilingual comparable corpus constructed from Wikipedia articles in English and Malay. A context-dependent multilingual lexical

lookup module was implemented, using the cosine similarity score between the vector of the input sentence and those of candidate translation sets to rank the latter in order of relevance. This had a precision score of 0.650 (compared to baseline 0.550) and MRR score of 0.810 (compared to baseline 0.771). The precision and MRR scores rise to 0.690 and 0.845 for medium- and under-resourced language test inputs, outperforming Google Translate's lexical selection. The LSI-backed translation context knowledge vectors, mined from bilingual comparable corpora, thus provide a fast and affordable data source for building intelligent reading aids. An additional advantage of adopting a bag-of-words model to translation context modelling is that it can be reused for languages not in the bilingual corpora: this means under-resourced languages would benefit as well.

While this research shows how allowing reuse of lexical resources can make them more useful, the issue of copyright and legality will also need to be discussed if the results are to be made available to the public (Bond & Paik, 2012). All data from external parties used in this research are available under open source licences, or used with permission for research purposes. However, only data issued under open source licenses were made available at the project website accompanying this research (`https://bitbucket.org/liantze/lexicontx`).

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

Multilingual lexicons are lexical databases that list LIs from different languages that convey a common meaning. As repositories of multilingual translational equivalents, multilingual lexicons are important resources for NLP applications and human users alike. However, since compiling multilingual lexicons manually from scratch is a time-consuming and labour-intensive undertaking, it would be much more feasible to devise methodologies for creating them automatically from existing resources. Most of such existing attempts require input lexical resources with rich content fields or large corpora. Unfortunately, such resources are often unavailable for under-resourced languages, and would take a long time to be developed.

The objective of this research is therefore to propose a framework for constructing multilingual lexicons using low cost means and resources, such that under-resourced languages can be rapidly connected to richer, more dominant languages. The framework should be flexible enough to allow initial construction with shallow data, with semantics and other deeper levels of information to be added in later stages. The constructed multilingual lexicon may then be used to extract new bilingual dictionaries, or used in a context-dependent lexical lookup module.

The main contribution of this research is to demonstrate that despite limitations facing under-resourced languages, it is possible to rapidly generate a multilingual lexicon of 'draft quality' from simple dictionaries in the form of bilingual translation lists. It is also possible to produce a context-dependent lexical lookup module for any member language of the lexicon, using a comparable corpus of a specific language pair as training data.

The following sections will recap on the work carried out, in relation to the

research objectives (ROn) and research contributions (RCn) set out in Chapter 1. The final section will outline some ideas of possible further improvements and extensions.

6.1 Study of Multilingual Lexicon Projects

A study of current multilingual lexicon projects in Chapter 2 categorised multilingual lexicon architectures into either 'deep' or 'shallow' approaches, depending on whether a holistic interlingual framework was used to model lexical meaning (RO1). In addition, to design and construct a multilingual lexicon that would include under-resourced languages, the following requirements and constraints were identified:

- mechanisms for handling linguistic phenomena like lexical ambiguity, diversification, lexical gaps and MWEs;
- low linguistics expertise requirement on volunteers to optimise the pool of available speakers of under-resourced languages;
- low requirement on input bilingual data resources to avoid data acquisition bottleneck for under-resourced languages.

6.2 Design and Rapid Construction of a Multilingual Lexicon

Following from these requirements, Lexicon+TX, a 'shallow' multilingual lexicon has been designed in Chapter 3 (RC1, RO1). In Lexicon+TX, translation equivalents from different languages are grouped into translation sets, so that a translation set corresponds to a (coarse-grained) meaning or concept. This scheme does not presume linguistic expertise on human contributors, so that any polyglot can contribute content without having to learn about underlying semantic or linguistic frameworks. Tree structures (generated with parser tools) are used to model MWEs, while gloss-like phrases are allowed in cases of lexical gaps. A new procedure, the modified one-time inverse consultation (OTIC), has been proposed for automatically generating 'draft' multilingual translation sets using only simple bilingual dictionaries (RC2, RO2). The input dictionary only needs to contain a simple list of SL–TL translation mappings and the POS.

By enforcing the principle of 'minimum requirements' on linguistic expertise and input data richness, the proposed design and construction procedure allowed a Lexicon+TX prototype containing six member languages (English, Chinese, Malay, French, Thai and Iban) to be created quickly and with minimum cost (RO2, Chapter 5). In particular, Iban is an under-resourced language, for which a wordnet system, thesaurus and large corpus are non-existent or still in development. As far as the author is aware, this is the first time Iban is connected to French, Chinese and Thai in a lexicon or dictionary.

From the implementation results, the modified OTIC filtering mechanism achieved best F_1 scores of 0.725 and 0.660 for Malay–Chinese and Iban–Malay respectively. 91.2 % of 500 random multilingual entries from Lexicon+TX require minimal or no human correction. Human volunteers who evaluated translation pairings (against which results of the modified OTIC procedure were later checked) were able to work through the data quickly, with many of them finishing 500 pairs within 2–4 hours.

Overall, the results are highly satisfactory, considering the simplicity of the input data required in comparison to related work. Specifically, the proposed modified OTIC procedure provides a fast, cheap and effective way for generating a first draft of a multilingual dictionary, which will then be improved by human evaluators. However, the effectiveness of the modified OTIC procedures relies a great deal on the extent of overlap of the input dictionaries and in how the entries were worded. Filtering threshold parameters must also be adjusted for each run of the procedure to ensure the best trade-off between precision and recall of the final results.

6.3 Context-Dependent Lexical Lookup using Translation Context Knowledge

Comparable bilingual corpora, which are more readily available than aligned parallel corpora, were processed with LSI to extract distributional information about the context of translation equivalents in Chapter 4. This information, in the form of numerical vectors, acts as a form of translation context knowledge for the multilingual translation sets in Lexicon+TX. Under-resourced language LIs members of the

translation sets therefore benefit from the richer-resourced languages (i.e. those of the comparable corpus), which would otherwise lack any usable data to support translation selection. This translation context knowledge is then used to perform context-dependent lexical lookup on new input texts (RC3, RO3).

A new annotation schema, the SSTC+L, was also proposed in Chapter 4 for marking up LI occurrences in natural language text, together with the links to their canonical lemma entries in a given lexicon (RC4, RO3). Examples have been given to demonstrate how the SSTC+L is capable of handling syntactically flexible MWEs, as well as annotating translational lexical gaps effectively when used in tandem with the S-SSTC.

As an implementation result, Lexicon+TX was enriched with translation context knowledge, in the form of vectors resulting from running LSI on bilingual comparable corpus constructed from English and Malay Wikipedia articles. A context-dependent multilingual lexical lookup module was implemented, using the cosine similarity score between the vector of the input sentence and those of candidate translation sets to rank the latter in order of relevance (RC3, RO3, Chapter 5). This had a precision score of 0.650 (compared to baseline 0.550) and MRR score of 0.810 (compared to baseline 0.771). The precision and MRR scores rise to 0.690 and 0.845 for medium- and under-resourced language test inputs, outperforming Google Translate's lexical selection. While this lookup module is not yet sufficiently accurate to be used directly in a MT system, it is helpful in producing a list of ranked multilingual translation sets depending on the input context, as part of an intelligent reading aid. In addition, it accepts input text in languages other than English and Malay (languages of the training corpus), as has been shown in the test results (Chapter 5, Appendix H).

6.4 Future Work

There are many areas for further work and extensions. These may broadly to grouped into future work on Lexicon+TX itself, as well as on its applications.

6.4.1 Future Work on Lexicon+TX

Apart from adding more languages, there are certain aspects on Lexicon+TX itself that would be interesting to work on. This section briefly describes some of the possible areas for further investigation.

6.4.1 (a) Linking to Other Multilingual Lexicons

Translation sets from Lexicon+TX can be linked or aligned to entries in Princeton's WordNet (Miller et al., 1990), Papillon (Boitet et al., 2002) or the UNL (Uchida, Zhu, & Senta, 2005) UW dictionary. This has mutual benefits for both Lexicon+TX and the external multilingual lexicons:

- All lexicons would acquire connections to new languages after the alignment. In particular, the external multilingual lexicons would now have access to less frequent language-pairs or under-resourced languages.
- Some of these lexicons, such as WordNet (Miller et al., 1990), have many other language resources developed around them. Examples include syntactic–semantic relations for verbs in VerbNet (Shi & Mihalcea, 2005; Kipper et al., 2008); case semantics and semantic roles in FrameNet (Fontenelle, 2003; Shi & Mihalcea, 2005); subject field labels (Magnini & Cavaglià, 2000); and ontology class labels from the SUMO (Niles & Pease, 2001, 2003). By linking Lexicon+TX to WordNet (for example), Lexicon+TX would gain access to these rich resources too.

6.4.1 (b) Automatic Acquisition of Diversification and Other Semantic Relations

The architecture of Lexicon+TX already allows diversification (section 2.2.1 to be modelled (section 3.1.1 (b)) as relations among the axies. This may also be used for modelling other types of semantic relations (such as hypernymy, causation and entailment) in future. Automatic means for extracting such relations from various

sources may be investigated further: a rough idea for detecting possible diversification from a bilingual corpus has been sketched in Lim, Ranaivo-Malançon, and Tang (2011a).

6.4.1 (c) Introducing Deeper Semantics

Lexicon+TX, in its current state, adopts a 'shallow' approach to modelling translational equivalence across languages. As discussed in section 2.3, 'shallow' multilingual lexicons are faster and easier to implement, and may be used in certain NLP applications that require only shallow semantic processing. However, if Lexicon+TX was to be used in systems that require deeper language understanding, richer ('deeper') semantic constructs must be introduced into its framework.

How this may be done is open to investigation. One possible approach is to link Lexicon+TX translation sets to external knowledge resources (section 6.4.1 (a) above), as has been done for Princeton WordNet and provisioned in the LMF (ISO24613, 2008). Alternatively, a holistic semantic framework, e.g. in the form of a wordnet or a more formal usage label ontology, may be devised to underlie the current Lexicon+TX architecture (if feasible).

6.4.2 Future Work on Applications

Another general direction for further work on this research is the applications of Lexicon+TX.

6.4.2 (a) Improving Translation Selection and Context-Dependent Lexical Lookup Accuracy

As the results in section 5.2 show, while the current strategy for content-dependent lexical lookup performs better than the baseline strategy, there is still room for improvement for the accuracy. Specifically, more information, such as syntactic relations and other cues, should be incorporated into the selection and ranking procedure, while ensuring that the algorithm is applicable to as many languages as possible.

6.4.2 (b) Advanced Recognition of MWEs

Currently, MWEs in Lexicon+TX are modelled as a string-tree correspondence structure (the SSTC, section 3.1.2 (b) and Appendix C), which records the canonical form and tree representation of a MWE. In view of the wide range of flexibility of MWEs (see section 2.2.3 and Sag et al., 2002), it would be interesting to study if and how generation templates of valid MWEs can be learned, as well as automatically recognising passivisation and topicalisation of MWEs in a text.

6.4.2 (c) Integration with an MT System

Section 4.3 presented the SSTC+L annotation schema for 'packaging' context-dependent lexical lookup results for machine-tractable consumption by other NLP systems, which is especially effective for non-contiguous MWEs. This paves the way for integrating Lexicon+TX into an MT system (namely SiSTeC-EBMT, Boitet et al., 2011), such that the content-dependent lexical lookup results can be used to construct the final translation output.

6.4.2 (d) Interactive User Interface for Symbiotic Updates

An interactive user interface can be provided to capture user actions in correcting or modifying the content-dependent lexical lookup results while working with the MT system that the lookup module is embedded in. The user interactions may provide further training data for the lookup module, as well as correcting lexicographic data errors in Lexicon+TX as described briefly in (Lim, Ranaivo-Malançon, & Tang, 2011b).

6.5 Conclusion

The research presented in this thesis is concerned with the rapid design (RO1), construction (RO2) and application (RO3) of a multilingual lexicon that takes into consideration the constraints faced by under-resourced languages, using low cost methodologies and human expertise.

The outcome of a study into existing multilingual lexicon projects yielded a lex-

icon design with a 'shallow' translation equivalence model, and is capable of modelling several linguistic phenomena, including diversification, lexical gaps and syntactically flexible MWEs (RC1). Using the methodology (modified OTIC) proposed in this thesis (RC2), Lexicon+TX, a prototype multilingual lexicon containing six languages (English, Malay, Chinese, French, Thai and Iban) was successfully constructed using simple input bilingual dictionaries. As far as the author is aware, this is the first time that Iban, a under-resourced languages, is connected to more widely spoken languages like Chinese, French and Thai.

A context-dependent lexical lookup module based on Lexicon+TX has also been developed, using translation context knowledge extracted via LSI from a bilingual comparable corpus. Using the methodology proposed in this thesis, the lookup module is able to process input text in any member language of Lexicon+TX, including Iban, which otherwise lack NLP resources for building a similar tool (RC3). A new annotation schema, the SSTC+L, has also been proposed for annotating the lexical lookup results (which may contain lexical gaps and flexible MWEs), so that they may be used by other NLP systems (RC4).

The results have shown that by using simple, easy-to-acquire input data and minimum linguistics expertise on human volunteers, it is possible to connect under-resourced languages to more dominant, richer-resourced languages via a multilingual lexicon with highly satisfactory results in a relatively short time. This paves the important first step for developing more NLP resources and processing tools for these under-resourced languages, thus helping more communities gain access to information that may be previously unintelligible.

APPENDIX A

ISO 639-1 AND ISO 639-3 LANGUAGE CODES

ISO 639-3 (Codes for the representation of names of languages – Part 3: Alpha-3 code) are 3-letter codes that attempt to 'provide as complete an enumeration of languages as possible, including living, extinct, ancient, and constructed languages, whether major or minor, written or unwritten'. (See `http://www.sil.org/iso639-3/` for the full list).

ISO 639-1, a list of 2-letter codes, is a subset of ISO 639-3. It was devised primarily for use in terminology, lexicography and linguistics.

ISO 639-1	ISO 639-3	Language name
de	deu	German
en	eng	English
fr	fra	French
–	iba	Iban
ko	kor	Korean
it	ita	Italian
ja	jpn	Japanese
ms	msa	Malay
ru	rus	Russian
th	tha	Thai
zh	zho	Chinese

APPENDIX B

LIST OF PART-OF-SPEECH CODES

Code	POS
N	noun
V	verb
A	adjective
ADV	adverb
DET	determiner
PREP	preposition
PRON	pronoun
PRON_POSS	possessive pronoun

APPENDIX C

STRUCTURED STRING-TREE CORRESPONDENCE ANNOTATION FRAMEWORKS: FORMAL DEFINITIONS

C.1 Structured String-Tree Correspondence

The Structured String-Tree Correspondence (SSTC) (Boitet & Zaharin, 1988) is an annotation schema for declaratively specifying multi-level correspondences between a string and its tree representation structure of arbitrary choice.

An SSTC comprises a string *st*, its tree representation structure *tr*, and the correspondences between them, *co*. Substrings of S are identified by *intervals*, which serve as mechanisms for specifying the correspondences between *st* and *tr* on two levels:

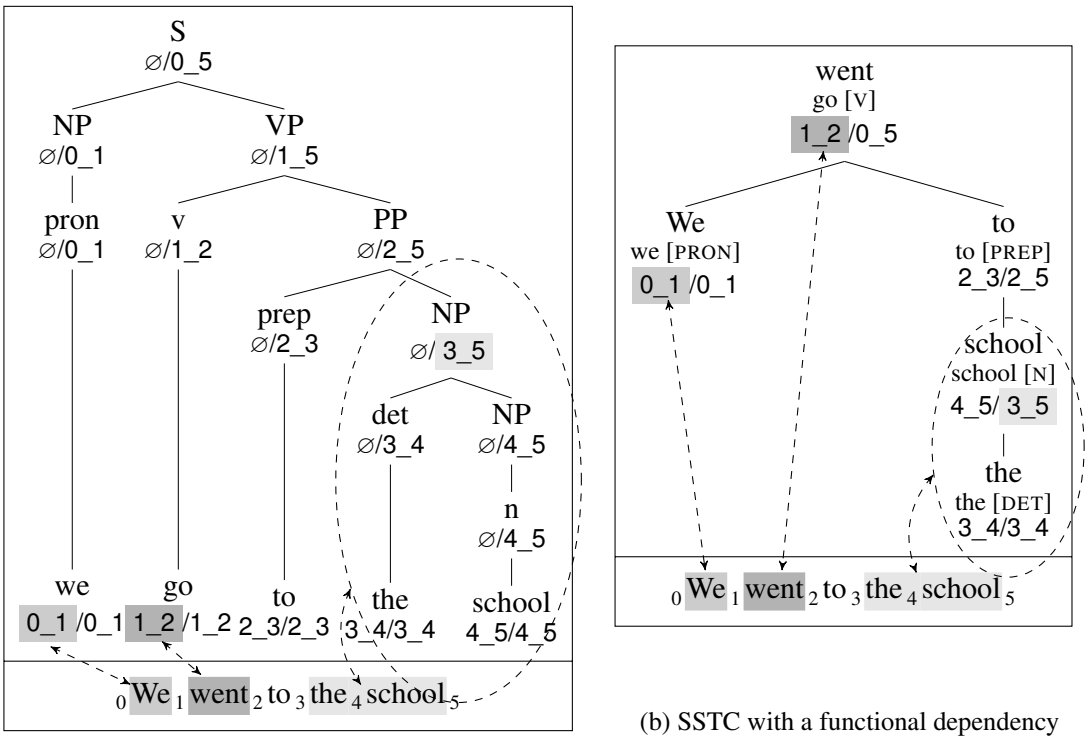
- lexical level, i.e. between (possibly discontiguous) substrings of *st* and tree nodes of *tr*, using SNODE intervals; and
- phrase level, i.e. between (possibly discontiguous) substrings of *st* and (possibly incomplete) subtrees of *tr*, using STREE intervals.

Formally, an SSTC is a triple (st, tr, co) where

- *st* is a string in one language,
- *tr* is its associated tree structure,
- *co* is the correspondence between *st* and *tr*.
- *co* can be encoded on the tree by attaching to each node N in *tr* two sequences of intervals:
  - SNODE(N): an interval of the substring in *st* that corresponds to the node N in *tr*.
  - STREE(N): an interval of the substring in *st* that corresponds to the subtree in *tr* having the node N as root.

Intervals are written as a minimal list, from left to right. That means that any occurrence of $n_1\_n_2 + n_2\_n_3$ is replaced by $n_1\_n_3$, n_i being a position between two typographical words (word-based),

or more generally (to handle writing systems without word delimiters such as Chinese, Japanese, Korean, Vietnamese, Thai, Lao, or Khmer), between two characters (character-based).



(a) SSTC with a phrase structure tree

(b) SSTC with a functional dependency tree

Figure C.1: SSTCs with different tree representation structures

The SSTC schema allows the annotator to choose any arbitrary tree representation model to be associated with a string, e.g. phrase structure trees or dependency trees, or other syntagmatic, functional and logical structures, to suit the needs of the task at hand. Figure C.1(a) shows an SSTC adopting a phrase structure tree representation, while Figure C.1(b) shows another SSTC using a functional dependency tree.

C.2 Synchronous Structured String-Tree Correspondence

The SSTC is a highly flexible structure, such that non-standard language phenomena, such as non-projectivity and ellipsis, can be captured declaratively. Its extension, the Synchronous SSTC (S-SSTC) schema (Al-Adhaileh et al., 2002), consists of a pair of SSTCs. Its formal definition is given below.

Let S and T be two SSTCs. An S-SSTC is a triple $(S, T, \varphi_{S,T})$ where $\varphi_{S,T}$ is a set of links

defining the synchronous correspondences between S and T at different internal levels of the two SSTC structures.

A synchronous correspondence link $\ell \in \varphi_{S,T}$ can be of type $\underset{sn}{\ell}$ or $\underset{st}{\ell}$.

- $\underset{sn}{\ell}$ records the synchronous correspondences at level of nodes in S and T (i.e. lexical correspondences between specified nodes) and normally $\underset{sn}{\ell} = (X_1, X_2)$ where X_1 and X_2 are sequences of SNODE correspondences in co, which may be empty.
- More specifically, $\underset{sn}{\ell}$ is a pair ($\underset{sn_S}{\ell}$, $\underset{sn_T}{\ell}$) where $\underset{sn_S}{\ell}$ is from the first SSTC (S) and $\underset{sn_T}{\ell}$ is from the second SSTC (T).
- $\underset{sn}{\ell}$ is represented by sets of intervals such that:
  - $\underset{sn_S}{\ell} = \{i_1\_j_1 + \cdots + i_k\_j_k + \cdots + i_p\_j_p\}$
    where $i_k\_j_k \in X$: SNODE correspondence in co of S
  - $\underset{sn_T}{\ell} = \{i_1\_j_1 + \cdots + i_k\_j_k + \cdots + i_p\_j_p\}$
    where $i_k\_j_k \in X$: SNODE correspondence in co of T
- $\underset{st}{\ell}$ records the synchronous correspondences at level of nodes in S and T (i.e. structural correspondences between specified nodes) and normally $\underset{st}{\ell} = (Y_1, Y_2)$ where Y_1 and Y_2 are sequences of STREE correspondences in co, which may be empty.
- More specifically, $\underset{st}{\ell}$ is a pair ($\underset{st_S}{\ell}$, $\underset{st_T}{\ell}$) where $\underset{st_S}{\ell}$ is from the first SSTC (S) and $\underset{st_T}{\ell}$ is from the second SSTC (T) as defined below:
  - $\underset{st_S}{\ell} = \{i_1\_j_1 + \cdots + i_k\_j_k + \cdots + i_p\_j_p\}$
    where $i_k\_j_k \in Y$: STREE correspondence in co of S, or
    $(i_k\_j_k) = (i_k\_j_k) - (i_u\_j_v) \quad | \quad i_u \geq i_k \wedge j_v \geq j_k;$
    i.e. $(i_u\_j_v) \subseteq (i_k\_j_k)$ which corresponds to an incomplete subtree.
  - $\underset{st_T}{\ell} = \{i_1\_j_1 + \cdots + i_k\_j_k + \cdots + i_p\_j_p\}$
    where $i_k\_j_k \in Y$: STREE correspondence in co of T, or
    $(i_k\_j_k) = (i_k\_j_k) - (i_u\_j_v) \quad | \quad i_u \geq i_k \wedge j_v \geq j_k;$
    i.e. $(i_u\_j_v) \subseteq (i_k\_j_k)$ which corresponds to an incomplete subtree.
- The synchronous correspondence between terminal nodes with X : SNODE $= Y$: STREE will be of both $\underset{sn}{\ell}$ and $\underset{st}{\ell}$ such that $\underset{sn}{\ell} = \underset{st}{\ell}$.

APPENDIX D

A MANUAL FOR LEXICON+TX CONSTRUCTION AND EXPANSION

Multilingual Dictionary Representation and Generation

and Extracting Bilingual Dictionaries from Them

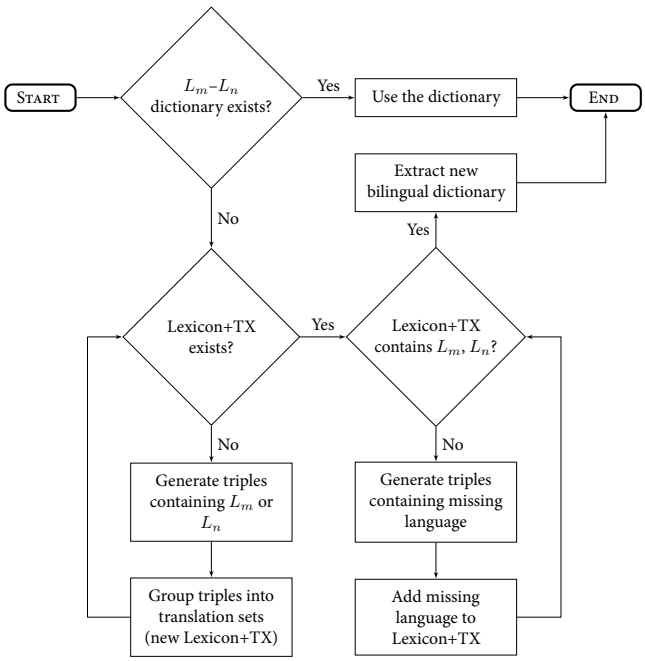
Lim Lian Tze (liantze@gmail.com)

SiSTeC Training Workshop

Abstract

SiSTeC-ebmt requires a bilingual dictionary to look up translation equivalents. However, a dictionary may not be available for a particular language pair, especially those involving under-resource languages. In this training, you will learn how to build or enrich a prototype multilingual dictionary using simple bilingual translation lists that *do* exist, using a modified one-time inverse consultation (OTIC) procedure. A bilingual dictionary of the required language pair can then be extracted from this multilingual dictionary.

Use the following flowchart to decide what needs to be done for producing a (draft quality) bilingual dictionary for the language pair L_m–L_n.



1

1. SiSTeC-ebmt Bilingual Dictionary Format

SiSTeC-ebmt uses a bilingual dictionary for looking up translation equivalents of words or lexical items (LIs) that it cannot find in the `tree_corr` and `node_corr` tables.

If you have some existing bilingual dictionary data for your required language pair, you can import the data into a MySQL table for use with SiSTeC-ebmt, similar to `kimd_entries` in SiSTeC-ebmt English–Malay system. The table definition of `kimd_entries` is:

```
mysql> describe kimd_entries;
+----------+-------------+------+-----+---------+-------+
| Field    | Type        | Null | Key | Default | Extra |
+----------+-------------+------+-----+---------+-------+
| KE_ID    | int(11)     | NO   | PRI | 0       |       |
| KE_ENTRY | varchar(50) | NO   | MUL |         |       |
| KE_POS   | varchar(15) | NO   |     |         |       |
| KE_EQUIV | text        | YES  |     | NULL    |       |
+----------+-------------+------+-----+---------+-------+
```

There must be fields for source language (SL) LIs, part of speech (POS), and target language (TL) equivalent(s). A few sample rows from the table:

```
+-------+-----------+--------+---------------------------+
| KE_ID | KE_ENTRY  | KE_POS | KE_EQUIV                  |
+-------+-----------+--------+---------------------------+
|     4 | abacus    | n      | sepua, sempoa, dekak-dekak |
|     6 | abaft     | prep   | di sebelah buritan dr      |
|     7 | abalone   | n      | abalone                    |
|    14 | abashed   | adj    | silu, malu                 |
|    15 | abashment | n      | rasa malu, perasaan malu   |
+-------+-----------+--------+---------------------------+
```

Note that if `KE_EQUIV` is a comma-delimited list, or if multiple rows exist for a `KE_ENTRY` value, SiSTeC-ebmt will only select the first given translation equivalent at present.

If you use different names for your bilingual dictionary table and fields, you may have to modify the code in `ebmt.BilingualLexicon.java` accordingly.

2. Wait, But I Don't Have a Dictionary for My Language Pair

Suppose we need a bilingual dictionary for the language pair L_m–L_n, but have no such data available. Suppose now we had a multilingual dictionary with languages $\{L_1, L_2, \ldots, L_k\}$:

- If the multilingual dictionary already contains L_m and L_n, we can extract the bilingual dictionary. (See section 8.)
- If the multilingual dictionary contains only one (or neither) of L_m and L_n, we add L_m and/or L_n to it.

  To add a new language L_m:

  1. Choose any L_x, L_y already present in the multilingual dictionary, such that you *do* have bilingual dictionaries for L_m–L_x, L_x–L_y and L_y–L_x.

These bilingual dictionaries are only required to take the form of simple bilingual translation lists. The POS field is not compulsory, but are useful as a filtering mechanism if present. Here's an example English–Chinese bilingual translation list, suitable for our purposes:

LI	pos	gloss
coast	n	沿海地区
coast	v	滑行
coast	v	滑翔
coastal	a	海岸的
coastal	a	沿海的
coastal	a	沿岸的
coaster	n	沿岸贸易船
coaster	n	居民
coaster	n	撬

Note how each row contains only one translation pair. See section 4.1 on preparing the input bilingual dictionary tables.

2. Generate a list of translation triples $(w_{L_m}, w_{L_x}, w_{L_y})$ using the modified OTIC procedure. (See section 6.)

3. Add the filtered list of w_{L_m} to existing translation sets of the multilingual dictionary, by matching w_{L_x} and w_{L_y}. (See section 7.)

- If there was no multilingual dictionary to begin with, regroup the translation triples to produce a new multilingual dictionary with initial languages L_m, L_x and L_y. (See section 6.)

The full algorithms can be found in the appendix and (Lim, Ranaivo-Malançon, & Tang, 2011); it may be helpful to grasp the main ideas of the principle, in case you need to modify the sample Java code to suit your own needs.

> Important Disclaimer
>
> This method requires only very simple data, and is therefore suitable for under-resourced languages. However, due to the same reason, *you should not expect outstanding high accuracy or coverage of the output!* Our purpose here is to quickly produce a *'draft' copy* of a multilingual dictionary, which should be manually verified later. The emphasis is to produce *something* that can be used in larger NLP systems (especially in research projects short on funding), as opposed to having *nothing* to go on at all.

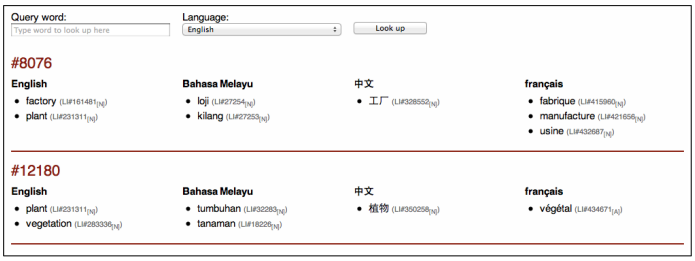
3. Multilingual Dictionary Representation

If there is no existing bilingual dictionary for a language pair L_m–L_n, it may be possible to extract one from a multilingual dictionary (e.g. Lexicon+TX) that comprises languages $L_1, L_2, \ldots, L_k$ if $1 \leq m, n \leq k$; i.e. languages L_m and L_n are already present in the multilingual dictionary. If L_m or L_n are missing, it is possible to add them to Lexicon+TX. The approach described here requires only simple bilingual translation mapping lists, and is therefore especially suitable for under-resourced languages.

3.1. Importing Lexicon+TX Database

`LexiconTX.sql` contains Lexicon+TX, a coarse-grained multilingual lexicon of 'draft' quality,[1] containing English, Chinese, Malay and French (Lim et al., 2011). **(Lexicon+TX is only available for academic research purposes at present.)**

In addition, if you have Apache and PHP installed, a very basic online interface (`search.php` is available for quick look-ups. `randomsets.php` randomly outputs some translation sets for sampling.

> Have a Go
>
> Create a new user `sistec` (password `sistecdict`) and load the `LexiconTX` data:
>
> **Windows:**
>
> ```
> prompt>  mysql -u root -p < sql\CreateLexiconTX.sql
> ```
> (Key in your MySQL `root` password)
>
> ```
> prompt>  mysql -u sistec -p LexiconTX < sql\biling-dicts.sql
> prompt>  mysql -u sistec -p LexiconTX < sql\LexiconTX.sql
> ```
> (Key in `sistecdict` as password)
>
> ***nix and Mac:**
>
> ```
> prompt$  mysql -u root -p < sql/CreateLexiconTX.sql
> prompt$  mysql -u sistec -p LexiconTX < sql/biling-dicts.sql
> prompt$  mysql -u sistec -p LexiconTX < sql/LexiconTX.sql
> ```
>
> You may substitute any username and password suitable for your setup. Remember to change the parameters in the properties file later if you use a different username and password.
>
> If Apache and PHP (with `mysqli` extension enabled) are installed on your machine, copy the files in the `php` folder to the relevant Apache web documents folder (e.g. `<Apache>\htdocs\Lexicon+TX\`), then access them in a browser at e.g. `http://localhost/Lexicon+TX/search.php`.
>
> 
>

3.2. Lexicon+TX Architecture

The rest of this section briefly describes the architecture design of Lexicon+TX, a multilingual dictionary. Figure 1 shows the conceptual view of a translation set in Lexicon+TX. Each translation set corresponds to a

[1] i.e. it may contain erroneous entries and/or mappings, but usable as a prototype when there is absolutely no other available resources

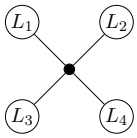


Figure 1: Conceptual structure of a multilingual translation set

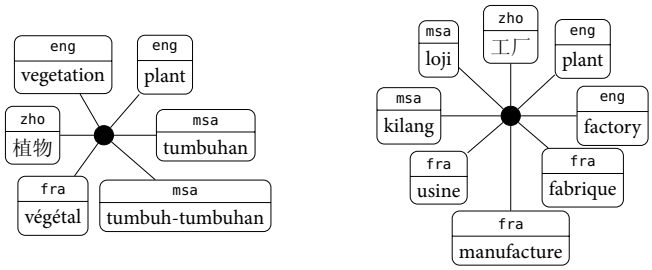


Figure 2: Example translation set for two different senses of English «plant» with lexical items from English, Chinese, Malay and French.

coarse-grained lexical sense or concept, and is accessed by a language-independent axis node. Translation equivalents expressing the same sense are connected to the axis. See (Lim et al., 2011) for further information about the dictionary architecture, which is heavily inspired by Papillong (Boitet, Mangeot, & Sérasset, 2002) and the Lexical Markup Framework (International Organization for Standardization [ISO], 2008; Francopoulo et al., 2009, multilingual extension).

As an illustration, Figure 2 shows two translation sets, each containing a different sense (meaning) of English «plant»: one for vegetation life, and the other for factories. Implementation wise, Figure 3 shows the ER diagram of the Lexicon+TX database tables (simplified view).

`Language` Languages are identified by their ISO 639-3 codes (http://sil.org/iso639-3/codes.asp), e.g. `eng` (English), `msa` (Malay), `fra` (French), `tha` (Thai).

`Axis` A language-independent mechanism for connecting translation equivalents from different languages together. Each `Axis` corresponds roughly to a lexicalised concept in some language, i.e. a coarse-grained lexical sense. If desired, each `Axis` can be further annotated with a `domain`, `concept label`, etc.

`LexicalItem` A single word, or chain of words, that make up a language's vocabulary and understood to convey a single meaning. Each `LexicalItem` has (at least) the fields for its language, lemma and (sometimes missing) POS. The internal tree structures of multi-word expressions (MWEs) may be recorded in a `tree` field if desired (c.f. SSTC).

`Gloss` An explanation of a `LexicalItem`, usually a phrasal construction (although valid `LexicalItem`s are often included redundantly). To illustrate: «lass» and 'young girl' are both translations of «少女», but «lass» would be recorded in `LexicalItem`, and 'young girl' in `Gloss`.

`TransEquiv` relates `LexicalItem`s conveying the same (coarse-grained) meaning to a common `Axis`.

`TransGlossEquiv` relates `Gloss`es conveying the same (coarse-grained) meaning to a common `Axis`.

As an example, the following shows the table rows corresponding to translations of «plant» as meaning

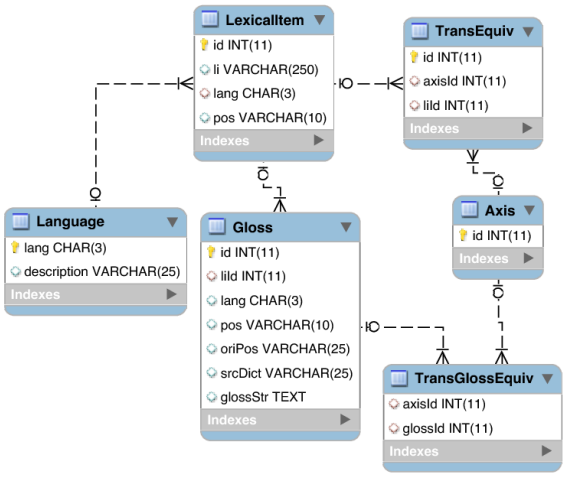


Figure 3: ER diagram of multilingual lexicon

'vegetation' (`Axis.id=11121`) and 'factory' (`Axis.id=7368`) respectively.

```
Axis          TransEquiv              LexicalItem
+-------+     +--------+--------+     +--------+------+-----------------+------+
| id    |     | Axisid | LIid   |     | id     | lang | li              | pos  |
+-------+     +--------+--------+     +--------+------+-----------------+------+
| 11121 |     |  11121 | 241793 |     | 241793 | eng  | plant           | N    |
+-------+     |  11121 | 293818 |     | 293818 | eng  | vegetation      | N    |
              |  11121 |  59675 |     |  59675 | zho  | 植物            | N    |
              |  11121 | 333582 |     | 333582 | msa  | tumbuhan        | N    |
              |  11121 | 428629 |     | 428629 | fra  | végétal         | N    |
              |  11121 | 319415 |     | 319415 | msa  | tanaman         | N    |
              |  11121 | 333347 |     | 333347 | msa  | tumbuh-tumbuhan | N    |
              +--------+--------+     +--------+------+-----------------+------+
```

```
Axis          TransEquiv              LexicalItem
+------+      +--------+--------+     +--------+------+-------------+------+
| id   |      | Axisid | LIid   |     | id     | lang | li          | pos  |
+------+      +--------+--------+     +--------+------+-------------+------+
| 7368 |      |   7368 | 171963 |     | 171963 | eng  | factory     | N    |
+------+      |   7368 | 241793 |     | 241793 | eng  | plant       | N    |
              |   7368 |  37969 |     |  37969 | zho  | 工厂        | N    |
              |   7368 | 328495 |     | 328495 | msa  | kilang      | N    |
              |   7368 | 328496 |     | 328496 | msa  | loji        | N    |
              |   7368 | 409918 |     | 409918 | fra  | fabrique    | N    |
              |   7368 | 415614 |     | 415614 | fra  | manufacture | N    |
              |   7368 | 426645 |     | 426645 | fra  | usine       | N    |
              +--------+--------+     +--------+------+-------------+------+
```

6

4. Preparing the Inputs and Configurations

4.1. Preparing the Input Bilingual Dictionary Tables

Each bilingual dictionary can be organised as one single flat table; or you could normalise them into multiple tables e.g. a table for the SL items and another table for the TL translations. The important thing is that you are able to retrieve

- a list of unique SL `LI`s,
- for each of these `LI`s, a list of TL `gloss`es and (optionally) the `pos`,
- for each unique `(LI, gloss, pos)` translation mapping row, there is a numerical identifier `id`.

For example, if these are your tables for an English–Chinese dictionary, with foreign keys defined and working correctly (`InnoDB table types`):

```
xdict-LI                xdict-gloss
+-------+---------+      +-------+------+--------------+-------+
| id    | LI      |      | id    | pos  | gloss        | LI_id |
+-------+---------+      +-------+------+--------------+-------+
| 32173 | coast   |      | 52198 | n    | 海岸         | 32173 |
| 32174 | coastal |      | 52199 | n    | 海滨         | 32173 |
| ...   | ...     |      | 52200 | n    | 沿海地区     | 32173 |
                           | 52201 | v    | 滑行         | 32173 |
                           | 52202 | v    | 滑翔         | 32173 |
                           | 52203 | a    | 海岸的       | 32174 |
                           | 52204 | a    | 沿海的       | 32174 |
                           | 52205 | a    | 沿岸的       | 32174 |
                           | ...   | ...  | ...          | ...   |
```

Then you may create a view as follows:

```
CREATE ALGORITHM=MERGE VIEW `xdict-en-zh` AS
SELECT L.id LI_id, L.LI LI, G.id gloss_id, G.pos, G.gloss, G.polished_gloss
FROM `xdict-LI` L INNER JOIN `xdict-gloss` G ON (L.id = G.LI_id);
```

```
xdict-en-zh
+-------+---------+----------+------+--------------+
| LI_id | LI      | gloss_id | pos  | gloss        |
+-------+---------+----------+------+--------------+
| 32173 | coast   |    52198 | n    | 海岸         |
| 32173 | coast   |    52199 | n    | 海滨         |
| 32173 | coast   |    52200 | n    | 沿海地区     |
| 32173 | coast   |    52201 | v    | 滑行         |
| 32173 | coast   |    52202 | v    | 滑翔         |
| 32174 | coastal |    52203 | a    | 海岸的       |
| 32174 | coastal |    52204 | a    | 沿海的       |
| 32174 | coastal |    52205 | a    | 沿岸的       |
| ...   | ...     |    ...   | ...  | ...          |
```

The fields of interest here are:

- `xdict-en-zh` is the English–Chinese (`eng-zho`) table/view.
- `gloss_id` is the column containing the unique identifier for each translation pair.

- `LI` is the column containing the LI.
- `pos` is the column containing the POS.
- `gloss` is the column containing the gloss string.

The Lexicon+TX database contains several input bilingual dictionaries. Some have a public license; others are available for research purposes. See Appendix A for the list.

> Have a Go
>
> We will add Thai to Lexicon+TX, using Yaitron, an open source Thai–English dictionary. Import the Yaitron bilingual dictionary into MySQL:
>
> **Windows:**
>
> ```
> prompt>  mysql -u sistec -p LexiconTX < sql\yaitron.sql
> ```
>
> ***nix and Mac:**
>
> ```
> prompt$  mysql -u sistec -p LexiconTX < sql/yaitron.sql
> ```

4.2. Mappings and Normalisations

- Add information about your input dictionaries in `liantze.struct.DictSource` if applicable. All we really need is a Java `Enum` representing your input dictionaries, e.g. `DictSource.XDict`.

- Map the `pos` codes in your dictionaries to Java `PartOfSpeech` Enums in `liantze.struct.PartOfSpeech.mapPos()`. At the very least, make sure that nouns, verbs, adjectives and adverbs are identified.

- Map the `pos` codes to SiSTeC-ebmt POS codes in a MySQL table `xxxPOS`, where `xxx` is a prefix code you assign to your bilingual dictionary. For example:

  ```
  sistecemPOS
  +--------------+--------+
  | pos          | stdpos |
  +--------------+--------+
  | kn.          | N      |
  | kn.&kk.i.    | N      |
  | kn.&kk.i.    | V      |
  | kk.t./i.     | V      |
  | ks.          | A      |
  | kn.&ks.      | A      |
  | kn.&ks.      | N      |
  | ...          | ...    |
  ```

- Implement string normalisation routines for the gloss strings in `liantze.utils.DictUtils.normalise(lang, gloss)`, if necessary.

  For example, you may want to normalise the English verb 'to cultivate' to simply 'cultivate'; or remove the adjectival article '的' from Chinese '茂盛的', so that it's easier to look up these words in the next dictionary in the chain.

4.3. Preparing the Configuration Properties File

Create a `.properties` file for your L_1–L_2–L_3 processing task. The fields concern the following:

- Database connection details (UTF-8 assumed)
  - `dbHost` e.g. `localhost`
  - `dhPort` e.g. `3306`
  - `dbName` e.g. `LexiconTX`
  - `dbUsername` e.g. `sistec`
  - `dbPassword` e.g. `sistectdict`
- Languages and input dictionaries (must match `DictSource Enums`)
  - `lang1` e.g. `tha`
  - `lang2` e.g. `eng`
  - `lang3` e.g. `zho`
  - `new_lang` e.g. `tha` (Thai is the new language to be added)
  - `langX_langY_Dict` e.g. `XDICT`
  - `langX_langY_DictPrefix` e.g. `xdict` (prefix of table `xxxPOS`
- Column names of interest in bilingual dictionary tables
  - `langX_langY_dict_table` e.g. `xdict-en-zh`
  - `langX_langY_LI_col` e.g. `LI`
  - `langX_langY_gloss_id_col` e.g. `gloss_id`
  - `langX_langY_gloss_col` e.g. `gloss`
  - `langX_langY_pos_col` e.g. `pos`
  - `langX_langY_qsuffix` any additional conditions for restricting the set of LIs to be considered in the OTIC process
- Prefix for temporary processing table and log file
  - `triplesPrefix` e.g. `tez`
  - `logPrefix` e.g. `tez`
- Filter threshold values
  - `threshold_ALPHA` e.g. `0.6`, filters triples with score $\geq \alpha \times \max(\text{score}_{L_1-L_3})$
  - `threshold_BETA` e.g. `0.2`, filters triples with score^2 $\geq \beta \times \max(\text{score}_{L_1-L_3})$

> **Have a Go**
>
> See `props\tha-eng-zho.properties` for a sample properties file of generating Thai–English–Chinese triples from Thai–English (Yaitron), English–Chinese (XDict) and Chinese–English (CC-CEDICT) input dictionaries, for the purpose of adding Thai to Lexicon+TX.

4.4. Importing Input Dictionary Entries

LIs and glosses from the input dictionaries should be copied into the `LexicalItem` and `Gloss` tables, with SiSTeC-ebmt standard POS codes. A `BilingDictImporter` tool is provided:

Windows:

```
prompt>  java -cp .;bin;lib\mysql-connector-java-5.1.12-bin.jar ^
           liantze.build.BilingDictImporter <task-props> <head|mid|tail|all>
```

***nix and Mac:**

```
prompt$  java -cp .:bin:lib/mysql-connector-java-5.1.12-bin.jar \
           liantze.build.BilingDictImporter <task-props> <head|mid|tail|all>
```

- `task-props`: `<task-props>.properties` is the file you prepared in the previous subsection.
- Specify which dictionary you want to import
  - `head`: the L_1–L_2 dictionary (`lang1_lang2_dict_table`)
  - `mid`: the L_2–L_3 dictionary (`lang2_lang3_dict_table`)
  - `tail`: the L_3–L_2 dictionary (`lang3_lang2_dict_table`)
  - `all`: all three dictionaries

> Have a Go
>
> Import Yaitron LIs and glosses into `LexiconTX`:
>
> **Windows:**
>
> ```
> prompt> java -cp .;bin;lib\mysql-connector-java-5.1.12-bin.jar ^
>  liantze.build.BilingDictImporter props\tha-eng-zho head
> ```
>
> ***nix and Mac:**
>
> ```
> prompt$ java -cp .:bin:lib/mysql-connector-java-5.1.12-bin.jar \
>  liantze.build.BilingDictImporter props/tha-eng-zho head
> ```

5. Generating Translation Triples from Bilingual Lists

Given L_1–L_2, L_2–L_3 and L_3–L_2 bilingual dictionaries (translation lists), we generate a list of translation triples $(w_{L_1}, w_{L_2}, w_{L_3})$ using the modified OTIC procedure. The algorithm is in Appendix B.1; or read more from (Lim et al., 2011). After running the `TripleGenerator` tool, all triples with non-zero scores will be placed in a table `<triplesPrefix>_temp`.

Windows:

```
prompt>  java -cp .;bin;lib\mysql-connector-java-5.1.12-bin.jar ^
  liantze.build.TripleGenerator <task-props> [filterPos]
```

***nix and Mac:**

```
prompt$  java -cp .:bin:lib/mysql-connector-java-5.1.12-bin.jar \
  liantze.build.TripleGenerator <task-props> [filterPos]
```

- `<task-props>.properties` is the properties file prepared earlier
- `filterPOS` is optional, to indicate that only LIs of matching POS should be considered when generating triples. Recommended for avoiding many frivolous mappings.

> Have a Go
>
> Generate Thai–English–Chinese translation triples:
>
> **Windows:**
>
> ```
> prompt>  java -cp .;bin;lib\mysql-connector-java-5.1.12-bin.jar ^
>  liantze.build.TripleGenerator props\tha-eng-zho filterPOS
> ```
>
> ***nix and Mac:**
>
> ```
> prompt$  java -cp .:bin:lib/mysql-connector-java-5.1.12-bin.jar \
>  liantze.build.TripleGenerator props/tha-eng-zho filterPOS
> ```

6. Grouping Translation Triples into Trilingual Translation Sets

If you have no existing multilingual lexicon data, use the generated translation triples to create a new one by aggregating the translation triples into translation sets, after filtering them with the threshold values (second half of the modified OTIC algorithm, Appendix B.1).

Windows:

```
prompt>  java -cp .;bin;lib\mysql-connector-java-5.1.12-bin.jar ^
           liantze.build.LexiconExpander <task-props> create-new
```

***nix and Mac:**

```
prompt$  java -cp .:bin:lib/mysql-connector-java-5.1.12-bin.jar \
           liantze.build.LexiconExpander <task-props> create-new
```

7. Adding New Languages to the Multilingual Dictionary

If you already have some multilingual lexicon data, and want to add a new language to it using the generated translation triples, run the `LexiconExpander` tool with the `add-lang` option. The algorithm is in Appendix B.2.

Windows:

```
prompt>  java -cp .;bin;lib\mysql-connector-java-5.1.12-bin.jar ^
           liantze.build.LexiconExpander <task-props> add-lang
```

***nix and Mac:**

```
prompt$  java -cp .:bin:lib/mysql-connector-java-5.1.12-bin.jar \
           liantze.build.LexiconExpander <task-props> add-lang
```

After inspecting the results (e.g. using `search.php` or `randomsets.php`), you may now delete the tables `<triplesPrefix>_temp` and `<triplesPrefix>_triple`.

> Have a Go
>
> Add Thai translation equivalents to `LexiconTX`:
>
> **Windows:**
>
> ```
> prompt>  java -cp .;bin;lib\mysql-connector-java-5.1.12-bin.jar ^
>  liantze.build.LexiconExpander props\tha-eng-zho add-lang
> ```
>
> ***nix or Mac:**
>
> ```
> prompt$  java -cp .:bin:lib/mysql-connector-java-5.1.12-bin.jar \
>  liantze.build.LexiconExpander props/tha-eng-zho add-lang
> ```
>
> You may now delete the tables `tez_temp` and `tez_triple`.

8. Extracting Bilingual Dictionaries from the Multilingual Dictionary

Use the `BilingExtractor` tool to extract a bilingual dictionary, suitable for use with SiSTeC-ebmt, from Lexicon+TX. The syntax is:

Windows:

```
prompt>  java -cp .;bin;lib\mysql-connector-java-5.1.12-bin.jar ^
           liantze.build.BilingDictExtractor ^
           <task-props> <src-lang> <tgt-lang> <output-file>
```

***nix and Mac:**

```
prompt$  java -cp .:bin:lib/mysql-connector-java-5.1.12-bin.jar \
           liantze.build.BilingDictExtractor \
           <task-props> <src-lang> <tgt-lang> <output-file>
```

- Any `<task-props>.properties` file containing the database connection details can be used here.
- `src-lang` and `tgt-lang` are the source and target languages.
- `output-file` will contain the extracted bilingual dictionary. **Note that the translation mappings in this file do not distinguish between senses.**

Have a Go

Extract a French–Thai bilingual dictionary from Lexicon+TX.

Windows:

```
prompt>  java -cp .;bin;lib\mysql-connector-java-5.1.12-bin.jar ^
           liantze.build.BilingDictExtractor props\tha-eng-zho ^
           fra tha fra-tha-lexicon.txt
```

***nix and Mac:**

```
prompt$  java -cp .:bin:lib/mysql-connector-java-5.1.12-bin.jar \
           liantze.build.BilingDictExtractor props\tha-eng-zho \
           fra tha fra-tha-lexicon.txt
```

Excerpt of `fra-tha-lexicon.txt`:

```
...
abattement     N     ความอ่อนเพลีย
abattre        V     ย่อหย่อน
abattre        V     อ่อนกำลัง
abattre        V     อ่อนปวกเปียก
abattre        V     อ่อนเปียก
abbaye         N     พิหาร
abbaye         N     สังฆาวาส
abbé           N     ทชี
abbé           N     นักบวช
abbé           N     บาทหลวง
abbé           N     ใบฎีกา
...
```

References

Boitet, C., Mangeot, M., & Sérasset, G. (2002). The PAPILLON project: Cooperatively building a multilingual lexical database to derive open source dictionaries & lexicons. In *Proceedings of the 2nd Workshop on NLP and XML (NLPXML'02)* (pp. 1–3). Taipei, Taiwan.

Francopoulo, G., Bel, N., George, M., Calzolari, N., Monachini, M., Pet, M., & Soria, C. (2009). Multilingual resources for NLP in the lexical markup framework (LMF). *Language Resources and Evaluation*, *43*(1), 57–70. doi:10.1007/s10579-008-9077-5

International Organization for Standardization. (2008). *ISO 24613:2008 Language resource management – Lexical Markup Framework (LMF)*.

Lim, L. T., Ranaivo-Malançon, B., & Tang, E. K. (2011). Low cost construction of a multilingual lexicon from bilingual lists. *Polibits*, *43*, 45–51.

A. Included Bilingual Dictionaries

SiSTeC-EMDict Available for research. English–Malay, but used as Malay–English.
XDict Fu Jianjun. Included in many GNU/Linux systems, open source. English–Chinese.
CC-CEDICT Collaborative effort, open source. Chinese–English.
FeM Dictionary DBP/USM/GETA-CLIPS. Available for research. French–English–Malay.
Yaitron NecTec, open source. Thai–English.

B. Algorithms

B.1. Generating trilingual translation triples from bilingual translation lists

GenerateTriples($\mathbb{L}_{L_1-L_2}, \mathbb{L}_{L_2-L_3}, \mathbb{L}_{L_2-L_3}$)
FilterSets(T, α, β)
MergeSets(T)

procedure GenerateTriples($\mathbb{L}_{L_1-L_2}, \mathbb{L}_{L_2-L_3}, \mathbb{L}_{L_2-L_3}$)
  $T \leftarrow$ empty set
  for all lexical items $w_h \in L_1$ **do**
    $\mathbb{W}_m \leftarrow$ translations of w_h in L_2 (from $\mathbb{L}_{L_1-L_2}$)
    for all $w_m \in \mathbb{W}_m$ **do**
      $\mathbb{W}_t \leftarrow$ translations of w_m in L_3 (from $\mathbb{L}_{L_2-L_3}$
      for all $w_t \in \mathbb{W}_t$ **do**
        Add translation triple (w_h, w_m, w_t) to T
        $\mathbb{W}_{m_r} \leftarrow$ translations of w_t in L_2 (from $\mathbb{L}_{L_3-L_2}$)
        $$\text{score}(w_h, w_m, w_t) \leftarrow \sum_{w \in \mathbb{W}_m} \frac{\text{no. of common words in } w_{m_r} \in \mathbb{W}_{m_r} \text{ and } w}{\text{no. of words in } w_{m_r} \in \mathbb{W}_{m_r}}$$
      end for
      $$\text{score}(w_h, w_t) \leftarrow 2 \times \frac{\sum_{w \in \mathbb{W}_m} \text{score}(w_h, w, w_t)}{|\mathbb{W}_m| + |\mathbb{W}_{m_r}|}$$
    end for
  end for
end procedure

13

```
procedure FilterTriples(T, α, β)        ▷ T is a set of translation triples (w_h, w_m, w_t) with a score
    for all lexical items w_h ∈ L_1 do
        X ← max_{w_t ∈ 𝕎_t} score(w_h, w_t)
        for all distinct translation pairs (w_h, w_t) do
            if score(w_h, w_t) ≥ αX or (score(w_h, w_t))^2 ≥ βX then
                Place w_h ∈ L_1, w_m ∈ L_2, w_t ∈ L_3 from all triples (w_h, w_..., w_t) in same translation set
                Record score(w_h, w_t) and score(w_h, w_m, w_t)
            else
                Discard all triples (w_h, w_..., w_t)
            end if
        end for
    end for                                         ▷ The sets are now grouped by (w_h, w_t)
end procedure

procedure MergeSets(T)
    Merge all translation sets containing triples with same (w_h, w_m)
    Merge all translation sets containing triples with same (w_m, w_t)
end procedure
```

B.2. Adding L_{k+1} to multilingual lexicon $\mathbb{L}$ of $\{L_1, L_2, \ldots, L_k\}$

```
GenerateTriples(𝕃_{L_{k+1}−L_m}, 𝕃_{L_m−L_n}, 𝕃_{L_n−L_m})        ▷ Other permutations are possible
FilterSets(T, α, β)
AddLang(T, 𝕃_{{L_1,...,L_k}})

procedure AddLang(T, 𝕃_{{L_1,...,L_k}})
    repeat
        cnt ← |T|
        for all (w_{L_{k+1}}, w_{L_m}, w_{L_n}) ∈ T do
            if there exists translation sets in 𝕃 that contains both w_{L_m} and w_{L_n} then
                Add w_{L_{k+1}} to all these translation sets
                Delete (w_{L_{k+1}}, w_{L_m}, w_{L_n}) from T
            end if
        end for
        cnt′ ← |T|
    until cnt = cnt′
    MergeSets(T)
    Add new translation sets to 𝕃_{{L_1,...,L_k}}
end procedure
```

14

APPENDIX E

OTIC FILTERING EVALUATION RESULTS

E.1 Precision, Recall and F_1 for Malay–Chinese Filtering

tp = true positive, fp = false positive, tn = true negative, fn = false negative

Threshold parameters		tp	fp	tn	fn	Precision	Recall	F_1
α	β							
0.0	0.0	274	226	0	0	0.548	1.000	0.708
0.0	0.2	274	226	0	0	0.548	1.000	0.708
0.0	0.4	274	226	0	0	0.548	1.000	0.708
0.0	0.6	274	226	0	0	0.548	1.000	0.708
0.0	0.8	274	226	0	0	0.548	1.000	0.708
0.0	1.0	274	226	0	0	0.548	1.000	0.708
0.2	0.0	274	226	0	0	0.548	1.000	0.708
0.2	0.2	262	188	38	12	0.582	0.956	0.724
0.2	0.4	262	188	38	12	0.582	0.956	0.724
0.2	0.6	262	188	38	12	0.582	0.956	0.724
0.2	0.8	262	188	38	12	0.582	0.956	0.724
0.2	1.0	262	188	38	12	0.582	0.956	0.724
0.4	0.0	274	226	0	0	0.548	1.000	0.708
0.4	0.2	231	132	94	43	0.636	0.843	0.725
0.4	0.4	231	132	94	43	0.636	0.843	0.725
0.4	0.6	231	132	94	43	0.636	0.843	0.725
0.4	0.8	231	132	94	43	0.636	0.843	0.725
0.4	1.0	231	132	94	43	0.636	0.843	0.725
0.6	0.0	274	226	0	0	0.548	1.000	0.708
0.6	0.2	220	120	106	54	0.647	0.803	0.717
0.6	0.4	173	72	154	101	0.706	0.631	0.667
0.6	0.6	173	72	154	101	0.706	0.631	0.667
0.6	0.8	173	72	154	101	0.706	0.631	0.667
0.6	1.0	173	72	154	101	0.706	0.631	0.667
0.8	0.0	274	226	0	0	0.548	1.000	0.708
0.8	0.2	220	120	106	54	0.647	0.803	0.717
0.8	0.4	165	70	156	109	0.702	0.602	0.648
0.8	0.6	140	57	169	134	0.711	0.511	0.594
0.8	0.8	132	56	170	142	0.702	0.482	0.571
0.8	1.0	132	56	170	142	0.702	0.482	0.571
1.0	0.0	274	226	0	0	0.548	1.000	0.708
1.0	0.2	220	120	106	54	0.647	0.803	0.717
1.0	0.4	165	70	156	109	0.702	0.602	0.648
1.0	0.6	140	57	169	134	0.711	0.511	0.594
1.0	0.8	114	38	188	160	0.750	0.416	0.535
1.0	1.0	104	31	195	170	0.770	0.380	0.509

E.2 Precision, Recall and F_1 for Iban–Malay Filtering

tp = true positive, fp = false positive, tn = true negative, fn = false negative

Threshold parameters		tp	fp	tn	fn	Precision	Recall	F_1
α	β							
0.0	0.0	246	254	0	0	0.492	1.000	0.660
0.0	0.2	246	254	0	0	0.492	1.000	0.660
0.0	0.4	246	254	0	0	0.492	1.000	0.660
0.0	0.6	246	254	0	0	0.492	1.000	0.660
0.0	0.8	246	254	0	0	0.492	1.000	0.660
0.0	1.0	246	254	0	0	0.492	1.000	0.660
0.2	0.0	246	254	0	0	0.492	1.000	0.660
0.2	0.2	230	234	20	16	0.496	0.935	0.648
0.2	0.4	230	234	20	16	0.496	0.935	0.648
0.2	0.6	230	234	20	16	0.496	0.935	0.648
0.2	0.8	230	234	20	16	0.496	0.935	0.648
0.2	1.0	230	234	20	16	0.496	0.935	0.648
0.4	0.0	246	254	0	0	0.492	1.000	0.660
0.4	0.2	195	188	66	51	0.509	0.793	0.620
0.4	0.4	195	188	66	51	0.509	0.793	0.620
0.4	0.6	195	188	66	51	0.509	0.793	0.620
0.4	0.8	195	188	66	51	0.509	0.793	0.620
0.4	1.0	195	188	66	51	0.509	0.793	0.620
0.6	0.0	246	254	0	0	0.492	1.000	0.660
0.6	0.2	185	169	85	61	0.523	0.752	0.617
0.6	0.4	143	126	128	103	0.532	0.581	0.555
0.6	0.6	143	126	128	103	0.532	0.581	0.555
0.6	0.8	143	126	128	103	0.532	0.581	0.555
0.6	1.0	143	126	128	103	0.532	0.581	0.555
0.8	0.0	246	254	0	0	0.492	1.000	0.660
0.8	0.2	185	169	85	61	0.523	0.752	0.617
0.8	0.4	135	114	140	111	0.542	0.549	0.545
0.8	0.6	108	90	164	138	0.545	0.439	0.486
0.8	0.8	108	86	168	138	0.557	0.439	0.491
0.8	1.0	108	86	168	138	0.557	0.439	0.491
1.0	0.0	246	254	0	0	0.492	1.000	0.660
1.0	0.2	185	169	85	61	0.523	0.752	0.617
1.0	0.4	135	114	140	111	0.542	0.549	0.545
1.0	0.6	108	90	164	138	0.545	0.439	0.486
1.0	0.8	87	67	187	159	0.565	0.354	0.435
1.0	1.0	77	64	190	169	0.546	0.313	0.398

E.3 Human Judgements and OTIC Filtering Decisions on Malay–Chinese Translation Pairings

Legend: ○ Accept × Reject − Unsure

		Human decision						Decision by OTIC Filtering																																			
Malay	Chinese	Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard	OTIC score	$\alpha = 0, \beta = 0$	$\alpha = 0, \beta = 0.2$	$\alpha = 0, \beta = 0.4$	$\alpha = 0, \beta = 0.6$	$\alpha = 0, \beta = 0.8$	$\alpha = 0, \beta = 1.0$	$\alpha = 0.2, \beta = 0$	$\alpha = 0.2, \beta = 0.2$	$\alpha = 0.2, \beta = 0.4$	$\alpha = 0.2, \beta = 0.6$	$\alpha = 0.2, \beta = 0.8$	$\alpha = 0.2, \beta = 1.0$	$\alpha = 0.4, \beta = 0$	$\alpha = 0.4, \beta = 0.2$	$\alpha = 0.4, \beta = 0.4$	$\alpha = 0.4, \beta = 0.6$	$\alpha = 0.4, \beta = 0.8$	$\alpha = 0.4, \beta = 1.0$	$\alpha = 0.6, \beta = 0$	$\alpha = 0.6, \beta = 0.2$	$\alpha = 0.6, \beta = 0.4$	$\alpha = 0.6, \beta = 0.6$	$\alpha = 0.6, \beta = 0.8$	$\alpha = 0.6, \beta = 1.0$	$\alpha = 0.8, \beta = 0$	$\alpha = 0.8, \beta = 0.2$	$\alpha = 0.8, \beta = 0.4$	$\alpha = 0.8, \beta = 0.6$	$\alpha = 0.8, \beta = 0.8$	$\alpha = 0.8, \beta = 1.0$	$\alpha = 1.0, \beta = 0$	$\alpha = 1.0, \beta = 0.2$	$\alpha = 1.0, \beta = 0.4$	$\alpha = 1.0, \beta = 0.6$	$\alpha = 1.0, \beta = 0.8$	$\alpha = 1.0, \beta = 1.0$
penangguhan	缓刑	×	×	×	○	×	0.182	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
penangguhan	耽搁	○	−	○	○	○	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×
penangguhan	迁延	○	○	−	○	○	0.091	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
penangguhan	悬挂	×	−	−	×	×	0.045	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
melampau	过度	○	○	○	−	○	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
melampau	极端	○	○	○	×	○	0.111	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
melampau	险峻	×	−	×	×	×	0.111	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
melampau	过分	○	○	○	○	○	0.105	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
melampau	奢侈	×	−	−	×	×	0.100	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
melampau	挥霍	×	−	−	×	×	0.100	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
beg kecil	小袋	○	○	○	○	○	1.000	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
perhatian	注意力	−	○	○	○	○	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
perhatian	敬意	×	×	○	×	×	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×
perhatian	备注	○	×	○	×	○	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×
perhatian	尊重	×	×	×	×	×	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×
perhatian	布告	○	×	×	×	×	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×
perhatian	通知	×	×	×	×	×	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×
perhatian	关怀	○	○	−	−	○	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×
satu pihak	单边	○	−	−	○	○	1.000	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
satu pihak	单方面	○	○	○	○	○	1.000	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

Malay	Chinese	Human decision					OTIC score	Decision by OTIC Filtering																																			
		Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard		$\alpha=0,\beta=0$	$\alpha=0,\beta=0.2$	$\alpha=0,\beta=0.4$	$\alpha=0,\beta=0.6$	$\alpha=0,\beta=0.8$	$\alpha=0,\beta=1.0$	$\alpha=0.2,\beta=0$	$\alpha=0.2,\beta=0.2$	$\alpha=0.2,\beta=0.4$	$\alpha=0.2,\beta=0.6$	$\alpha=0.2,\beta=0.8$	$\alpha=0.2,\beta=1.0$	$\alpha=0.4,\beta=0$	$\alpha=0.4,\beta=0.2$	$\alpha=0.4,\beta=0.4$	$\alpha=0.4,\beta=0.6$	$\alpha=0.4,\beta=0.8$	$\alpha=0.4,\beta=1.0$	$\alpha=0.6,\beta=0$	$\alpha=0.6,\beta=0.2$	$\alpha=0.6,\beta=0.4$	$\alpha=0.6,\beta=0.6$	$\alpha=0.6,\beta=0.8$	$\alpha=0.6,\beta=1.0$	$\alpha=0.8,\beta=0$	$\alpha=0.8,\beta=0.2$	$\alpha=0.8,\beta=0.4$	$\alpha=0.8,\beta=0.6$	$\alpha=0.8,\beta=0.8$	$\alpha=0.8,\beta=1.0$	$\alpha=1.0,\beta=0$	$\alpha=1.0,\beta=0.2$	$\alpha=1.0,\beta=0.4$	$\alpha=1.0,\beta=0.6$	$\alpha=1.0,\beta=0.8$	$\alpha=1.0,\beta=1.0$
satu pihak	片面	×	×	○	×	×	0.667	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
memasang	安装	○	○	○	○	○	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
memasang	装置	○	○	○	○	○	0.182	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
memasang	适合	×	×	×	×	×	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
memasang	附属物	×	×	×	×	×	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
memasang	修理	×	×	○	×	×	0.125	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×
memasang	配合	×	×	×	×	×	0.111	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
memasang	系	—	×	○	×	×	0.105	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
memasang	装配	○	○	○	×	○	0.083	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
memasang	扎营	○	×	—	—	○	0.077	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
memasang	附件	×	×	×	×	×	0.051	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
memasang	确定	×	×	×	×	×	0.026	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
menandatangani	签名	○	○	○	○	○	0.563	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
rumbai	小瀑布	×	—	—	—	×	0.667	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
rumbai	缨	—	—	—	—	×	0.440	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
baji	楔形物	—	—	—	—	×	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
baji	楔子	—	—	○	—	○	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
baji	楔形文字	—	—	—	—	×	0.125	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
mendesak	激励	×	×	—	×	×	0.174	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
mendesak	驱使	○	—	○	—	○	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×
mendesak	逼迫	○	○	○	○	○	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×
mendesak	推进	×	×	○	—	×	0.154	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
mendesak	压	×	×	×	×	×	0.120	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
mendesak	推	×	×	×	×	×	0.094	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
mendesak	鼓励	×	×	—	×	×	0.091	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×

Malay	Chinese	Human decision					OTIC score	Decision by OTIC Filtering																																					
		Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard		$\alpha=0,\beta=0$	$\alpha=0,\beta=0.2$	$\alpha=0,\beta=0.4$	$\alpha=0,\beta=0.6$	$\alpha=0,\beta=0.8$	$\alpha=0,\beta=1.0$	$\alpha=0.2,\beta=0$	$\alpha=0.2,\beta=0.2$	$\alpha=0.2,\beta=0.4$	$\alpha=0.2,\beta=0.6$	$\alpha=0.2,\beta=0.8$	$\alpha=0.2,\beta=1.0$	$\alpha=0.4,\beta=0$	$\alpha=0.4,\beta=0.2$	$\alpha=0.4,\beta=0.4$	$\alpha=0.4,\beta=0.6$	$\alpha=0.4,\beta=0.8$	$\alpha=0.4,\beta=1.0$	$\alpha=0.6,\beta=0$	$\alpha=0.6,\beta=0.2$	$\alpha=0.6,\beta=0.4$	$\alpha=0.6,\beta=0.6$	$\alpha=0.6,\beta=0.8$	$\alpha=0.6,\beta=1.0$	$\alpha=0.8,\beta=0$	$\alpha=0.8,\beta=0.2$	$\alpha=0.8,\beta=0.4$	$\alpha=0.8,\beta=0.6$	$\alpha=0.8,\beta=0.8$	$\alpha=0.8,\beta=1.0$	$\alpha=1.0,\beta=0$	$\alpha=1.0,\beta=0.2$	$\alpha=1.0,\beta=0.4$	$\alpha=1.0,\beta=0.6$	$\alpha=1.0,\beta=0.8$	$\alpha=1.0,\beta=1.0$
mendesak	迫切	○	-	○	×	○	0.091	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
mendesak	胁迫	○	—	○	○	○	0.091	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
mendesak	紧迫	○	×	○	×	○	0.091	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
mendesak	紧急	×	×	○	×	×	0.087	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
mendesak	驾驶	×	×	×	×	×	0.087	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
mendesak	劝诫	×	—	×	×	×	0.087	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
mendesak	强迫	○	○	○	○	○	0.087	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
mendesak	推动	×	×	×	×	×	0.051	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
mendesak	坚持	×	×	×	×	×	0.038	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
mendesak	开车	×	×	×	×	×	0.030	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
melakukan	履行	○	—	○	○	○	0.208	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
melakukan	表演	×	×	×	×	×	0.182	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
melakukan	实行	○	○	○	×	○	0.182	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
melakukan	演出	×	×	×	×	×	0.182	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
melakukan	实现	×	○	○	×	○	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	○	○	○	○	×	×
melakukan	执行	○	○	○	○	○	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	○	○	○	○	×	×
menunjukkan	表现	○	—	×	○	○	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
menunjukkan	显示	○	○	○	○	○	0.214	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
menunjukkan	指示	○	○	○	×	○	0.185	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
menunjukkan	展示	○	—	○	○	○	0.179	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
menunjukkan	展览	×	×	×	○	×	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
menunjukkan	陈列	×	—	○	—	○	0.154	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	×	×	×	×
menunjukkan	演示	×	○	○	—	○	0.154	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	×	×	×	×
menunjukkan	表示	○	×	○	×	○	0.133	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
menunjukkan	指出	○	○	○	○	○	0.115	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×

		Human decision						Decision by OTIC Filtering																																			
Malay	Chinese	Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard	OTIC score	$\alpha = 0, \beta = 0$	$\alpha = 0, \beta = 0.2$	$\alpha = 0, \beta = 0.4$	$\alpha = 0, \beta = 0.6$	$\alpha = 0, \beta = 0.8$	$\alpha = 0, \beta = 1.0$	$\alpha = 0.2, \beta = 0$	$\alpha = 0.2, \beta = 0.2$	$\alpha = 0.2, \beta = 0.4$	$\alpha = 0.2, \beta = 0.6$	$\alpha = 0.2, \beta = 0.8$	$\alpha = 0.2, \beta = 1.0$	$\alpha = 0.4, \beta = 0$	$\alpha = 0.4, \beta = 0.2$	$\alpha = 0.4, \beta = 0.4$	$\alpha = 0.4, \beta = 0.6$	$\alpha = 0.4, \beta = 0.8$	$\alpha = 0.4, \beta = 1.0$	$\alpha = 0.6, \beta = 0$	$\alpha = 0.6, \beta = 0.2$	$\alpha = 0.6, \beta = 0.4$	$\alpha = 0.6, \beta = 0.6$	$\alpha = 0.6, \beta = 0.8$	$\alpha = 0.6, \beta = 1.0$	$\alpha = 0.8, \beta = 0$	$\alpha = 0.8, \beta = 0.2$	$\alpha = 0.8, \beta = 0.4$	$\alpha = 0.8, \beta = 0.6$	$\alpha = 0.8, \beta = 0.8$	$\alpha = 0.8, \beta = 1.0$	$\alpha = 1.0, \beta = 0$	$\alpha = 1.0, \beta = 0.2$	$\alpha = 1.0, \beta = 0.4$	$\alpha = 1.0, \beta = 0.6$	$\alpha = 1.0, \beta = 0.8$	$\alpha = 1.0, \beta = 1.0$
menunjukkan	指明	○	○	○	○	○	0.111	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
menunjukkan	展出	○	—	×	—	○	0.103	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
menunjukkan	示范	○	—	×	○	○	0.092	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
menunjukkan	阅读	×	×	×	×	×	0.080	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
menunjukkan	领导	×	×	×	×	×	0.080	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
menunjukkan	反射	×	×	×	×	×	0.080	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
menunjukkan	标明	○	×	○	—	○	0.080	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
menunjukkan	显出	○	○	○	○	○	0.077	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
menunjukkan	呈现	○	○	○	○	○	0.074	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
menunjukkan	看懂	×	×	×	×	×	0.074	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
menunjukkan	读	×	×	×	×	×	0.074	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
menunjukkan	描写	○	○	×	×	○	0.074	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
menunjukkan	表明	○	○	○	—	○	0.071	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
menunjukkan	象征	○	○	×	×	○	0.071	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
menunjukkan	预示	×	○	○	×	○	0.071	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
menunjukkan	反映	○	○	×	×	○	0.069	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
menunjukkan	吹	×	×	×	×	×	0.063	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
menunjukkan	看	×	×	×	×	×	0.053	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
menunjukkan	思考	×	×	×	×	×	0.038	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
menunjukkan	招致	×	×	×	×	×	0.038	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
menunjukkan	赠送	×	×	×	×	×	0.038	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
menunjukkan	论证	×	—	×	×	×	0.037	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
menunjukkan	出现	×	×	×	—	×	0.036	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
menunjukkan	示威	×	×	×	○	×	0.016	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
cemerlang	优秀	○	○	○	○	○	0.235	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

Malay	Chinese	Human decision					OTIC score	Decision by OTIC Filtering																																			
		Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard		$\alpha=0,\beta=0$	$\alpha=0,\beta=0.2$	$\alpha=0,\beta=0.4$	$\alpha=0,\beta=0.6$	$\alpha=0,\beta=0.8$	$\alpha=0,\beta=1.0$	$\alpha=0.2,\beta=0$	$\alpha=0.2,\beta=0.2$	$\alpha=0.2,\beta=0.4$	$\alpha=0.2,\beta=0.6$	$\alpha=0.2,\beta=0.8$	$\alpha=0.2,\beta=1.0$	$\alpha=0.4,\beta=0$	$\alpha=0.4,\beta=0.2$	$\alpha=0.4,\beta=0.4$	$\alpha=0.4,\beta=0.6$	$\alpha=0.4,\beta=0.8$	$\alpha=0.4,\beta=1.0$	$\alpha=0.6,\beta=0$	$\alpha=0.6,\beta=0.2$	$\alpha=0.6,\beta=0.4$	$\alpha=0.6,\beta=0.6$	$\alpha=0.6,\beta=0.8$	$\alpha=0.6,\beta=1.0$	$\alpha=0.8,\beta=0$	$\alpha=0.8,\beta=0.2$	$\alpha=0.8,\beta=0.4$	$\alpha=0.8,\beta=0.6$	$\alpha=0.8,\beta=0.8$	$\alpha=0.8,\beta=1.0$	$\alpha=1.0,\beta=0$	$\alpha=1.0,\beta=0.2$	$\alpha=1.0,\beta=0.4$	$\alpha=1.0,\beta=0.6$	$\alpha=1.0,\beta=0.8$	$\alpha=1.0,\beta=1.0$
cemerlang	辉煌	○	×	○	○	○	0.235	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
cemerlang	光辉	○	×	○	×	○	0.235	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
cemerlang	灿烂	○	×	×	×	×	0.235	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
cemerlang	壮丽	×	×	×	×	×	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×
cemerlang	卓越	○	○	○	○	○	0.211	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×
cemerlang	光亮	×	×	×	×	×	0.125	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
cemerlang	才华横溢	—	○	○	—	○	0.125	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
cemerlang	突出	×	×	○	○	○	0.118	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
cemerlang	胜过	×	×	×	×	×	0.118	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
cemerlang	极好	×	×	○	○	○	0.111	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
cemerlang	显著	×	×	×	○	×	0.111	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
cemerlang	鲜明	×	×	×	×	×	0.111	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
cemerlang	杰出	○	○	○	○	○	0.105	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
cemerlang	聪明	○	○	×	×	○	0.100	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×
setiap masa	始终	×	○	×	×	×	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
setiap masa	日夜	×	×	○	×	×	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
setiap masa	昼夜	×	—	○	×	×	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
setiap masa	一贯	×	—	—	○	○	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
setiap masa	一直	○	○	○	○	○	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
penggulung benang	纺锤	—	—	—	—	×	1.000	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
otot	腱	—	○	—	—	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
otot	肌肉发达	×	×	—	—	×	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
otot	肌肉	○	○	○	○	○	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
otot	膂力	○	○	—	—	○	0.182	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×

Malay	Chinese	Human decision					OTIC score	Decision by OTIC Filtering																																			
		Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard		$\alpha=0,\beta=0$	$\alpha=0,\beta=0.2$	$\alpha=0,\beta=0.4$	$\alpha=0,\beta=0.6$	$\alpha=0,\beta=0.8$	$\alpha=0,\beta=1.0$	$\alpha=0.2,\beta=0$	$\alpha=0.2,\beta=0.2$	$\alpha=0.2,\beta=0.4$	$\alpha=0.2,\beta=0.6$	$\alpha=0.2,\beta=0.8$	$\alpha=0.2,\beta=1.0$	$\alpha=0.4,\beta=0$	$\alpha=0.4,\beta=0.2$	$\alpha=0.4,\beta=0.4$	$\alpha=0.4,\beta=0.6$	$\alpha=0.4,\beta=0.8$	$\alpha=0.4,\beta=1.0$	$\alpha=0.6,\beta=0$	$\alpha=0.6,\beta=0.2$	$\alpha=0.6,\beta=0.4$	$\alpha=0.6,\beta=0.6$	$\alpha=0.6,\beta=0.8$	$\alpha=0.6,\beta=1.0$	$\alpha=0.8,\beta=0$	$\alpha=0.8,\beta=0.2$	$\alpha=0.8,\beta=0.4$	$\alpha=0.8,\beta=0.6$	$\alpha=0.8,\beta=0.8$	$\alpha=0.8,\beta=1.0$	$\alpha=1.0,\beta=0$	$\alpha=1.0,\beta=0.2$	$\alpha=1.0,\beta=0.4$	$\alpha=1.0,\beta=0.6$	$\alpha=1.0,\beta=0.8$	$\alpha=1.0,\beta=1.0$
resonans	谐振	○	×	—	—	○	0.667	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
resonans	共鸣	○	○	—	—	○	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
resonans	共振	×	—	—	—	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
perjanjian	条约	○	○	○	○	○	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
perjanjian	协议	○	○	○	○	○	0.308	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×
perjanjian	协商	○	×	×	○	○	0.182	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
perjanjian	债券	×	×	×	×	×	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
perjanjian	契约	○	○	○	○	○	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
perjanjian	展览	×	×	×	×	×	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
perjanjian	婚约	×	×	×	○	×	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
perjanjian	约定	○	×	○	×	○	0.158	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
perjanjian	约会	×	×	×	×	×	0.154	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
perjanjian	演出	×	×	×	×	×	0.154	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
perjanjian	表演	×	×	×	×	×	0.143	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×
perjanjian	表现	×	×	×	×	×	0.133	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×
perjanjian	便宜货	×	×	×	×	×	0.083	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×
perjanjian	交易	×	×	×	×	×	0.083	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×
perjanjian	协定	○	○	○	○	○	0.033	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×
arak	酒精	○	×	×	○	○	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
arak	酒	○	○	○	○	○	0.120	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
arak	杜松子酒	○	—	×	—	○	0.118	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
arak	灵魂	×	×	×	×	×	0.111	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
arak	气概	×	×	×	×	×	0.105	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
arak	神灵	×	×	×	×	×	0.100	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
arak	精神	×	×	×	×	×	0.080	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×

Malay	Chinese	Human decision					OTIC score	Decision by OTIC Filtering																																					
		Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard		$\alpha=0,\beta=0$	$\alpha=0,\beta=0.2$	$\alpha=0,\beta=0.4$	$\alpha=0,\beta=0.6$	$\alpha=0,\beta=0.8$	$\alpha=0,\beta=1.0$	$\alpha=0.2,\beta=0$	$\alpha=0.2,\beta=0.2$	$\alpha=0.2,\beta=0.4$	$\alpha=0.2,\beta=0.6$	$\alpha=0.2,\beta=0.8$	$\alpha=0.2,\beta=1.0$	$\alpha=0.4,\beta=0$	$\alpha=0.4,\beta=0.2$	$\alpha=0.4,\beta=0.4$	$\alpha=0.4,\beta=0.6$	$\alpha=0.4,\beta=0.8$	$\alpha=0.4,\beta=1.0$	$\alpha=0.6,\beta=0$	$\alpha=0.6,\beta=0.2$	$\alpha=0.6,\beta=0.4$	$\alpha=0.6,\beta=0.6$	$\alpha=0.6,\beta=0.8$	$\alpha=0.6,\beta=1.0$	$\alpha=0.8,\beta=0$	$\alpha=0.8,\beta=0.2$	$\alpha=0.8,\beta=0.4$	$\alpha=0.8,\beta=0.6$	$\alpha=0.8,\beta=0.8$	$\alpha=0.8,\beta=1.0$	$\alpha=1.0,\beta=0$	$\alpha=1.0,\beta=0.2$	$\alpha=1.0,\beta=0.4$	$\alpha=1.0,\beta=0.6$	$\alpha=1.0,\beta=0.8$	$\alpha=1.0,\beta=1.0$
arak	葡萄酒	○	○	×	—	○	0.029	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
arak	烧酒	○	○	×	—	○	0.017	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
tulang	骨头	○	○	○	○	○	0.667	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
tulang	骨	○	○	○	○	○	0.667	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
menggerudi	操练	×	×	×	—	×	0.667	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
menggerudi	训练	×	×	×	—	×	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
menggerudi	钻	○	○	○	—	○	0.182	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
kumpulan	帮	○	○	○	○	○	0.211	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
kumpulan	群	○	○	○	○	○	0.132	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×
kumpulan	聚会	×	×	×	×	×	0.121	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
kumpulan	阀	×	—	—	—	×	0.118	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
kumpulan	集合体	○	○	○	×	○	0.118	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
kumpulan	组	○	○	○	○	○	0.118	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
kumpulan	队	○	×	○	○	○	0.118	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
kumpulan	团体	○	×	○	○	○	0.118	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
kumpulan	积累	○	×	○	○	○	0.063	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
kumpulan	人群	○	○	×	○	○	0.063	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
kumpulan	收藏	×	×	—	×	×	0.063	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
kumpulan	水塘	×	×	×	×	×	0.063	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
kumpulan	随行人员	×	×	×	×	×	0.061	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
kumpulan	派系	×	×	○	×	×	0.061	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
kumpulan	蜂群	×	×	×	×	×	0.061	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
kumpulan	束	×	×	○	×	×	0.061	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
kumpulan	包装	×	×	×	×	×	0.061	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
kumpulan	结	×	×	×	×	×	0.059	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×

Malay	Chinese	Human decision					OTIC score	Decision by OTIC Filtering																																			
		Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard		$\alpha = 0, \beta = 0$	$\alpha = 0, \beta = 0.2$	$\alpha = 0, \beta = 0.4$	$\alpha = 0, \beta = 0.6$	$\alpha = 0, \beta = 0.8$	$\alpha = 0, \beta = 1.0$	$\alpha = 0.2, \beta = 0$	$\alpha = 0.2, \beta = 0.2$	$\alpha = 0.2, \beta = 0.4$	$\alpha = 0.2, \beta = 0.6$	$\alpha = 0.2, \beta = 0.8$	$\alpha = 0.2, \beta = 1.0$	$\alpha = 0.4, \beta = 0$	$\alpha = 0.4, \beta = 0.2$	$\alpha = 0.4, \beta = 0.4$	$\alpha = 0.4, \beta = 0.6$	$\alpha = 0.4, \beta = 0.8$	$\alpha = 0.4, \beta = 1.0$	$\alpha = 0.6, \beta = 0$	$\alpha = 0.6, \beta = 0.2$	$\alpha = 0.6, \beta = 0.4$	$\alpha = 0.6, \beta = 0.6$	$\alpha = 0.6, \beta = 0.8$	$\alpha = 0.6, \beta = 1.0$	$\alpha = 0.8, \beta = 0$	$\alpha = 0.8, \beta = 0.2$	$\alpha = 0.8, \beta = 0.4$	$\alpha = 0.8, \beta = 0.6$	$\alpha = 0.8, \beta = 0.8$	$\alpha = 0.8, \beta = 1.0$	$\alpha = 1.0, \beta = 0$	$\alpha = 1.0, \beta = 0.2$	$\alpha = 1.0, \beta = 0.4$	$\alpha = 1.0, \beta = 0.6$	$\alpha = 1.0, \beta = 0.8$	$\alpha = 1.0, \beta = 1.0$
kumpulan	背包	×	×	×	×	×	0.056	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
kumpulan	种类	○	×	×	×	×	0.051	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
kumpulan	小组	○	×	○	○	○	0.031	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
kumpulan	剧团	×	—	×	○	×	0.031	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
kumpulan	杂集	×	×	×	×	×	0.031	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
kumpulan	政党	×	×	×	—	×	0.031	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
kumpulan	杂录	×	×	×	×	×	0.030	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
kumpulan	一套	×	×	○	—	×	0.029	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
kumpulan	许多	×	×	×	×	×	0.021	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
kumpulan	串	×	×	×	×	×	0.020	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
kumpulan	暴民	×	×	×	×	×	0.016	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
kumpulan	流派	×	×	×	×	×	0.016	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
mengaratkan	生锈	○	○	×	○	○	0.667	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
mengaratkan	侵蚀	×	×	○	×	×	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×
mengaratkan	腐蚀	×	×	○	×	×	0.100	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
indikator	指示剂	○	○	×	×	○	1.000	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
indikator	指示器	○	—	○	○	○	1.000	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
ayam hutan	鹧鸪	○	—	—	—	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
ayam hutan	水鸡	×	×	—	—	×	0.083	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
pengabaian	省略	×	—	○	○	○	0.143	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
kadensa	韵律	○	—	—	—	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
kadensa	节奏	○	—	—	—	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
kadensa	调子	○	—	—	—	○	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
kadensa	抑扬	×	—	—	—	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
seperinduk	担架	○	—	—	—	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

Malay	Chinese	Human decision					OTIC score	Decision by OTIC Filtering																																			
		Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard		$\alpha=0,\beta=0$	$\alpha=0,\beta=0.2$	$\alpha=0,\beta=0.4$	$\alpha=0,\beta=0.6$	$\alpha=0,\beta=0.8$	$\alpha=0,\beta=1.0$	$\alpha=0.2,\beta=0$	$\alpha=0.2,\beta=0.2$	$\alpha=0.2,\beta=0.4$	$\alpha=0.2,\beta=0.6$	$\alpha=0.2,\beta=0.8$	$\alpha=0.2,\beta=1.0$	$\alpha=0.4,\beta=0$	$\alpha=0.4,\beta=0.2$	$\alpha=0.4,\beta=0.4$	$\alpha=0.4,\beta=0.6$	$\alpha=0.4,\beta=0.8$	$\alpha=0.4,\beta=1.0$	$\alpha=0.6,\beta=0$	$\alpha=0.6,\beta=0.2$	$\alpha=0.6,\beta=0.4$	$\alpha=0.6,\beta=0.6$	$\alpha=0.6,\beta=0.8$	$\alpha=0.6,\beta=1.0$	$\alpha=0.8,\beta=0$	$\alpha=0.8,\beta=0.2$	$\alpha=0.8,\beta=0.4$	$\alpha=0.8,\beta=0.6$	$\alpha=0.8,\beta=0.8$	$\alpha=0.8,\beta=1.0$	$\alpha=1.0,\beta=0$	$\alpha=1.0,\beta=0.2$	$\alpha=1.0,\beta=0.4$	$\alpha=1.0,\beta=0.6$	$\alpha=1.0,\beta=0.8$	$\alpha=1.0,\beta=1.0$
bertekak	划船	×	×	×	×	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
bertekak	划	×	×	○	×	×	0.154	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
jawapan	答案	○	○	○	○	○	0.444	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
jawapan	反应	○	×	○	×	○	0.364	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
jawapan	溶解	×	×	×	×	×	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
jawapan	响应	○	×	○	×	○	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
jawapan	解答	○	×	○	○	○	0.240	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
jawapan	关键	×	×	×	×	×	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
jawapan	要害	×	×	×	×	×	0.125	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
jawapan	回答	○	○	○	○	○	0.125	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
jawapan	溶液	×	×	×	×	×	0.083	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
jawapan	键	×	×	×	×	×	0.081	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
jawapan	抗辩	×	×	×	×	×	0.037	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
jawapan	基调	×	×	×	×	×	0.032	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
akne	痤疮	×	○	○	○	○	1.000	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
akne	粉刺	○	○	○	○	○	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
nama	标签	×	—	×	○	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
nama	称号	○	○	○	○	○	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
nama	名称	○	○	○	○	○	0.133	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
nama	姓名	○	○	○	○	○	0.133	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
nama	面额	×	—	×	×	×	0.067	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
nama	名义	○	○	○	○	○	0.048	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
penyambung	桥	×	×	×	—	×	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
penyambung	舰桥	×	×	×	—	×	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
penyambung	桥梁	×	×	×	—	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×

Malay	Chinese	Human decision					OTIC score	Decision by OTIC Filtering																																			
		Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard		$\alpha = 0, \beta = 0$	$\alpha = 0, \beta = 0.2$	$\alpha = 0, \beta = 0.4$	$\alpha = 0, \beta = 0.6$	$\alpha = 0, \beta = 0.8$	$\alpha = 0, \beta = 1.0$	$\alpha = 0.2, \beta = 0$	$\alpha = 0.2, \beta = 0.2$	$\alpha = 0.2, \beta = 0.4$	$\alpha = 0.2, \beta = 0.6$	$\alpha = 0.2, \beta = 0.8$	$\alpha = 0.2, \beta = 1.0$	$\alpha = 0.4, \beta = 0$	$\alpha = 0.4, \beta = 0.2$	$\alpha = 0.4, \beta = 0.4$	$\alpha = 0.4, \beta = 0.6$	$\alpha = 0.4, \beta = 0.8$	$\alpha = 0.4, \beta = 1.0$	$\alpha = 0.6, \beta = 0$	$\alpha = 0.6, \beta = 0.2$	$\alpha = 0.6, \beta = 0.4$	$\alpha = 0.6, \beta = 0.6$	$\alpha = 0.6, \beta = 0.8$	$\alpha = 0.6, \beta = 1.0$	$\alpha = 0.8, \beta = 0$	$\alpha = 0.8, \beta = 0.2$	$\alpha = 0.8, \beta = 0.4$	$\alpha = 0.8, \beta = 0.6$	$\alpha = 0.8, \beta = 0.8$	$\alpha = 0.8, \beta = 1.0$	$\alpha = 1.0, \beta = 0$	$\alpha = 1.0, \beta = 0.2$	$\alpha = 1.0, \beta = 0.4$	$\alpha = 1.0, \beta = 0.6$	$\alpha = 1.0, \beta = 0.8$	$\alpha = 1.0, \beta = 1.0$
penyambung	桥牌	×	×	×	—	×	0.080	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
penyambung	链接	○	○	○	○	○	0.080	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
ukuran	尺寸	○	○	○	○	○	0.545	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
ukuran	大小	○	×	○	○	○	0.429	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	○	○	○	○	×	×
ukuran	测量	○	×	○	○	○	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
ukuran	措施	×	×	×	×	×	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
ukuran	尺码	○	○	○	○	○	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
ukuran	标准	×	×	×	×	×	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
ukuran	轨范	×	×	×	×	×	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
ukuran	等级	×	×	×	×	×	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
ukuran	度量	×	×	×	○	×	0.182	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
ukuran	速度	×	×	×	×	×	0.182	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
ukuran	比率	×	×	×	×	×	0.182	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
ukuran	率	×	×	×	×	×	0.182	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
sedih	悲惨	×	○	○	—	○	0.138	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
sedih	悲伤	○	○	○	○	○	0.071	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
sedih	不幸	○	×	×	—	×	0.069	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
sedih	遗憾	×	×	×	○	×	0.069	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
sedih	怜悯	×	×	×	—	×	0.069	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
sedih	哀伤	○	○	○	○	○	0.067	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
sedih	沮丧	○	○	○	○	○	0.067	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
sedih	哀怨	×	×	○	○	○	0.067	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
sedih	可惜	×	×	×	—	×	0.023	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
tetuang	号角	—	—	○	—	○	0.667	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
tetuang	角	—	—	○	—	○	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×

		Human decision						Decision by OTIC Filtering																																	
Malay	Chinese	Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard	OTIC score	$\alpha = 0, \beta = 0$	$\alpha = 0, \beta = 0.2$	$\alpha = 0, \beta = 0.4$	$\alpha = 0, \beta = 0.6$	$\alpha = 0, \beta = 0.8$	$\alpha = 0, \beta = 1.0$	$\alpha = 0.2, \beta = 0$	$\alpha = 0.2, \beta = 0.2$	$\alpha = 0.2, \beta = 0.4$	$\alpha = 0.2, \beta = 0.6$	$\alpha = 0.2, \beta = 0.8$	$\alpha = 0.2, \beta = 1.0$	$\alpha = 0.4, \beta = 0$	$\alpha = 0.4, \beta = 0.2$	$\alpha = 0.4, \beta = 0.4$	$\alpha = 0.4, \beta = 0.6$	$\alpha = 0.4, \beta = 0.8$	$\alpha = 0.4, \beta = 1.0$	$\alpha = 0.6, \beta = 0$	$\alpha = 0.6, \beta = 0.2$	$\alpha = 0.6, \beta = 0.4$	$\alpha = 0.6, \beta = 0.6$	$\alpha = 0.6, \beta = 0.8$	$\alpha = 0.6, \beta = 1.0$	$\alpha = 0.8, \beta = 0$	$\alpha = 0.8, \beta = 0.2$	$\alpha = 0.8, \beta = 0.4$	$\alpha = 0.8, \beta = 0.6$	$\alpha = 0.8, \beta = 0.8$	$\alpha = 0.8, \beta = 1.0$	$\alpha = 1.0, \beta = 0$	$\alpha = 1.0, \beta = 0.2$	$\alpha = 1.0, \beta = 0.4$	$\alpha = 1.0, \beta = 0.6$	$\alpha = 1.0, \beta = 0.8$	$\alpha = 1.0, \beta = 1.0$
tetuang	喇叭	—	—	○	—	○	0.033	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
memperdayakan	欺骗	○	○	○	—	○	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
memperdayakan	行骗	○	○	○	—	○	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
memperdayakan	哄骗	○	○	○	○	○	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
bunga primrose	报春花	○	—	○	—	○	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
yuran	贡献	×	×	×	—	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
yuran	订阅	×	○	×	—	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
yuran	捐献	×	×	×	—	×	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
panggil	打电话	×	○	○	○	○	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
panggil	访问	×	×	×	—	×	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
senyap	寂静	○	○	○	—	○	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
senyap	无声	○	○	○	○	○	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
senyap	安静	○	○	○	○	○	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
senyap	静悄悄	○	○	○	○	○	0.143	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
mengimport	进口	○	○	○	○	○	0.667	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
mengimport	输入	○	○	○	○	○	0.667	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
lapan	八字形	×	—	×	×	×	0.133	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
di tengah-tengah	心脏	×	×	×	×	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
di tengah-tengah	心	×	×	×	×	×	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
di tengah-tengah	内心	×	×	×	×	×	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
di tengah-tengah	中心	○	○	○	○	○	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×

Malay	Chinese	Human decision					OTIC score	Decision by OTIC Filtering																																			
		Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard		$\alpha=0,\beta=0$	$\alpha=0,\beta=0.2$	$\alpha=0,\beta=0.4$	$\alpha=0,\beta=0.6$	$\alpha=0,\beta=0.8$	$\alpha=0,\beta=1.0$	$\alpha=0.2,\beta=0$	$\alpha=0.2,\beta=0.2$	$\alpha=0.2,\beta=0.4$	$\alpha=0.2,\beta=0.6$	$\alpha=0.2,\beta=0.8$	$\alpha=0.2,\beta=1.0$	$\alpha=0.4,\beta=0$	$\alpha=0.4,\beta=0.2$	$\alpha=0.4,\beta=0.4$	$\alpha=0.4,\beta=0.6$	$\alpha=0.4,\beta=0.8$	$\alpha=0.4,\beta=1.0$	$\alpha=0.6,\beta=0$	$\alpha=0.6,\beta=0.2$	$\alpha=0.6,\beta=0.4$	$\alpha=0.6,\beta=0.6$	$\alpha=0.6,\beta=0.8$	$\alpha=0.6,\beta=1.0$	$\alpha=0.8,\beta=0$	$\alpha=0.8,\beta=0.2$	$\alpha=0.8,\beta=0.4$	$\alpha=0.8,\beta=0.6$	$\alpha=0.8,\beta=0.8$	$\alpha=0.8,\beta=1.0$	$\alpha=1.0,\beta=0$	$\alpha=1.0,\beta=0.2$	$\alpha=1.0,\beta=0.4$	$\alpha=1.0,\beta=0.6$	$\alpha=1.0,\beta=0.8$	$\alpha=1.0,\beta=1.0$
mencelahi	拦截	○	×	○	—	○	0.667	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
mencelahi	介入	○	×	○	—	○	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
mencelahi	干预	×	×	○	○	○	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
pengucapan	表情	×	—	×	×	×	0.278	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
pengucapan	发音	×	—	×	×	×	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×
pengucapan	措辞	×	—	○	○	○	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
pengucapan	读法	×	—	○	×	×	0.037	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
melodi	曲调	○	○	○	○	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
melodi	曲子	○	○	○	○	○	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
melodi	空气	×	×	×	×	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
melodi	空军	×	×	×	×	×	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
hasrat	心愿	○	○	○	○	○	0.467	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
hasrat	愿望	○	○	○	○	○	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
hasrat	祝愿	×	×	○	○	○	0.182	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
hasrat	憧憬	○	×	○	×	○	0.167	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
hasrat	雄心	×	×	○	×	×	0.167	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
hasrat	野心	×	×	○	×	×	0.154	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
seligi	投射	—	—	×	—	×	0.667	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
catatan	注解	○	—	○	○	○	0.231	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
catatan	记录	○	○	○	○	○	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
catatan	编年史	×	—	×	○	×	0.182	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	○	○	○	○	×	×
catatan	备忘录	○	×	○	○	○	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
catatan	档案	×	×	×	○	×	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
catatan	评注	○	×	○	×	○	0.154	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
catatan	唱片	×	×	×	×	×	0.091	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×

Malay	Chinese	Human decision					OTIC score	Decision by OTIC Filtering																																			
		Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard		$\alpha = 0, \beta = 0$	$\alpha = 0, \beta = 0.2$	$\alpha = 0, \beta = 0.4$	$\alpha = 0, \beta = 0.6$	$\alpha = 0, \beta = 0.8$	$\alpha = 0, \beta = 1.0$	$\alpha = 0.2, \beta = 0$	$\alpha = 0.2, \beta = 0.2$	$\alpha = 0.2, \beta = 0.4$	$\alpha = 0.2, \beta = 0.6$	$\alpha = 0.2, \beta = 0.8$	$\alpha = 0.2, \beta = 1.0$	$\alpha = 0.4, \beta = 0$	$\alpha = 0.4, \beta = 0.2$	$\alpha = 0.4, \beta = 0.4$	$\alpha = 0.4, \beta = 0.6$	$\alpha = 0.4, \beta = 0.8$	$\alpha = 0.4, \beta = 1.0$	$\alpha = 0.6, \beta = 0$	$\alpha = 0.6, \beta = 0.2$	$\alpha = 0.6, \beta = 0.4$	$\alpha = 0.6, \beta = 0.6$	$\alpha = 0.6, \beta = 0.8$	$\alpha = 0.6, \beta = 1.0$	$\alpha = 0.8, \beta = 0$	$\alpha = 0.8, \beta = 0.2$	$\alpha = 0.8, \beta = 0.4$	$\alpha = 0.8, \beta = 0.6$	$\alpha = 0.8, \beta = 0.8$	$\alpha = 0.8, \beta = 1.0$	$\alpha = 1.0, \beta = 0$	$\alpha = 1.0, \beta = 0.2$	$\alpha = 1.0, \beta = 0.4$	$\alpha = 1.0, \beta = 0.6$	$\alpha = 1.0, \beta = 0.8$	$\alpha = 1.0, \beta = 1.0$
catatan	照会	×	×	×	×	×	0.067	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
catatan	条目	×	×	×	×	×	0.045	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
tidak ketara	无形	—	×	×	—	×	0.381	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
tidak ketara	看不见	×	×	○	—	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
setan	魔鬼	○	○	○	—	○	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
berjanji	允诺	○	○	○	○	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
berjanji	答应	○	○	○	○	○	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
berjanji	约定	○	○	○	○	○	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
menggambarkan	描写	○	○	○	○	○	0.545	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
menggambarkan	描绘	○	○	○	○	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
menggambarkan	反射	×	×	×	×	×	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
menggambarkan	描述	○	○	○	○	○	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
menggambarkan	刻画	×	—	○	○	○	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
menggambarkan	反映	○	○	○	×	○	0.154	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
menggambarkan	思考	○	×	×	×	×	0.100	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
menggambarkan	拍摄	○	×	○	○	○	0.050	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
penyelamatan	救出	○	×	○	○	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
penurut	从者	○	○	○	○	○	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
kendala	障碍物	—	—	○	—	○	0.667	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
kendala	障碍	—	—	○	—	○	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
susunan	排列	○	○	○	○	○	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
susunan	队列	○	○	○	×	○	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×
susunan	顺序	○	○	○	○	○	0.182	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
susunan	命令	×	×	×	×	×	0.182	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
susunan	次序	○	○	○	○	○	0.182	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×

Malay	Chinese	Human decision					OTIC score	Decision by OTIC Filtering																																			
		Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard		$\alpha=0,\beta=0$	$\alpha=0,\beta=0.2$	$\alpha=0,\beta=0.4$	$\alpha=0,\beta=0.6$	$\alpha=0,\beta=0.8$	$\alpha=0,\beta=1.0$	$\alpha=0.2,\beta=0$	$\alpha=0.2,\beta=0.2$	$\alpha=0.2,\beta=0.4$	$\alpha=0.2,\beta=0.6$	$\alpha=0.2,\beta=0.8$	$\alpha=0.2,\beta=1.0$	$\alpha=0.4,\beta=0$	$\alpha=0.4,\beta=0.2$	$\alpha=0.4,\beta=0.4$	$\alpha=0.4,\beta=0.6$	$\alpha=0.4,\beta=0.8$	$\alpha=0.4,\beta=1.0$	$\alpha=0.6,\beta=0$	$\alpha=0.6,\beta=0.2$	$\alpha=0.6,\beta=0.4$	$\alpha=0.6,\beta=0.6$	$\alpha=0.6,\beta=0.8$	$\alpha=0.6,\beta=1.0$	$\alpha=0.8,\beta=0$	$\alpha=0.8,\beta=0.2$	$\alpha=0.8,\beta=0.4$	$\alpha=0.8,\beta=0.6$	$\alpha=0.8,\beta=0.8$	$\alpha=0.8,\beta=1.0$	$\alpha=1.0,\beta=0$	$\alpha=1.0,\beta=0.2$	$\alpha=1.0,\beta=0.4$	$\alpha=1.0,\beta=0.6$	$\alpha=1.0,\beta=0.8$	$\alpha=1.0,\beta=1.0$
susunan	建筑物	×	×	×	×	×	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
susunan	阵容	×	×	×	×	×	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
susunan	结构	×	×	×	×	×	0.154	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
sekejap-sekejap	间歇	×	—	○	—	○	1.000	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
sekejap-sekejap	断断续续	○	○	○	—	○	0.667	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
tupai	松鼠	○	○	○	—	○	1.000	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
berlatih	排练	○	○	○	○	○	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
berlatih	实习	○	○	○	○	○	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
berlatih	实践	×	—	×	×	×	0.444	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
berlatih	预演	○	○	○	○	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
berlatih	开业	×	×	×	×	×	0.133	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
berlatih	排演	×	○	○	○	○	0.125	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
berlatih	实行	○	×	×	×	×	0.111	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
berlatih	行使	×	—	×	×	×	0.100	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
berlatih	锻炼	○	○	○	○	○	0.083	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
birokrat	官僚	○	—	○	○	○	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
mendung	阴沉	○	○	○	○	○	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
mendung	云	×	×	×	—	×	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
mendung	萧瑟	×	—	×	×	×	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
mendung	阴暗	×	○	○	×	○	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
mendung	阴冷	×	○	×	×	×	0.083	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
mencelah	打断	○	—	○	—	○	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
mencelah	插嘴	×	—	○	—	○	0.056	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×

Malay	Chinese	Human decision					OTIC score	Decision by OTIC Filtering																																			
		Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard		$\alpha=0,\beta=0$	$\alpha=0,\beta=0.2$	$\alpha=0,\beta=0.4$	$\alpha=0,\beta=0.6$	$\alpha=0,\beta=0.8$	$\alpha=0,\beta=1.0$	$\alpha=0.2,\beta=0$	$\alpha=0.2,\beta=0.2$	$\alpha=0.2,\beta=0.4$	$\alpha=0.2,\beta=0.6$	$\alpha=0.2,\beta=0.8$	$\alpha=0.2,\beta=1.0$	$\alpha=0.4,\beta=0$	$\alpha=0.4,\beta=0.2$	$\alpha=0.4,\beta=0.4$	$\alpha=0.4,\beta=0.6$	$\alpha=0.4,\beta=0.8$	$\alpha=0.4,\beta=1.0$	$\alpha=0.6,\beta=0$	$\alpha=0.6,\beta=0.2$	$\alpha=0.6,\beta=0.4$	$\alpha=0.6,\beta=0.6$	$\alpha=0.6,\beta=0.8$	$\alpha=0.6,\beta=1.0$	$\alpha=0.8,\beta=0$	$\alpha=0.8,\beta=0.2$	$\alpha=0.8,\beta=0.4$	$\alpha=0.8,\beta=0.6$	$\alpha=0.8,\beta=0.8$	$\alpha=0.8,\beta=1.0$	$\alpha=1.0,\beta=0$	$\alpha=1.0,\beta=0.2$	$\alpha=1.0,\beta=0.4$	$\alpha=1.0,\beta=0.6$	$\alpha=1.0,\beta=0.8$	$\alpha=1.0,\beta=1.0$
keanjalan	弹性	○	○	○	—	○	0.444	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
keanjalan	机动性	×	○	○	—	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×
keanjalan	灵活性	○	○	○	—	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×
pameran	展览	○	○	○	○	○	0.444	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
pameran	展览会	○	○	○	○	○	0.444	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
pameran	表演	×	×	×	×	×	0.364	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
pameran	表现	×	×	×	×	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
pameran	演出	○	×	×	×	×	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
pameran	排列	×	×	×	×	×	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
pameran	庙会	×	×	×	×	×	0.125	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
penggabungan	巩固	×	—	×	—	×	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
penggabungan	盟友	○	×	×	—	×	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
penggabungan	同盟国	○	×	×	—	×	0.182	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×
gadis	处女	○	×	○	—	○	0.182	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
gadis	少女	○	○	○	○	○	0.182	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
gadis	女孩子	○	○	○	○	○	0.182	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
gadis	闺女	×	—	○	○	○	0.154	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
gadis	姑娘	○	○	○	○	○	0.143	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	○	○	○	○	×	×
koma bertitik	分号	×	○	○	—	○	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
memeriksa	校验	○	—	○	○	○	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
memeriksa	检查	○	○	○	○	○	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
memeriksa	检验	○	○	○	○	○	0.235	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×
memeriksa	查看	○	○	○	○	○	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
memeriksa	参观	×	×	×	×	×	0.176	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×

Malay	Chinese	Human decision					OTIC score	Decision by OTIC Filtering																																			
		Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard		$\alpha=0,\beta=0$	$\alpha=0,\beta=0.2$	$\alpha=0,\beta=0.4$	$\alpha=0,\beta=0.6$	$\alpha=0,\beta=0.8$	$\alpha=0,\beta=1.0$	$\alpha=0.2,\beta=0$	$\alpha=0.2,\beta=0.2$	$\alpha=0.2,\beta=0.4$	$\alpha=0.2,\beta=0.6$	$\alpha=0.2,\beta=0.8$	$\alpha=0.2,\beta=1.0$	$\alpha=0.4,\beta=0$	$\alpha=0.4,\beta=0.2$	$\alpha=0.4,\beta=0.4$	$\alpha=0.4,\beta=0.6$	$\alpha=0.4,\beta=0.8$	$\alpha=0.4,\beta=1.0$	$\alpha=0.6,\beta=0$	$\alpha=0.6,\beta=0.2$	$\alpha=0.6,\beta=0.4$	$\alpha=0.6,\beta=0.6$	$\alpha=0.6,\beta=0.8$	$\alpha=0.6,\beta=1.0$	$\alpha=0.8,\beta=0$	$\alpha=0.8,\beta=0.2$	$\alpha=0.8,\beta=0.4$	$\alpha=0.8,\beta=0.6$	$\alpha=0.8,\beta=0.8$	$\alpha=0.8,\beta=1.0$	$\alpha=1.0,\beta=0$	$\alpha=1.0,\beta=0.2$	$\alpha=1.0,\beta=0.4$	$\alpha=1.0,\beta=0.6$	$\alpha=1.0,\beta=0.8$	$\alpha=1.0,\beta=1.0$
memeriksa	看	×	×	×	×	×	0.143	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
memeriksa	经历	×	×	×	×	×	0.125	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
memeriksa	证实	×	×	×	○	×	0.125	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
memeriksa	调查	○	○	○	○	○	0.125	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
memeriksa	看见	×	×	×	×	×	0.125	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
memeriksa	纠正	×	×	×	×	×	0.125	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
memeriksa	检阅	○	○	○	○	○	0.125	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
memeriksa	显得	×	×	×	×	×	0.118	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
memeriksa	矫正	×	×	×	×	×	0.111	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
memeriksa	过去	×	×	×	×	×	0.105	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×
memeriksa	瞧	×	×	×	○	×	0.067	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
memeriksa	望	×	×	×	×	×	0.053	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
memeriksa	核对	×	×	○	○	○	0.042	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
memeriksa	校正	×	×	○	○	○	0.042	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
mencongak	计算	×	—	○	—	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
silikat	硅酸盐	×	—	○	—	○	1.000	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
menghad	限制	○	○	○	○	○	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
menghad	约束	○	○	○	○	○	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
menghad	限定	○	○	○	○	○	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
cekap	能干	○	○	○	○	○	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
cekap	熟练	○	○	○	○	○	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
cekap	灵巧	○	○	○	—	○	0.133	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
cekap	精通	○	○	○	—	○	0.133	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
cekap	内行	×	×	×	—	×	0.105	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×
cekap	胜任	×	×	×	—	×	0.042	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×

Malay	Chinese	Human decision					OTIC score	Decision by OTIC Filtering																																			
		Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard		$\alpha=0,\beta=0$	$\alpha=0,\beta=0.2$	$\alpha=0,\beta=0.4$	$\alpha=0,\beta=0.6$	$\alpha=0,\beta=0.8$	$\alpha=0,\beta=1.0$	$\alpha=0.2,\beta=0$	$\alpha=0.2,\beta=0.2$	$\alpha=0.2,\beta=0.4$	$\alpha=0.2,\beta=0.6$	$\alpha=0.2,\beta=0.8$	$\alpha=0.2,\beta=1.0$	$\alpha=0.4,\beta=0$	$\alpha=0.4,\beta=0.2$	$\alpha=0.4,\beta=0.4$	$\alpha=0.4,\beta=0.6$	$\alpha=0.4,\beta=0.8$	$\alpha=0.4,\beta=1.0$	$\alpha=0.6,\beta=0$	$\alpha=0.6,\beta=0.2$	$\alpha=0.6,\beta=0.4$	$\alpha=0.6,\beta=0.6$	$\alpha=0.6,\beta=0.8$	$\alpha=0.6,\beta=1.0$	$\alpha=0.8,\beta=0$	$\alpha=0.8,\beta=0.2$	$\alpha=0.8,\beta=0.4$	$\alpha=0.8,\beta=0.6$	$\alpha=0.8,\beta=0.8$	$\alpha=0.8,\beta=1.0$	$\alpha=1.0,\beta=0$	$\alpha=1.0,\beta=0.2$	$\alpha=1.0,\beta=0.4$	$\alpha=1.0,\beta=0.6$	$\alpha=1.0,\beta=0.8$	$\alpha=1.0,\beta=1.0$
mencucuh	点燃	—	—	○	○	○	0.800	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
mencucuh	解雇	—	—	×	—	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
mencucuh	开枪	—	—	○	—	○	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
mencucuh	放枪	—	—	○	—	○	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
mencucuh	烧制	—	—	×	—	×	0.125	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
hujung minggu	周末	○	○	○	○	○	1.000	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
orang gaji	女管家	×	○	×	○	○	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
orang gaji	仆人	○	○	○	○	○	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
orang gaji	保姆	×	○	×	—	×	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
orang gaji	女仆	○	○	○	○	○	0.125	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
setiap orang	每人	○	○	○	○	○	0.273	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
setiap orang	各自	○	○	○	○	○	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
setiap orang	人人	○	○	○	○	○	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
memperbalah	争执	○	○	○	—	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
tanggung	支撑	○	○	○	—	○	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
tanggung	支援	○	×	—	○	○	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
tanggung	支持	×	○	○	○	○	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
tanggung	供养	○	○	○	○	○	0.133	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
larva	幼虫	○	○	○	—	○	0.667	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
puncak	峰	○	○	○	○	○	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
puncak	顶	○	○	○	○	○	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
puncak	尖峰	○	○	○	○	○	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	○	○	○	○	×	×
puncak	顶端	○	—	○	○	○	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	○	○	○	○	×	×
puncak	顶点	○	○	○	○	○	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	○	○	○	○	×	×

Malay	Chinese	Human decision					OTIC score	Decision by OTIC Filtering																																			
		Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard		$\alpha=0,\beta=0$	$\alpha=0,\beta=0.2$	$\alpha=0,\beta=0.4$	$\alpha=0,\beta=0.6$	$\alpha=0,\beta=0.8$	$\alpha=0,\beta=1.0$	$\alpha=0.2,\beta=0$	$\alpha=0.2,\beta=0.2$	$\alpha=0.2,\beta=0.4$	$\alpha=0.2,\beta=0.6$	$\alpha=0.2,\beta=0.8$	$\alpha=0.2,\beta=1.0$	$\alpha=0.4,\beta=0$	$\alpha=0.4,\beta=0.2$	$\alpha=0.4,\beta=0.4$	$\alpha=0.4,\beta=0.6$	$\alpha=0.4,\beta=0.8$	$\alpha=0.4,\beta=1.0$	$\alpha=0.6,\beta=0$	$\alpha=0.6,\beta=0.2$	$\alpha=0.6,\beta=0.4$	$\alpha=0.6,\beta=0.6$	$\alpha=0.6,\beta=0.8$	$\alpha=0.6,\beta=1.0$	$\alpha=0.8,\beta=0$	$\alpha=0.8,\beta=0.2$	$\alpha=0.8,\beta=0.4$	$\alpha=0.8,\beta=0.6$	$\alpha=0.8,\beta=0.8$	$\alpha=0.8,\beta=1.0$	$\alpha=1.0,\beta=0$	$\alpha=1.0,\beta=0.2$	$\alpha=1.0,\beta=0.4$	$\alpha=1.0,\beta=0.6$	$\alpha=1.0,\beta=0.8$	$\alpha=1.0,\beta=1.0$
puncak	高峰	○	○	○	○	○	0.211	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
puncak	冠	×	×	○	○	○	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
puncak	顶峰	○	○	○	○	○	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
puncak	天顶	×	×	×	—	×	0.118	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
puncak	王冠	×	×	×	×	×	0.118	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
puncak	顶部	×	×	○	○	○	0.117	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
puncak	头	×	×	×	×	×	0.107	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
puncak	矛	×	×	×	×	×	0.105	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
puncak	脑袋	×	×	×	×	×	0.100	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
puncak	高潮	×	○	○	×	○	0.091	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
puncak	头顶	×	×	×	×	×	0.059	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
puncak	尖端	×	×	×	×	×	0.050	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
rapat	示威运动	×	×	×	×	×	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
rapat	亲密	○	○	○	○	○	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
rapat	亲昵	○	○	○	○	○	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
rapat	熟悉	○	○	○	×	○	0.083	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
seteru	对手	—	○	○	○	○	0.667	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
seteru	仇敌	—	○	○	—	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×
seteru	敌人	—	○	○	—	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×
seteru	敌手	—	—	○	—	○	0.375	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
seteru	敌军	—	—	○	—	○	0.167	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
prajadian	先例	×	—	○	—	○	1.000	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
prajadian	前例	○	—	○	—	○	0.667	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
prajadian	前事	○	—	○	○	○	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
menyamun	洗劫	○	○	○	○	○	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

Malay	Chinese	Human decision					OTIC score	Decision by OTIC Filtering																																			
		Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard		$\alpha=0,\beta=0$	$\alpha=0,\beta=0.2$	$\alpha=0,\beta=0.4$	$\alpha=0,\beta=0.6$	$\alpha=0,\beta=0.8$	$\alpha=0,\beta=1.0$	$\alpha=0.2,\beta=0$	$\alpha=0.2,\beta=0.2$	$\alpha=0.2,\beta=0.4$	$\alpha=0.2,\beta=0.6$	$\alpha=0.2,\beta=0.8$	$\alpha=0.2,\beta=1.0$	$\alpha=0.4,\beta=0$	$\alpha=0.4,\beta=0.2$	$\alpha=0.4,\beta=0.4$	$\alpha=0.4,\beta=0.6$	$\alpha=0.4,\beta=0.8$	$\alpha=0.4,\beta=1.0$	$\alpha=0.6,\beta=0$	$\alpha=0.6,\beta=0.2$	$\alpha=0.6,\beta=0.4$	$\alpha=0.6,\beta=0.6$	$\alpha=0.6,\beta=0.8$	$\alpha=0.6,\beta=1.0$	$\alpha=0.8,\beta=0$	$\alpha=0.8,\beta=0.2$	$\alpha=0.8,\beta=0.4$	$\alpha=0.8,\beta=0.6$	$\alpha=0.8,\beta=0.8$	$\alpha=0.8,\beta=1.0$	$\alpha=1.0,\beta=0$	$\alpha=1.0,\beta=0.2$	$\alpha=1.0,\beta=0.4$	$\alpha=1.0,\beta=0.6$	$\alpha=1.0,\beta=0.8$	$\alpha=1.0,\beta=1.0$
menyamun	抢劫	○	○	○	○	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
menyamun	掠夺	○	—	○	○	○	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
sengsara	受苦	○	○	○	○	○	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
sengsara	苦难	○	○	○	○	○	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
sengsara	悲惨	○	○	○	○	○	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
sengsara	苦痛	○	○	○	○	○	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
sengsara	遭受	○	×	×	×	×	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
sengsara	痛苦	○	○	○	○	○	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
sengsara	受到	○	×	×	×	×	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
pelengkap	补足物	—	○	○	—	○	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
pelengkap	增刊	—	×	×	—	×	0.083	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
pelengkap	余角	○	×	×	—	×	0.050	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
menewaskan	打败	○	○	○	○	○	0.444	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
menewaskan	击败	○	○	○	○	○	0.308	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
menewaskan	压倒	○	○	○	×	○	0.154	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
menewaskan	打	×	×	×	×	×	0.080	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
menewaskan	反击	○	×	×	×	×	0.071	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
menewaskan	克服	×	×	×	×	×	0.017	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
samping	方面	○	—	×	—	○	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
samping	边	○	—	○	—	○	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	○	○	○	○	×	×
samping	面	×	—	×	—	×	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
samping	侧面	×	—	×	—	×	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
samping	一方	×	—	×	○	×	0.143	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×

Malay	Chinese	Human decision					OTIC score	Decision by OTIC Filtering																																			
		Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard		$\alpha=0,\beta=0$	$\alpha=0,\beta=0.2$	$\alpha=0,\beta=0.4$	$\alpha=0,\beta=0.6$	$\alpha=0,\beta=0.8$	$\alpha=0,\beta=1.0$	$\alpha=0.2,\beta=0$	$\alpha=0.2,\beta=0.2$	$\alpha=0.2,\beta=0.4$	$\alpha=0.2,\beta=0.6$	$\alpha=0.2,\beta=0.8$	$\alpha=0.2,\beta=1.0$	$\alpha=0.4,\beta=0$	$\alpha=0.4,\beta=0.2$	$\alpha=0.4,\beta=0.4$	$\alpha=0.4,\beta=0.6$	$\alpha=0.4,\beta=0.8$	$\alpha=0.4,\beta=1.0$	$\alpha=0.6,\beta=0$	$\alpha=0.6,\beta=0.2$	$\alpha=0.6,\beta=0.4$	$\alpha=0.6,\beta=0.6$	$\alpha=0.6,\beta=0.8$	$\alpha=0.6,\beta=1.0$	$\alpha=0.8,\beta=0$	$\alpha=0.8,\beta=0.2$	$\alpha=0.8,\beta=0.4$	$\alpha=0.8,\beta=0.6$	$\alpha=0.8,\beta=0.8$	$\alpha=0.8,\beta=1.0$	$\alpha=1.0,\beta=0$	$\alpha=1.0,\beta=0.2$	$\alpha=1.0,\beta=0.4$	$\alpha=1.0,\beta=0.6$	$\alpha=1.0,\beta=0.8$	$\alpha=1.0,\beta=1.0$
memberi sebab-musabab	说明	○	○	○	○	○	0.667	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
memberi sebab-musabab	解释	○	○	○	○	○	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
memberi sebab-musabab	辩解	○	○	○	○	○	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
sah	合法	○	○	○	○	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
sah	法定	×	○	○	○	○	0.143	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
sah	有效	○	○	○	○	○	0.143	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
sah	跑	×	×	×	×	×	0.143	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
sah	跑步	×	×	×	×	×	0.133	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
sah	逃跑	×	×	×	×	×	0.067	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
sah	奔	×	×	×	×	×	0.063	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
pemidang tilik	取景器	○	×	○	—	○	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
mara	前进	○	○	○	○	○	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
mara	向前	○	○	○	○	○	0.182	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
mara	不幸	×	×	×	×	×	0.182	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
mara	打制	×	×	×	×	×	0.182	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
mara	进步	×	×	○	—	×	0.154	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×	×
mara	伪造	×	×	×	×	×	0.154	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×	×
mara	进展	×	×	○	—	×	0.083	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
mara	继续	○	×	○	○	○	0.077	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×

		Human decision						Decision by OTIC Filtering																																			
Malay	Chinese	Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard	OTIC score	$\alpha = 0, \beta = 0$	$\alpha = 0, \beta = 0.2$	$\alpha = 0, \beta = 0.4$	$\alpha = 0, \beta = 0.6$	$\alpha = 0, \beta = 0.8$	$\alpha = 0, \beta = 1.0$	$\alpha = 0.2, \beta = 0$	$\alpha = 0.2, \beta = 0.2$	$\alpha = 0.2, \beta = 0.4$	$\alpha = 0.2, \beta = 0.6$	$\alpha = 0.2, \beta = 0.8$	$\alpha = 0.2, \beta = 1.0$	$\alpha = 0.4, \beta = 0$	$\alpha = 0.4, \beta = 0.2$	$\alpha = 0.4, \beta = 0.4$	$\alpha = 0.4, \beta = 0.6$	$\alpha = 0.4, \beta = 0.8$	$\alpha = 0.4, \beta = 1.0$	$\alpha = 0.6, \beta = 0$	$\alpha = 0.6, \beta = 0.2$	$\alpha = 0.6, \beta = 0.4$	$\alpha = 0.6, \beta = 0.6$	$\alpha = 0.6, \beta = 0.8$	$\alpha = 0.6, \beta = 1.0$	$\alpha = 0.8, \beta = 0$	$\alpha = 0.8, \beta = 0.2$	$\alpha = 0.8, \beta = 0.4$	$\alpha = 0.8, \beta = 0.6$	$\alpha = 0.8, \beta = 0.8$	$\alpha = 0.8, \beta = 1.0$	$\alpha = 1.0, \beta = 0$	$\alpha = 1.0, \beta = 0.2$	$\alpha = 1.0, \beta = 0.4$	$\alpha = 1.0, \beta = 0.6$	$\alpha = 1.0, \beta = 0.8$	$\alpha = 1.0, \beta = 1.0$
mara	转寄	×	×	×	×	×	0.026	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
menuruti	跟随	○	○	○	—	○	0.667	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
menuruti	遵循	○	—	○	○	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×
menuruti	接着	×	×	×	×	×	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
menuruti	遵从	○	○	○	○	○	0.133	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
menuruti	服从	○	○	○	○	○	0.100	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
pengantar	引言	×	—	○	×	×	0.667	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
pengantar	导言	○	×	○	×	○	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
pengantar	入门	×	—	○	×	×	0.125	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
memulakan semula	重新开始	○	○	○	○	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
memulakan semula	恢复	○	×	○	○	○	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×
memulakan semula	更新	○	×	×	—	×	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×
makan air	褥疮	—	×	×	×	×	1.000	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
timbun	积聚	○	○	○	—	○	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
lari	飞	×	×	×	×	×	0.154	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
lari	飘扬	×	×	×	×	×	0.154	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
lari	飞翔	×	×	×	×	×	0.154	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
lari	逃之夭夭	○	○	○	○	○	0.154	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
menabung	储蓄	○	○	○	○	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
menabung	挽救	×	×	×	×	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
menabung	救	×	×	×	×	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
menabung	节省	×	×	○	×	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×

E.4 Human Judgements and OTIC Filtering Decisions on Iban–Malay Translation Pairings

Legend: ○ Accept × Reject — Unsure

		Human decision						Decision by OTIC Filtering																																			
Iban	Malay	Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard	OTIC score	$\alpha = 0, \beta = 0$	$\alpha = 0, \beta = 0.2$	$\alpha = 0, \beta = 0.4$	$\alpha = 0, \beta = 0.6$	$\alpha = 0, \beta = 0.8$	$\alpha = 0, \beta = 1.0$	$\alpha = 0.2, \beta = 0$	$\alpha = 0.2, \beta = 0.2$	$\alpha = 0.2, \beta = 0.4$	$\alpha = 0.2, \beta = 0.6$	$\alpha = 0.2, \beta = 0.8$	$\alpha = 0.2, \beta = 1.0$	$\alpha = 0.4, \beta = 0$	$\alpha = 0.4, \beta = 0.2$	$\alpha = 0.4, \beta = 0.4$	$\alpha = 0.4, \beta = 0.6$	$\alpha = 0.4, \beta = 0.8$	$\alpha = 0.4, \beta = 1.0$	$\alpha = 0.6, \beta = 0$	$\alpha = 0.6, \beta = 0.2$	$\alpha = 0.6, \beta = 0.4$	$\alpha = 0.6, \beta = 0.6$	$\alpha = 0.6, \beta = 0.8$	$\alpha = 0.6, \beta = 1.0$	$\alpha = 0.8, \beta = 0$	$\alpha = 0.8, \beta = 0.2$	$\alpha = 0.8, \beta = 0.4$	$\alpha = 0.8, \beta = 0.6$	$\alpha = 0.8, \beta = 0.8$	$\alpha = 0.8, \beta = 1.0$	$\alpha = 1.0, \beta = 0$	$\alpha = 1.0, \beta = 0.2$	$\alpha = 1.0, \beta = 0.4$	$\alpha = 1.0, \beta = 0.6$	$\alpha = 1.0, \beta = 0.8$	$\alpha = 1.0, \beta = 1.0$
bing	bank	×	○	×	○	○	1.000	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
bing	tebing	×	×	×	×	×	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
bing	permatang	×	×	×	×	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
ladang	tebing	×	×	×	×	×	0.444	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
ladang	daik	×	×	×	×	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
ladang	bank	×	×	×	×	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
ladang	permatang	×	×	×	×	×	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
bekening	bank	×	—	×	×	×	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
bekening	cerun	×	—	×	×	×	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×
bekening	tebing	×	—	×	×	×	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×
bekening	permatang	×	—	×	×	×	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
bekening	lereng	×	—	×	×	×	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
pekung	pertumbuhan	×	—	×	—	×	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
pekung	mengepung	×	—	×	—	×	0.364	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×
pekung	tumbuhan	×	—	×	—	×	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
pekung	mengelilingi	×	—	×	—	×	0.211	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
pekung	menyelubungi	○	—	×	—	○	0.083	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
pengawa	pendudukan	×	—	×	×	×	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
pengawa	pekerjaan	○	○	○	○	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
pengawa	upacara amal	○	×	×	○	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

Iban	Malay	Human decision					OTIC score	Decision by OTIC Filtering																																			
		Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard		$\alpha = 0, \beta = 0$	$\alpha = 0, \beta = 0.2$	$\alpha = 0, \beta = 0.4$	$\alpha = 0, \beta = 0.6$	$\alpha = 0, \beta = 0.8$	$\alpha = 0, \beta = 1.0$	$\alpha = 0.2, \beta = 0$	$\alpha = 0.2, \beta = 0.2$	$\alpha = 0.2, \beta = 0.4$	$\alpha = 0.2, \beta = 0.6$	$\alpha = 0.2, \beta = 0.8$	$\alpha = 0.2, \beta = 1.0$	$\alpha = 0.4, \beta = 0$	$\alpha = 0.4, \beta = 0.2$	$\alpha = 0.4, \beta = 0.4$	$\alpha = 0.4, \beta = 0.6$	$\alpha = 0.4, \beta = 0.8$	$\alpha = 0.4, \beta = 1.0$	$\alpha = 0.6, \beta = 0$	$\alpha = 0.6, \beta = 0.2$	$\alpha = 0.6, \beta = 0.4$	$\alpha = 0.6, \beta = 0.6$	$\alpha = 0.6, \beta = 0.8$	$\alpha = 0.6, \beta = 1.0$	$\alpha = 0.8, \beta = 0$	$\alpha = 0.8, \beta = 0.2$	$\alpha = 0.8, \beta = 0.4$	$\alpha = 0.8, \beta = 0.6$	$\alpha = 0.8, \beta = 0.8$	$\alpha = 0.8, \beta = 1.0$	$\alpha = 1.0, \beta = 0$	$\alpha = 1.0, \beta = 0.2$	$\alpha = 1.0, \beta = 0.4$	$\alpha = 1.0, \beta = 0.6$	$\alpha = 1.0, \beta = 0.8$	$\alpha = 1.0, \beta = 1.0$
pengawa	karya	×	×	×	—	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
pengawa	amal	×	×	×	×	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
pengawa	urusan	○	○	○	○	○	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×
pengawa	perniagaan	○	×	×	×	×	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×
pengawa	hal	○	○	×	○	○	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
pengawa	kilang	×	×	×	×	×	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
pengawa	kerja	○	○	×	○	○	0.182	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
kempat	menebang	×	×	×	—	×	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
kempat	menuai	×	×	×	×	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
kempat	memangkas	○	×	×	○	○	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×
kempat	mengerat	○	×	×	○	○	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×
kempat	menggunting	×	○	×	○	○	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×
kempat	menghiris	×	×	×	○	×	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
kempat	mencantas	○	×	○	○	○	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
kempat	memotong	○	×	×	○	○	0.143	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	×
kempat	mengurangkan	×	×	×	×	×	0.048	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	×
dada	dada	×	○	○	○	○	0.571	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
dada	tetek	○	×	×	×	×	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
dada	buah dada	○	×	×	×	×	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
dada	susu	○	×	×	×	×	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
dada	peti	○	×	×	×	×	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
lelegong	putih	○	—	×	—	○	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
lelegong	comel	○	—	×	—	○	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
lelegong	cantik	○	—	×	—	○	0.174	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
lelegong	patut	○	—	×	—	○	0.133	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	×

Iban	Malay	Human decision					OTIC score	Decision by OTIC Filtering																																			
		Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard		$\alpha = 0, \beta = 0$	$\alpha = 0, \beta = 0.2$	$\alpha = 0, \beta = 0.4$	$\alpha = 0, \beta = 0.6$	$\alpha = 0, \beta = 0.8$	$\alpha = 0, \beta = 1.0$	$\alpha = 0.2, \beta = 0$	$\alpha = 0.2, \beta = 0.2$	$\alpha = 0.2, \beta = 0.4$	$\alpha = 0.2, \beta = 0.6$	$\alpha = 0.2, \beta = 0.8$	$\alpha = 0.2, \beta = 1.0$	$\alpha = 0.4, \beta = 0$	$\alpha = 0.4, \beta = 0.2$	$\alpha = 0.4, \beta = 0.4$	$\alpha = 0.4, \beta = 0.6$	$\alpha = 0.4, \beta = 0.8$	$\alpha = 0.4, \beta = 1.0$	$\alpha = 0.6, \beta = 0$	$\alpha = 0.6, \beta = 0.2$	$\alpha = 0.6, \beta = 0.4$	$\alpha = 0.6, \beta = 0.6$	$\alpha = 0.6, \beta = 0.8$	$\alpha = 0.6, \beta = 1.0$	$\alpha = 0.8, \beta = 0$	$\alpha = 0.8, \beta = 0.2$	$\alpha = 0.8, \beta = 0.4$	$\alpha = 0.8, \beta = 0.6$	$\alpha = 0.8, \beta = 0.8$	$\alpha = 0.8, \beta = 1.0$	$\alpha = 1.0, \beta = 0$	$\alpha = 1.0, \beta = 0.2$	$\alpha = 1.0, \beta = 0.4$	$\alpha = 1.0, \beta = 0.6$	$\alpha = 1.0, \beta = 0.8$	$\alpha = 1.0, \beta = 1.0$
lelegong	adil	○	—	×	—	○	0.118	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
lelegong	baik	○	—	×	—	○	0.065	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
bechakap	mencabar	○	—	×	—	○	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
bechakap	bercakap besar	○	—	×	—	○	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
dawat	dakwat	○	○	○	○	○	0.667	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
dawat	tinta	×	○	×	—	×	0.667	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
kulup	kulit khatan	○	×	×	—	×	1.000	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
kulup	kulup	○	○	○	—	○	0.667	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
wai	handai	○	×	—	×	×	0.667	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
wai	taulan	○	×	×	×	×	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
wai	sahabat	○	○	×	—	○	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
wai	teman	○	○	○	○	○	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
wai	kawan	○	○	○	○	○	0.143	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
wai	rakan	○	○	○	○	○	0.118	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
ringgat	kumpulan	×	—	×	—	×	0.063	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
penggang	parau	○	×	×	—	×	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
penggang	serak	○	×	×	—	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
buntis	memilih	○	—	×	—	○	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
layar	melayarkan	○	—	×	—	○	1.000	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
layar	layar	○	○	○	○	○	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
layar	belayar	○	×	×	—	×	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
rigai	nipis	○	—	×	—	○	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
rigai	tipis	○	—	×	—	○	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
rigai	kurus	○	—	×	—	○	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
rigai	kurus kering	○	—	○	—	○	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×

		Human decision						Decision by OTIC Filtering																																			
Iban	Malay	Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard	OTIC score	$\alpha = 0, \beta = 0$	$\alpha = 0, \beta = 0.2$	$\alpha = 0, \beta = 0.4$	$\alpha = 0, \beta = 0.6$	$\alpha = 0, \beta = 0.8$	$\alpha = 0, \beta = 1.0$	$\alpha = 0.2, \beta = 0$	$\alpha = 0.2, \beta = 0.2$	$\alpha = 0.2, \beta = 0.4$	$\alpha = 0.2, \beta = 0.6$	$\alpha = 0.2, \beta = 0.8$	$\alpha = 0.2, \beta = 1.0$	$\alpha = 0.4, \beta = 0$	$\alpha = 0.4, \beta = 0.2$	$\alpha = 0.4, \beta = 0.4$	$\alpha = 0.4, \beta = 0.6$	$\alpha = 0.4, \beta = 0.8$	$\alpha = 0.4, \beta = 1.0$	$\alpha = 0.6, \beta = 0$	$\alpha = 0.6, \beta = 0.2$	$\alpha = 0.6, \beta = 0.4$	$\alpha = 0.6, \beta = 0.6$	$\alpha = 0.6, \beta = 0.8$	$\alpha = 0.6, \beta = 1.0$	$\alpha = 0.8, \beta = 0$	$\alpha = 0.8, \beta = 0.2$	$\alpha = 0.8, \beta = 0.4$	$\alpha = 0.8, \beta = 0.6$	$\alpha = 0.8, \beta = 0.8$	$\alpha = 0.8, \beta = 1.0$	$\alpha = 1.0, \beta = 0$	$\alpha = 1.0, \beta = 0.2$	$\alpha = 1.0, \beta = 0.4$	$\alpha = 1.0, \beta = 0.6$	$\alpha = 1.0, \beta = 0.8$	$\alpha = 1.0, \beta = 1.0$
gemu	tambun	○	—	×	×	×	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
gemu	subur	○	×	○	×	○	0.235	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
gemu	tebal	○	×	×	×	×	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	○	○	○	○	×	×
gemu	gemuk	○	○	○	○	○	0.154	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
gemu	menguntungkan	○	×	×	×	×	0.154	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
tipah	kepak	○	—	×	—	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
tipah	sayap	○	—	×	—	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
ngenyadi	kikir	×	×	×	×	×	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
ngenyadi	lokek	○	×	×	×	×	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
ngenyadi	bakhil	×	×	×	×	×	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
ngenyadi	kedekut	○	×	×	×	×	0.125	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
asi	beras	×	×	×	○	×	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
asi	nasi	○	○	○	○	○	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
asi	kegemaran	×	×	×	×	×	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
asi	kanan	×	×	×	×	×	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
asi	patut	×	×	×	×	×	0.211	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
asi	betul	×	×	×	×	×	0.205	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
asi	putih	×	×	×	×	×	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
asi	padi	×	×	×	×	×	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
asi	adil	×	○	○	×	○	0.190	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
asi	mematuhi	○	—	×	○	○	0.154	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×
asi	mentaati	○	○	○	○	○	0.154	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×
asi	baik	—	×	×	×	×	0.129	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
asi	sesuai	×	×	×	×	×	0.118	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
asi	wajar	○	×	×	×	×	0.105	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×

		Human decision						Decision by OTIC Filtering																																			
Iban	Malay	Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard	OTIC score	$\alpha = 0, \beta = 0$	$\alpha = 0, \beta = 0.2$	$\alpha = 0, \beta = 0.4$	$\alpha = 0, \beta = 0.6$	$\alpha = 0, \beta = 0.8$	$\alpha = 0, \beta = 1.0$	$\alpha = 0.2, \beta = 0$	$\alpha = 0.2, \beta = 0.2$	$\alpha = 0.2, \beta = 0.4$	$\alpha = 0.2, \beta = 0.6$	$\alpha = 0.2, \beta = 0.8$	$\alpha = 0.2, \beta = 1.0$	$\alpha = 0.4, \beta = 0$	$\alpha = 0.4, \beta = 0.2$	$\alpha = 0.4, \beta = 0.4$	$\alpha = 0.4, \beta = 0.6$	$\alpha = 0.4, \beta = 0.8$	$\alpha = 0.4, \beta = 1.0$	$\alpha = 0.6, \beta = 0$	$\alpha = 0.6, \beta = 0.2$	$\alpha = 0.6, \beta = 0.4$	$\alpha = 0.6, \beta = 0.6$	$\alpha = 0.6, \beta = 0.8$	$\alpha = 0.6, \beta = 1.0$	$\alpha = 0.8, \beta = 0$	$\alpha = 0.8, \beta = 0.2$	$\alpha = 0.8, \beta = 0.4$	$\alpha = 0.8, \beta = 0.6$	$\alpha = 0.8, \beta = 0.8$	$\alpha = 0.8, \beta = 1.0$	$\alpha = 1.0, \beta = 0$	$\alpha = 1.0, \beta = 0.2$	$\alpha = 1.0, \beta = 0.4$	$\alpha = 1.0, \beta = 0.6$	$\alpha = 1.0, \beta = 0.8$	$\alpha = 1.0, \beta = 1.0$
asi	sepatutnya	○	×	○	×	○	0.100	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
asi	tepat	○	×	×	×	×	0.078	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
asi	cantik	×	×	×	×	×	0.074	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
asi	sihat	×	×	×	×	×	0.029	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
rarik	memukul cantas	○	—	×	—	○	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
rarik	menghiris	○	—	×	—	○	0.444	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
rarik	terbelah	○	—	×	—	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
rarik	membelah	○	—	×	—	○	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
rarik	memotong	○	—	×	—	○	0.057	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
tambit	mengancing	×	○	○	×	○	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
tambit	mengunci	○	○	○	○	○	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
tambit	menyelak	×	×	×	×	×	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
tambit	melekatkan	×	×	×	×	×	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	○	○	○	○	×	×
tambit	mengukuhkan	×	○	×	×	×	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
tambit	mengikat	×	○	×	×	×	0.087	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
empiar	selumbar	○	—	×	—	○	0.667	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
empiar	serpihan	○	—	○	—	○	0.154	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
runggu	gegaran	×	×	×	—	×	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
kih	nekad	○	—	×	○	○	0.182	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
kih	keras hati	○	○	○	○	○	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×
kih	degil	○	×	○	○	○	0.154	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
kih	keras kepala	○	○	○	○	○	0.125	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
gerawan	syaitan	×	—	×	—	×	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
gerawan	hantu	×	—	×	—	×	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

		Human decision						Decision by OTIC Filtering																																			
Iban	Malay	Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard	OTIC score	$\alpha=0,\beta=0$	$\alpha=0,\beta=0.2$	$\alpha=0,\beta=0.4$	$\alpha=0,\beta=0.6$	$\alpha=0,\beta=0.8$	$\alpha=0,\beta=1.0$	$\alpha=0.2,\beta=0$	$\alpha=0.2,\beta=0.2$	$\alpha=0.2,\beta=0.4$	$\alpha=0.2,\beta=0.6$	$\alpha=0.2,\beta=0.8$	$\alpha=0.2,\beta=1.0$	$\alpha=0.4,\beta=0$	$\alpha=0.4,\beta=0.2$	$\alpha=0.4,\beta=0.4$	$\alpha=0.4,\beta=0.6$	$\alpha=0.4,\beta=0.8$	$\alpha=0.4,\beta=1.0$	$\alpha=0.6,\beta=0$	$\alpha=0.6,\beta=0.2$	$\alpha=0.6,\beta=0.4$	$\alpha=0.6,\beta=0.6$	$\alpha=0.6,\beta=0.8$	$\alpha=0.6,\beta=1.0$	$\alpha=0.8,\beta=0$	$\alpha=0.8,\beta=0.2$	$\alpha=0.8,\beta=0.4$	$\alpha=0.8,\beta=0.6$	$\alpha=0.8,\beta=0.8$	$\alpha=0.8,\beta=1.0$	$\alpha=1.0,\beta=0$	$\alpha=1.0,\beta=0.2$	$\alpha=1.0,\beta=0.4$	$\alpha=1.0,\beta=0.6$	$\alpha=1.0,\beta=0.8$	$\alpha=1.0,\beta=1.0$
babal	empangan	×	—	×	—	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
babal	Empangan	×	—	×	—	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
babal	pelupa	○	—	×	—	○	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
babal	tambak	×	—	×	—	×	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
babal	menyekat	×	—	×	—	×	0.120	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
pusil	memetik	×	○	×	—	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
pusil	mencubit	×	○	○	—	○	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
pusil	memulas	○	○	×	—	○	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
pusil	mencabut	×	×	○	—	×	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
pusil	memilih	×	×	×	—	×	0.143	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
bungkur	membungkus	○	×	○	—	○	0.231	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
bungkur	membalut	○	×	×	×	×	0.231	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
bungkur	meliputi	○	○	×	○	○	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
bungkur	menutup	○	○	○	○	○	0.118	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
igur	diduduki	○	—	×	—	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
igur	diganggu oleh berbagai-bagai masalah	○	—	—	—	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
igur	sibuk	○	—	×	—	○	0.308	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
igur	leka	○	—	×	—	○	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
igur	asyik	○	—	×	—	○	0.267	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
kebut	berganjak	×	×	○	—	×	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
kebut	mengusulkan	×	×	×	—	×	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
kebut	bergerak	○	○	○	○	○	0.364	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
kebut	bertindak	○	○	×	—	○	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×

		Human decision						Decision by OTIC Filtering																																			
Iban	Malay	Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard	OTIC score	$\alpha=0,\beta=0$	$\alpha=0,\beta=0.2$	$\alpha=0,\beta=0.4$	$\alpha=0,\beta=0.6$	$\alpha=0,\beta=0.8$	$\alpha=0,\beta=1.0$	$\alpha=0.2,\beta=0$	$\alpha=0.2,\beta=0.2$	$\alpha=0.2,\beta=0.4$	$\alpha=0.2,\beta=0.6$	$\alpha=0.2,\beta=0.8$	$\alpha=0.2,\beta=1.0$	$\alpha=0.4,\beta=0$	$\alpha=0.4,\beta=0.2$	$\alpha=0.4,\beta=0.4$	$\alpha=0.4,\beta=0.6$	$\alpha=0.4,\beta=0.8$	$\alpha=0.4,\beta=1.0$	$\alpha=0.6,\beta=0$	$\alpha=0.6,\beta=0.2$	$\alpha=0.6,\beta=0.4$	$\alpha=0.6,\beta=0.6$	$\alpha=0.6,\beta=0.8$	$\alpha=0.6,\beta=1.0$	$\alpha=0.8,\beta=0$	$\alpha=0.8,\beta=0.2$	$\alpha=0.8,\beta=0.4$	$\alpha=0.8,\beta=0.6$	$\alpha=0.8,\beta=0.8$	$\alpha=0.8,\beta=1.0$	$\alpha=1.0,\beta=0$	$\alpha=1.0,\beta=0.2$	$\alpha=1.0,\beta=0.4$	$\alpha=1.0,\beta=0.6$	$\alpha=1.0,\beta=0.8$	$\alpha=1.0,\beta=1.0$
kebut	bergoyang	○	○	×	○	○	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
kebut	berpindah	×	×	○	—	×	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
kebut	memindahkan	×	×	×	—	×	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×
kebut	mengacau	×	×	×	—	×	0.077	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×
kebut	pergi	×	×	×	—	×	0.056	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×
keruin	gam	×	—	×	—	×	0.667	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
keruin	resin	○	—	×	—	○	0.667	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
keruin	gusi	×	—	×	—	×	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
keruin	damar	×	—	×	—	×	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
keruin	perekat	×	—	×	—	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
abis	segala	×	×	×	×	×	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
abis	menghabiskan	×	○	×	○	○	0.267	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×
abis	menyelesaikan	○	○	×	○	○	0.235	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
abis	melengkapkan	○	×	×	○	○	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×	×
abis	menamatkan	○	○	×	○	○	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×	×
abis	menyiapkan	○	○	×	—	○	0.182	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×	×
abis	semua	×	×	×	×	×	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
abis	habis	○	○	○	○	○	0.152	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
abis	berakhir	○	○	○	○	○	0.128	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
kanchin	butang	○	○	○	○	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
chekak	mencekik	○	○	○	○	○	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
tambak	keratan	×	○	○	—	○	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
tambak	menanam	○	○	○	—	○	0.133	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
tambak	potongan	×	—	○	○	○	0.125	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
silam	mendung	×	—	○	—	○	0.444	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

		Human decision						Decision by OTIC Filtering																																			
Iban	Malay	Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard	OTIC score	$\alpha = 0, \beta = 0$	$\alpha = 0, \beta = 0.2$	$\alpha = 0, \beta = 0.4$	$\alpha = 0, \beta = 0.6$	$\alpha = 0, \beta = 0.8$	$\alpha = 0, \beta = 1.0$	$\alpha = 0.2, \beta = 0$	$\alpha = 0.2, \beta = 0.2$	$\alpha = 0.2, \beta = 0.4$	$\alpha = 0.2, \beta = 0.6$	$\alpha = 0.2, \beta = 0.8$	$\alpha = 0.2, \beta = 1.0$	$\alpha = 0.4, \beta = 0$	$\alpha = 0.4, \beta = 0.2$	$\alpha = 0.4, \beta = 0.4$	$\alpha = 0.4, \beta = 0.6$	$\alpha = 0.4, \beta = 0.8$	$\alpha = 0.4, \beta = 1.0$	$\alpha = 0.6, \beta = 0$	$\alpha = 0.6, \beta = 0.2$	$\alpha = 0.6, \beta = 0.4$	$\alpha = 0.6, \beta = 0.6$	$\alpha = 0.6, \beta = 0.8$	$\alpha = 0.6, \beta = 1.0$	$\alpha = 0.8, \beta = 0$	$\alpha = 0.8, \beta = 0.2$	$\alpha = 0.8, \beta = 0.4$	$\alpha = 0.8, \beta = 0.6$	$\alpha = 0.8, \beta = 0.8$	$\alpha = 0.8, \beta = 1.0$	$\alpha = 1.0, \beta = 0$	$\alpha = 1.0, \beta = 0.2$	$\alpha = 1.0, \beta = 0.4$	$\alpha = 1.0, \beta = 0.6$	$\alpha = 1.0, \beta = 0.8$	$\alpha = 1.0, \beta = 1.0$
silam	kelam	×	—	×	—	×	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
silam	hitam	×	—	×	—	×	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
silam	tersembunyi	×	—	×	—	×	0.182	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
silam	gelap	×	—	○	—	○	0.182	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
silam	muram	×	—	○	—	○	0.173	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	×
silam	tua	×	—	×	—	×	0.154	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	×
silam	suram	×	—	×	—	×	0.133	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	×
sepi	merasa	○	○	○	○	○	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
sepi	berasa	×	×	×	×	×	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
sepi	terasa	×	×	×	×	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
sepi	mengalami	×	—	×	×	×	0.182	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	×
sepi	mencuba	○	×	○	○	○	0.182	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	×
rapan	kemalapan	×	—	—	—	×	1.000	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
rapan	kesamaran	×	—	—	—	×	0.200	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	×
rapan	kelemahan	×	—	—	—	×	0.111	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	×
nguman	menanah	×	—	×	—	×	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
nguman	membarah	×	—	×	—	×	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
nguman	mengutip	×	—	×	—	×	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
nguman	memungut	×	—	×	—	×	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
nguman	berkumpul	×	—	×	—	×	0.133	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	×
pengaroh	jampi	×	×	○	×	×	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
pengaroh	tangkal	○	○	×	○	○	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
pengaroh	sihir	×	×	○	—	×	0.182	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
liar	pemalu	×	—	×	×	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
liar	liar	×	—	○	○	○	0.308	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×

		Human decision						Decision by OTIC Filtering																																			
Iban	Malay	Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard	OTIC score	$\alpha = 0, \beta = 0$	$\alpha = 0, \beta = 0.2$	$\alpha = 0, \beta = 0.4$	$\alpha = 0, \beta = 0.6$	$\alpha = 0, \beta = 0.8$	$\alpha = 0, \beta = 1.0$	$\alpha = 0.2, \beta = 0$	$\alpha = 0.2, \beta = 0.2$	$\alpha = 0.2, \beta = 0.4$	$\alpha = 0.2, \beta = 0.6$	$\alpha = 0.2, \beta = 0.8$	$\alpha = 0.2, \beta = 1.0$	$\alpha = 0.4, \beta = 0$	$\alpha = 0.4, \beta = 0.2$	$\alpha = 0.4, \beta = 0.4$	$\alpha = 0.4, \beta = 0.6$	$\alpha = 0.4, \beta = 0.8$	$\alpha = 0.4, \beta = 1.0$	$\alpha = 0.6, \beta = 0$	$\alpha = 0.6, \beta = 0.2$	$\alpha = 0.6, \beta = 0.4$	$\alpha = 0.6, \beta = 0.6$	$\alpha = 0.6, \beta = 0.8$	$\alpha = 0.6, \beta = 1.0$	$\alpha = 0.8, \beta = 0$	$\alpha = 0.8, \beta = 0.2$	$\alpha = 0.8, \beta = 0.4$	$\alpha = 0.8, \beta = 0.6$	$\alpha = 0.8, \beta = 0.8$	$\alpha = 0.8, \beta = 1.0$	$\alpha = 1.0, \beta = 0$	$\alpha = 1.0, \beta = 0.2$	$\alpha = 1.0, \beta = 0.4$	$\alpha = 1.0, \beta = 0.6$	$\alpha = 1.0, \beta = 0.8$	$\alpha = 1.0, \beta = 1.0$
liar	malu	×	—	×	×	×	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
liar	bergelora	×	—	×	×	×	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
liar	terbeliak	×	—	×	×	×	0.143	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
liar	ganas	×	—	○	×	×	0.118	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
pupus	siap	×	—	○	×	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
pupus	seluruh	×	—	×	×	×	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
pupus	lengkap	○	—	×	×	×	0.176	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
pupus	semuanya	○	—	×	×	×	0.091	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
pupus	sempurna	○	—	×	×	×	0.087	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
kandung	rahim	×	○	×	○	○	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
kandung	memikirkan	×	×	×	×	×	0.095	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
linga	cuai	○	×	○	—	○	0.154	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
linga	bodoh	○	×	×	—	×	0.039	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
anchur	memencarkan	×	×	×	—	×	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
anchur	roboh	×	×	○	—	×	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
anchur	menjadi lembut	○	×	×	○	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
anchur	mencairkan	○	○	○	○	○	0.364	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×
anchur	melarutkan	○	○	×	○	○	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
anchur	menyuraikan	○	×	×	—	×	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
anchur	runtuh	×	×	×	—	×	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×
anchur	melenyapkan	○	×	×	—	×	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×	×
anchur	berselerak	×	○	○	×	○	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×	×
anchur	hilang	×	×	×	×	×	0.214	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×	×
anchur	meruntuhkan	×	×	×	×	×	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×	×

Iban	Malay	Human decision					OTIC score	Decision by OTIC Filtering																																			
		Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard		$\alpha = 0, \beta = 0$	$\alpha = 0, \beta = 0.2$	$\alpha = 0, \beta = 0.4$	$\alpha = 0, \beta = 0.6$	$\alpha = 0, \beta = 0.8$	$\alpha = 0, \beta = 1.0$	$\alpha = 0.2, \beta = 0$	$\alpha = 0.2, \beta = 0.2$	$\alpha = 0.2, \beta = 0.4$	$\alpha = 0.2, \beta = 0.6$	$\alpha = 0.2, \beta = 0.8$	$\alpha = 0.2, \beta = 1.0$	$\alpha = 0.4, \beta = 0$	$\alpha = 0.4, \beta = 0.2$	$\alpha = 0.4, \beta = 0.4$	$\alpha = 0.4, \beta = 0.6$	$\alpha = 0.4, \beta = 0.8$	$\alpha = 0.4, \beta = 1.0$	$\alpha = 0.6, \beta = 0$	$\alpha = 0.6, \beta = 0.2$	$\alpha = 0.6, \beta = 0.4$	$\alpha = 0.6, \beta = 0.6$	$\alpha = 0.6, \beta = 0.8$	$\alpha = 0.6, \beta = 1.0$	$\alpha = 0.8, \beta = 0$	$\alpha = 0.8, \beta = 0.2$	$\alpha = 0.8, \beta = 0.4$	$\alpha = 0.8, \beta = 0.6$	$\alpha = 0.8, \beta = 0.8$	$\alpha = 0.8, \beta = 1.0$	$\alpha = 1.0, \beta = 0$	$\alpha = 1.0, \beta = 0.2$	$\alpha = 1.0, \beta = 0.4$	$\alpha = 1.0, \beta = 0.6$	$\alpha = 1.0, \beta = 0.8$	$\alpha = 1.0, \beta = 1.0$
anchur	menyebarkan	×	×	×	×	×	0.133	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
japai	menyita	×	×	×	×	×	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
japai	menyambar	○	×	—	—	○	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
japai	menawan	×	×	×	×	×	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
japai	merampas	×	×	○	×	×	0.111	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
japai	menangkap	○	×	×	—	×	0.105	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
surik	mark	○	—	—	○	○	0.667	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
surik	markah	○	—	×	×	×	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
surik	tikas	○	—	×	×	×	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
surik	sasaran	○	—	×	×	×	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
surik	parut	○	×	×	×	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
surik	kesan	○	×	×	—	×	0.133	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
surik	tanda	○	×	○	—	○	0.095	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
mantuk	mematuk	○	○	○	○	○	1.000	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
merejuk	melompat	○	×	○	○	○	0.444	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
merejuk	meloncat	○	○	○	○	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×
merejuk	terkejut	×	×	×	×	×	0.057	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
merejuk	terperanjat	×	×	×	×	×	0.040	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
tajam	runcing	×	○	×	○	○	0.444	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
tajam	mudah teruja	×	×	×	×	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
tajam	teruja	×	×	×	×	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
tajam	tajam	○	○	○	○	○	0.183	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
tajam	nyaring	×	×	×	×	×	0.125	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
tajam	gigih	×	×	×	×	×	0.111	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
tajam	pintar	×	○	○	—	○	0.043	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×

		Human decision						Decision by OTIC Filtering																																			
Iban	Malay	Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard	OTIC score	$\alpha = 0, \beta = 0$	$\alpha = 0, \beta = 0.2$	$\alpha = 0, \beta = 0.4$	$\alpha = 0, \beta = 0.6$	$\alpha = 0, \beta = 0.8$	$\alpha = 0, \beta = 1.0$	$\alpha = 0.2, \beta = 0$	$\alpha = 0.2, \beta = 0.2$	$\alpha = 0.2, \beta = 0.4$	$\alpha = 0.2, \beta = 0.6$	$\alpha = 0.2, \beta = 0.8$	$\alpha = 0.2, \beta = 1.0$	$\alpha = 0.4, \beta = 0$	$\alpha = 0.4, \beta = 0.2$	$\alpha = 0.4, \beta = 0.4$	$\alpha = 0.4, \beta = 0.6$	$\alpha = 0.4, \beta = 0.8$	$\alpha = 0.4, \beta = 1.0$	$\alpha = 0.6, \beta = 0$	$\alpha = 0.6, \beta = 0.2$	$\alpha = 0.6, \beta = 0.4$	$\alpha = 0.6, \beta = 0.6$	$\alpha = 0.6, \beta = 0.8$	$\alpha = 0.6, \beta = 1.0$	$\alpha = 0.8, \beta = 0$	$\alpha = 0.8, \beta = 0.2$	$\alpha = 0.8, \beta = 0.4$	$\alpha = 0.8, \beta = 0.6$	$\alpha = 0.8, \beta = 0.8$	$\alpha = 0.8, \beta = 1.0$	$\alpha = 1.0, \beta = 0$	$\alpha = 1.0, \beta = 0.2$	$\alpha = 1.0, \beta = 0.4$	$\alpha = 1.0, \beta = 0.6$	$\alpha = 1.0, \beta = 0.8$	$\alpha = 1.0, \beta = 1.0$
seliah	menghindarkan	×	○	×	○	○	0.444	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
seliah	mengelakkan	○	○	○	○	○	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
sengak	lelah	○	○	○	○	○	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
nguyum	menanah	×	○	×	—	×	0.667	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
nguyum	membarah	×	—	×	—	×	0.667	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
gagai	mengejar	○	○	○	○	○	0.407	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
gagai	memburu	○	×	×	—	×	0.407	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
gagai	menghalau	×	×	○	—	×	0.154	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
mugar	menggosok	○	○	×	—	○	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
mugar	membersihkan	○	○	○	—	○	0.154	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
mugar	menggilap	○	×	○	—	○	0.111	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
ruang	palka	×	—	×	—	×	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
ruang	pegangan	×	—	×	×	×	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
ruang	pengaruh	×	—	×	×	×	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
pinggir	sempadan	×	—	×	—	×	0.111	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
pinggir	pinggir	×	—	○	—	○	0.105	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×
pinggir	tepi	×	—	○	—	○	0.100	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×
panggung	pail	×	—	×	×	×	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
panggung	longgokan	×	—	○	×	×	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
panggung	cerucuk	×	—	×	×	×	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
panggung	timbunan	×	—	×	×	×	0.364	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×
panggung	serombong	×	—	×	×	×	0.125	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
rintai	baris	○	○	×	—	○	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
rintai	talian	○	—	×	×	×	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
rintai	garisan	○	×	○	×	○	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×

		Human decision						Decision by OTIC Filtering																																			
Iban	Malay	Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard	OTIC score	$\alpha = 0, \beta = 0$	$\alpha = 0, \beta = 0.2$	$\alpha = 0, \beta = 0.4$	$\alpha = 0, \beta = 0.6$	$\alpha = 0, \beta = 0.8$	$\alpha = 0, \beta = 1.0$	$\alpha = 0.2, \beta = 0$	$\alpha = 0.2, \beta = 0.2$	$\alpha = 0.2, \beta = 0.4$	$\alpha = 0.2, \beta = 0.6$	$\alpha = 0.2, \beta = 0.8$	$\alpha = 0.2, \beta = 1.0$	$\alpha = 0.4, \beta = 0$	$\alpha = 0.4, \beta = 0.2$	$\alpha = 0.4, \beta = 0.4$	$\alpha = 0.4, \beta = 0.6$	$\alpha = 0.4, \beta = 0.8$	$\alpha = 0.4, \beta = 1.0$	$\alpha = 0.6, \beta = 0$	$\alpha = 0.6, \beta = 0.2$	$\alpha = 0.6, \beta = 0.4$	$\alpha = 0.6, \beta = 0.6$	$\alpha = 0.6, \beta = 0.8$	$\alpha = 0.6, \beta = 1.0$	$\alpha = 0.8, \beta = 0$	$\alpha = 0.8, \beta = 0.2$	$\alpha = 0.8, \beta = 0.4$	$\alpha = 0.8, \beta = 0.6$	$\alpha = 0.8, \beta = 0.8$	$\alpha = 0.8, \beta = 1.0$	$\alpha = 1.0, \beta = 0$	$\alpha = 1.0, \beta = 0.2$	$\alpha = 1.0, \beta = 0.4$	$\alpha = 1.0, \beta = 0.6$	$\alpha = 1.0, \beta = 0.8$	$\alpha = 1.0, \beta = 1.0$
rintai	garis	○	×	×	—	×	0.296	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
rintai	barisan	○	○	○	—	○	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
rintai	sempadan	○	×	×	×	×	0.037	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
simpan	menepati	×	×	×	×	×	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
simpan	tahan	×	×	○	×	×	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
simpan	menjaga	○	×	×	×	×	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
simpan	memelihara	○	×	×	×	×	0.154	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
simpan	menyimpan	○	○	○	○	○	0.118	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
simpan	menanggung	○	×	×	—	×	0.118	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
simpan	mengurung	○	○	○	×	○	0.118	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
buntut	akhir	○	×	×	×	×	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
buntut	tamat	○	×	×	×	×	0.238	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
buntut	matlamat	○	×	×	×	×	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
buntut	tujuan	○	×	×	×	×	0.118	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
sengah	lelah	○	×	—	—	○	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
sengah	parau	×	×	—	×	×	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
sengah	serak	×	×	—	×	×	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×
berumpu	menghimpunkan	○	×	—	○	○	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
berumpu	berhimpun	○	×	—	—	○	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
berumpu	mengumpulkan	○	×	—	○	○	0.143	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
berumpu	berkumpul	○	○	—	—	○	0.133	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
lagu	melodi	○	×	○	○	○	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
lagu	lagu	○	○	○	○	○	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
lagu	nyanyian	○	○	○	○	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
pusil	memetik	○	○	○	—	○	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

		Human decision						Decision by OTIC Filtering																																			
Iban	Malay	Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard	OTIC score	$\alpha = 0, \beta = 0$	$\alpha = 0, \beta = 0.2$	$\alpha = 0, \beta = 0.4$	$\alpha = 0, \beta = 0.6$	$\alpha = 0, \beta = 0.8$	$\alpha = 0, \beta = 1.0$	$\alpha = 0.2, \beta = 0$	$\alpha = 0.2, \beta = 0.2$	$\alpha = 0.2, \beta = 0.4$	$\alpha = 0.2, \beta = 0.6$	$\alpha = 0.2, \beta = 0.8$	$\alpha = 0.2, \beta = 1.0$	$\alpha = 0.4, \beta = 0$	$\alpha = 0.4, \beta = 0.2$	$\alpha = 0.4, \beta = 0.4$	$\alpha = 0.4, \beta = 0.6$	$\alpha = 0.4, \beta = 0.8$	$\alpha = 0.4, \beta = 1.0$	$\alpha = 0.6, \beta = 0$	$\alpha = 0.6, \beta = 0.2$	$\alpha = 0.6, \beta = 0.4$	$\alpha = 0.6, \beta = 0.6$	$\alpha = 0.6, \beta = 0.8$	$\alpha = 0.6, \beta = 1.0$	$\alpha = 0.8, \beta = 0$	$\alpha = 0.8, \beta = 0.2$	$\alpha = 0.8, \beta = 0.4$	$\alpha = 0.8, \beta = 0.6$	$\alpha = 0.8, \beta = 0.8$	$\alpha = 0.8, \beta = 1.0$	$\alpha = 1.0, \beta = 0$	$\alpha = 1.0, \beta = 0.2$	$\alpha = 1.0, \beta = 0.4$	$\alpha = 1.0, \beta = 0.6$	$\alpha = 1.0, \beta = 0.8$	$\alpha = 1.0, \beta = 1.0$
pusil	mencubit	×	×	×	—	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
pusil	memulas	○	○	×	—	○	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
pusil	mencabut	×	×	×	—	×	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
pusil	memilih	×	×	×	—	×	0.143	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×
pabu	tohor	×	—	×	—	×	1.000	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
pabu	dangkal	×	—	×	—	×	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
pabu	cetek	○	○	○	—	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×
lurus	lurus	○	○	○	○	○	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
lurus	langsung	×	×	×	—	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
lurus	terus	×	×	×	—	×	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
lurus	jujur	○	×	○	○	○	0.118	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
lurus	tepat	×	×	×	—	×	0.067	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
jarau	memangkas	○	○	○	○	○	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
jarau	mencantas	○	×	○	○	○	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
jarau	mengurangkan	○	×	×	—	×	0.100	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
ngeregau	menyelongkar	○	×	×	—	×	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
ngeregau	menggeledah	○	×	×	—	×	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
ngeregau	mengganggu	○	○	○	—	○	0.069	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
chemegah	bangga	○	—	—	—	○	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
chemegah	megah	○	—	—	—	○	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
chemegah	angkuh	○	—	—	—	○	0.119	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
chemegah	sombong	○	—	—	—	○	0.067	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
chemegah	bongkak	○	—	—	—	○	0.026	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
malu	memalukan	○	×	×	—	×	0.211	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
malu	malu	○	○	○	○	○	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×

		Human decision						Decision by OTIC Filtering																																			
Iban	Malay	Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard	OTIC score	$\alpha = 0, \beta = 0$	$\alpha = 0, \beta = 0.2$	$\alpha = 0, \beta = 0.4$	$\alpha = 0, \beta = 0.6$	$\alpha = 0, \beta = 0.8$	$\alpha = 0, \beta = 1.0$	$\alpha = 0.2, \beta = 0$	$\alpha = 0.2, \beta = 0.2$	$\alpha = 0.2, \beta = 0.4$	$\alpha = 0.2, \beta = 0.6$	$\alpha = 0.2, \beta = 0.8$	$\alpha = 0.2, \beta = 1.0$	$\alpha = 0.4, \beta = 0$	$\alpha = 0.4, \beta = 0.2$	$\alpha = 0.4, \beta = 0.4$	$\alpha = 0.4, \beta = 0.6$	$\alpha = 0.4, \beta = 0.8$	$\alpha = 0.4, \beta = 1.0$	$\alpha = 0.6, \beta = 0$	$\alpha = 0.6, \beta = 0.2$	$\alpha = 0.6, \beta = 0.4$	$\alpha = 0.6, \beta = 0.6$	$\alpha = 0.6, \beta = 0.8$	$\alpha = 0.6, \beta = 1.0$	$\alpha = 0.8, \beta = 0$	$\alpha = 0.8, \beta = 0.2$	$\alpha = 0.8, \beta = 0.4$	$\alpha = 0.8, \beta = 0.6$	$\alpha = 0.8, \beta = 0.8$	$\alpha = 0.8, \beta = 1.0$	$\alpha = 1.0, \beta = 0$	$\alpha = 1.0, \beta = 0.2$	$\alpha = 1.0, \beta = 0.4$	$\alpha = 1.0, \beta = 0.6$	$\alpha = 1.0, \beta = 0.8$	$\alpha = 1.0, \beta = 1.0$
malu	memalu	○	×	×	—	×	0.182	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
malu	membalun	×	×	×	○	×	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
malu	merempuh	×	×	×	×	×	0.154	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
malu	menggodam	×	×	×	×	×	0.154	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
malu	menewaskan	×	×	×	×	×	0.111	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
malu	menumbuk	×	×	×	—	×	0.111	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
malu	memukul	×	○	○	○	○	0.103	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
malu	mengalahkan	×	×	×	×	×	0.067	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
malu	sederhana	×	×	×	×	×	0.065	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
sepur	repui	○	—	×	—	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
sepur	habuk	○	○	○	—	○	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
sepur	debu	○	×	○	—	○	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
sepur	empuk	○	×	×	—	×	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×
sepur	lembik	×	×	×	—	×	0.182	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
sepur	lembut hati	×	×	×	—	×	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
sepur	ringan	×	×	×	—	×	0.154	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
sepur	lembut	×	×	×	—	×	0.100	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
nyembayar	lipan	○	×	○	—	○	1.000	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
migai	muat	×	—	×	—	×	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
migai	tunggu	×	—	×	—	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
migai	mengadakan	×	—	×	—	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
migai	memerintah	×	—	○	—	○	0.267	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
migai	mengawal	×	—	×	—	×	0.233	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
migai	memegang	×	—	○	—	○	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
migai	menguasai	×	—	○	—	○	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×

Iban	Malay	Human decision					OTIC score	Decision by OTIC Filtering																																	
		Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard		$\alpha = 0, \beta = 0$	$\alpha = 0, \beta = 0.2$	$\alpha = 0, \beta = 0.4$	$\alpha = 0, \beta = 0.6$	$\alpha = 0, \beta = 0.8$	$\alpha = 0, \beta = 1.0$	$\alpha = 0.2, \beta = 0$	$\alpha = 0.2, \beta = 0.2$	$\alpha = 0.2, \beta = 0.4$	$\alpha = 0.2, \beta = 0.6$	$\alpha = 0.2, \beta = 0.8$	$\alpha = 0.2, \beta = 1.0$	$\alpha = 0.4, \beta = 0$	$\alpha = 0.4, \beta = 0.2$	$\alpha = 0.4, \beta = 0.4$	$\alpha = 0.4, \beta = 0.6$	$\alpha = 0.4, \beta = 0.8$	$\alpha = 0.4, \beta = 1.0$	$\alpha = 0.6, \beta = 0$	$\alpha = 0.6, \beta = 0.2$	$\alpha = 0.6, \beta = 0.4$	$\alpha = 0.6, \beta = 0.6$	$\alpha = 0.6, \beta = 0.8$	$\alpha = 0.6, \beta = 1.0$	$\alpha = 0.8, \beta = 0$	$\alpha = 0.8, \beta = 0.2$	$\alpha = 0.8, \beta = 0.4$	$\alpha = 0.8, \beta = 0.6$	$\alpha = 0.8, \beta = 0.8$	$\alpha = 0.8, \beta = 1.0$	$\alpha = 1.0, \beta = 0$	$\alpha = 1.0, \beta = 0.2$	$\alpha = 1.0, \beta = 0.4$	$\alpha = 1.0, \beta = 0.6$	$\alpha = 1.0, \beta = 0.8$	$\alpha = 1.0, \beta = 1.0$
migai	menahan	×	—	×	—	×	0.056	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
migai	bertahan	×	—	×	—	×	0.044	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
siping	separa	○	—	—	—	○	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
siping	berat sebelah	○	—	○	—	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
berauh	mengaum	○	—	—	—	○	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
berauh	menggemakan	○	—	—	—	○	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
berauh	meraung	○	—	○	—	○	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
berauh	berteriak	×	—	○	—	○	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
berauh	bergema	○	—	×	—	○	0.182	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
berauh	memekik	×	—	×	—	×	0.154	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×
berauh	menjerit	×	—	○	—	○	0.133	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
berauh	meniru	×	—	×	—	×	0.100	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
berauh	bersorak	×	—	○	—	○	0.050	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
ngulih	memperoleh	○	○	○	○	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
ngulih	kena	×	×	×	×	×	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
ngulih	sampai	×	×	×	×	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
ngulih	mendapat	○	○	○	○	○	0.308	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
ngulih	mengambil	×	○	×	×	×	0.143	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
ngulih	menerima	×	×	×	—	×	0.143	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
ngulih	mengenakan	×	×	×	×	×	0.143	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
ngulih	memahami	×	×	×	×	×	0.037	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
kapur	cat kapur	○	○	×	—	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
kapur	limau nipis	×	×	×	—	×	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
kapur	kapur	○	○	○	○	○	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×

		Human decision						Decision by OTIC Filtering																																			
Iban	Malay	Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard	OTIC score	$\alpha = 0, \beta = 0$	$\alpha = 0, \beta = 0.2$	$\alpha = 0, \beta = 0.4$	$\alpha = 0, \beta = 0.6$	$\alpha = 0, \beta = 0.8$	$\alpha = 0, \beta = 1.0$	$\alpha = 0.2, \beta = 0$	$\alpha = 0.2, \beta = 0.2$	$\alpha = 0.2, \beta = 0.4$	$\alpha = 0.2, \beta = 0.6$	$\alpha = 0.2, \beta = 0.8$	$\alpha = 0.2, \beta = 1.0$	$\alpha = 0.4, \beta = 0$	$\alpha = 0.4, \beta = 0.2$	$\alpha = 0.4, \beta = 0.4$	$\alpha = 0.4, \beta = 0.6$	$\alpha = 0.4, \beta = 0.8$	$\alpha = 0.4, \beta = 1.0$	$\alpha = 0.6, \beta = 0$	$\alpha = 0.6, \beta = 0.2$	$\alpha = 0.6, \beta = 0.4$	$\alpha = 0.6, \beta = 0.6$	$\alpha = 0.6, \beta = 0.8$	$\alpha = 0.6, \beta = 1.0$	$\alpha = 0.8, \beta = 0$	$\alpha = 0.8, \beta = 0.2$	$\alpha = 0.8, \beta = 0.4$	$\alpha = 0.8, \beta = 0.6$	$\alpha = 0.8, \beta = 0.8$	$\alpha = 0.8, \beta = 1.0$	$\alpha = 1.0, \beta = 0$	$\alpha = 1.0, \beta = 0.2$	$\alpha = 1.0, \beta = 0.4$	$\alpha = 1.0, \beta = 0.6$	$\alpha = 1.0, \beta = 0.8$	$\alpha = 1.0, \beta = 1.0$
kapur	pokok limau nipis	×	×	×	—	×	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
ketak	menebang	○	×	×	—	×	0.667	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
ketak	mengerat	○	○	×	○	○	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
ketak	menggunting	×	○	×	○	○	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
ketak	menghiris	×	×	×	—	×	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×
ketak	mencantas	○	×	×	○	○	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×
ketak	memotong	○	○	×	○	○	0.091	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
ketak	mengurangkan	×	×	×	×	×	0.053	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
lambing	melendut	○	×	—	—	○	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
tuchi	tulen	○	○	○	○	○	0.364	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
tuchi	jati	○	×	×	—	×	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	○	○	○	○	×	×
tuchi	suci	○	○	○	○	○	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×
tuchi	bersih	○	○	○	○	○	0.216	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
tuchi	semata-mata	×	×	×	×	×	0.083	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
tuchi	kemas	○	×	×	—	×	0.078	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
tegap	kukuh	○	○	○	○	○	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
tegap	pekat	×	×	×	×	×	0.154	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×
tegap	tegas	○	×	×	×	×	0.130	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
tegap	teguh	○	○	○	○	○	0.111	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×
tegap	kuat	○	○	○	○	○	0.087	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
tegap	tetap	○	—	×	—	○	0.077	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
ngetam	rata	×	—	×	—	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
ngetam	licin	×	—	×	—	×	0.308	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×
bepok	bersembunyi	×	—	—	—	×	0.292	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

		Human decision						Decision by OTIC Filtering																																			
Iban	Malay	Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard	OTIC score	$\alpha=0,\beta=0$	$\alpha=0,\beta=0.2$	$\alpha=0,\beta=0.4$	$\alpha=0,\beta=0.6$	$\alpha=0,\beta=0.8$	$\alpha=0,\beta=1.0$	$\alpha=0.2,\beta=0$	$\alpha=0.2,\beta=0.2$	$\alpha=0.2,\beta=0.4$	$\alpha=0.2,\beta=0.6$	$\alpha=0.2,\beta=0.8$	$\alpha=0.2,\beta=1.0$	$\alpha=0.4,\beta=0$	$\alpha=0.4,\beta=0.2$	$\alpha=0.4,\beta=0.4$	$\alpha=0.4,\beta=0.6$	$\alpha=0.4,\beta=0.8$	$\alpha=0.4,\beta=1.0$	$\alpha=0.6,\beta=0$	$\alpha=0.6,\beta=0.2$	$\alpha=0.6,\beta=0.4$	$\alpha=0.6,\beta=0.6$	$\alpha=0.6,\beta=0.8$	$\alpha=0.6,\beta=1.0$	$\alpha=0.8,\beta=0$	$\alpha=0.8,\beta=0.2$	$\alpha=0.8,\beta=0.4$	$\alpha=0.8,\beta=0.6$	$\alpha=0.8,\beta=0.8$	$\alpha=0.8,\beta=1.0$	$\alpha=1.0,\beta=0$	$\alpha=1.0,\beta=0.2$	$\alpha=1.0,\beta=0.4$	$\alpha=1.0,\beta=0.6$	$\alpha=1.0,\beta=0.8$	$\alpha=1.0,\beta=1.0$
bepok	menyorok	×	—	—	—	×	0.292	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
sapit	kembar	○	○	○	○	○	1.000	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
jamah	perbuatan	×	×	×	—	×	0.182	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
jamah	menewaskan	×	×	×	—	×	0.125	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
jamah	melanggar	×	×	×	×	×	0.100	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
jamah	menyerang	○	×	○	—	○	0.083	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
jamah	mengalahkan	○	×	×	×	×	0.071	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
jamah	memukul	○	○	×	—	○	0.054	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
penau	ilmu pengetahuan	○	—	○	—	○	0.667	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
penau	keupayaan	○	—	○	○	○	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
penau	kebolehan	○	○	○	○	○	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
penau	kemampuan	○	×	×	○	○	0.250	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
mikir	fikir	○	×	○	—	○	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
mikir	fail	×	×	×	×	×	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
mikir	kikir	×	×	○	×	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
mikir	memikirkan	○	○	○	×	○	0.100	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
rajin	gigih	○	○	×	○	○	0.375	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
rajin	tekun	○	○	×	—	○	0.316	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
rajin	rajin	○	○	○	○	○	0.182	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
rajin	teliti	○	×	×	×	×	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×
rajin	tabah	○	○	×	×	○	0.125	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
rajin	tajam	×	×	×	×	×	0.091	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
benat	leper	○	×	—	—	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
benat	pipih	○	×	—	—	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

		Human decision						Decision by OTIC Filtering																																			
Iban	Malay	Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard	OTIC score	$\alpha = 0, \beta = 0$	$\alpha = 0, \beta = 0.2$	$\alpha = 0, \beta = 0.4$	$\alpha = 0, \beta = 0.6$	$\alpha = 0, \beta = 0.8$	$\alpha = 0, \beta = 1.0$	$\alpha = 0.2, \beta = 0$	$\alpha = 0.2, \beta = 0.2$	$\alpha = 0.2, \beta = 0.4$	$\alpha = 0.2, \beta = 0.6$	$\alpha = 0.2, \beta = 0.8$	$\alpha = 0.2, \beta = 1.0$	$\alpha = 0.4, \beta = 0$	$\alpha = 0.4, \beta = 0.2$	$\alpha = 0.4, \beta = 0.4$	$\alpha = 0.4, \beta = 0.6$	$\alpha = 0.4, \beta = 0.8$	$\alpha = 0.4, \beta = 1.0$	$\alpha = 0.6, \beta = 0$	$\alpha = 0.6, \beta = 0.2$	$\alpha = 0.6, \beta = 0.4$	$\alpha = 0.6, \beta = 0.6$	$\alpha = 0.6, \beta = 0.8$	$\alpha = 0.6, \beta = 1.0$	$\alpha = 0.8, \beta = 0$	$\alpha = 0.8, \beta = 0.2$	$\alpha = 0.8, \beta = 0.4$	$\alpha = 0.8, \beta = 0.6$	$\alpha = 0.8, \beta = 0.8$	$\alpha = 0.8, \beta = 1.0$	$\alpha = 1.0, \beta = 0$	$\alpha = 1.0, \beta = 0.2$	$\alpha = 1.0, \beta = 0.4$	$\alpha = 1.0, \beta = 0.6$	$\alpha = 1.0, \beta = 0.8$	$\alpha = 1.0, \beta = 1.0$
benat	rata	○	×	—	—	○	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×
benat	datar	○	×	—	—	○	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
benat	lembap	○	×	—	—	○	0.087	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
benat	kabur	○	×	—	—	○	0.069	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
kilas	meleding	○	—	×	—	○	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
peraia	major	○	—	—	—	○	0.667	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
peraia	luas	○	—	—	—	○	0.133	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
peraia	utama	○	—	—	—	○	0.091	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
peraia	besar	○	—	—	—	○	0.080	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
selempepai	pecah	○	—	—	—	○	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
selempepai	memecahkan	○	—	—	—	○	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
selempepai	meletup	○	—	—	—	○	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
penyadi	syarat	×	×	×	×	×	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
penyadi	keadaan	○	×	×	—	×	0.200	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×
penyadi	peristiwa	○	×	○	—	○	0.182	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
penyadi	kejadian	○	×	○	○	○	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
tesengki	berlanggar	○	○	×	—	○	0.667	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
sabana	sayu	×	×	×	×	×	0.364	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
sabana	mengadu	○	○	×	○	○	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×
sabana	bersungut	○	×	×	—	×	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×
sabana	sedih	○	×	○	—	○	0.160	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
sabana	menyedihkan	○	×	○	—	○	0.063	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
tarik	mengetatkan	○	○	×	×	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
tarik	mengheret	○	○	×	○	○	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
tarik	menarik	○	○	○	○	○	0.105	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×

		Human decision						Decision by OTIC Filtering																																	
Iban	Malay	Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard	OTIC score	$\alpha = 0, \beta = 0$	$\alpha = 0, \beta = 0.2$	$\alpha = 0, \beta = 0.4$	$\alpha = 0, \beta = 0.6$	$\alpha = 0, \beta = 0.8$	$\alpha = 0, \beta = 1.0$	$\alpha = 0.2, \beta = 0$	$\alpha = 0.2, \beta = 0.2$	$\alpha = 0.2, \beta = 0.4$	$\alpha = 0.2, \beta = 0.6$	$\alpha = 0.2, \beta = 0.8$	$\alpha = 0.2, \beta = 1.0$	$\alpha = 0.4, \beta = 0$	$\alpha = 0.4, \beta = 0.2$	$\alpha = 0.4, \beta = 0.4$	$\alpha = 0.4, \beta = 0.6$	$\alpha = 0.4, \beta = 0.8$	$\alpha = 0.4, \beta = 1.0$	$\alpha = 0.6, \beta = 0$	$\alpha = 0.6, \beta = 0.2$	$\alpha = 0.6, \beta = 0.4$	$\alpha = 0.6, \beta = 0.6$	$\alpha = 0.6, \beta = 0.8$	$\alpha = 0.6, \beta = 1.0$	$\alpha = 0.8, \beta = 0$	$\alpha = 0.8, \beta = 0.2$	$\alpha = 0.8, \beta = 0.4$	$\alpha = 0.8, \beta = 0.6$	$\alpha = 0.8, \beta = 0.8$	$\alpha = 0.8, \beta = 1.0$	$\alpha = 1.0, \beta = 0$	$\alpha = 1.0, \beta = 0.2$	$\alpha = 1.0, \beta = 0.4$	$\alpha = 1.0, \beta = 0.6$	$\alpha = 1.0, \beta = 0.8$	$\alpha = 1.0, \beta = 1.0$
tarik	mencabut	×	×	×	○	×	0.045	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
peinsa	menghidap	○	—	—	×	○	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
peinsa	mengalami	○	—	—	×	○	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
peinsa	menyeksa	○	—	—	×	○	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
sempulang	mengiringi	○	○	○	○	○	0.571	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
sempulang	menemani	○	×	○	—	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
sempulang	teman	○	×	×	—	×	0.182	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
manas	murka	○	—	×	—	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
manas	memupuk	×	—	×	×	×	0.286	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
manas	berang	○	—	×	—	○	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
manas	menanam	×	—	×	×	×	0.143	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
manas	marah	○	—	○	—	○	0.087	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
ngilau	ternampak-nampak	○	—	—	—	○	0.667	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
ngilau	mengunjungi	○	—	—	—	○	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
ngilau	menyaksikan	○	—	—	—	○	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×
ngilau	menengok	○	—	—	—	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
ngilau	nampak	○	—	—	—	○	0.364	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
ngilau	melawat	○	—	—	—	○	0.333	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
ngilau	memastikan	○	—	—	—	○	0.182	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
ngilau	mengetahui	○	—	—	—	○	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
ngilau	memeriksa	○	—	—	—	○	0.125	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
ngilau	melihat	○	—	—	—	○	0.105	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
megai	muat	×	×	×	×	×	0.667	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
megai	tunggu	×	×	×	×	×	0.417	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×

		Human decision						Decision by OTIC Filtering																																			
Iban	Malay	Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard	OTIC score	$\alpha=0,\beta=0$	$\alpha=0,\beta=0.2$	$\alpha=0,\beta=0.4$	$\alpha=0,\beta=0.6$	$\alpha=0,\beta=0.8$	$\alpha=0,\beta=1.0$	$\alpha=0.2,\beta=0$	$\alpha=0.2,\beta=0.2$	$\alpha=0.2,\beta=0.4$	$\alpha=0.2,\beta=0.6$	$\alpha=0.2,\beta=0.8$	$\alpha=0.2,\beta=1.0$	$\alpha=0.4,\beta=0$	$\alpha=0.4,\beta=0.2$	$\alpha=0.4,\beta=0.4$	$\alpha=0.4,\beta=0.6$	$\alpha=0.4,\beta=0.8$	$\alpha=0.4,\beta=1.0$	$\alpha=0.6,\beta=0$	$\alpha=0.6,\beta=0.2$	$\alpha=0.6,\beta=0.4$	$\alpha=0.6,\beta=0.6$	$\alpha=0.6,\beta=0.8$	$\alpha=0.6,\beta=1.0$	$\alpha=0.8,\beta=0$	$\alpha=0.8,\beta=0.2$	$\alpha=0.8,\beta=0.4$	$\alpha=0.8,\beta=0.6$	$\alpha=0.8,\beta=0.8$	$\alpha=0.8,\beta=1.0$	$\alpha=1.0,\beta=0$	$\alpha=1.0,\beta=0.2$	$\alpha=1.0,\beta=0.4$	$\alpha=1.0,\beta=0.6$	$\alpha=1.0,\beta=0.8$	$\alpha=1.0,\beta=1.0$
megai	mengadakan	×	×	×	×	×	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
megai	memegang	○	○	○	○	○	0.250	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
megai	memerintah	○	×	○	○	○	0.143	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
megai	mengawal	○	○	×	—	○	0.140	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
megai	menahan	×	×	×	×	×	0.057	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
megai	bertahan	×	×	×	×	×	0.048	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
sakit	sakit	○	○	○	○	○	0.545	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
sakit	uzur	○	○	×	○	○	0.462	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×
sakit	malang	○	×	×	—	×	0.111	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
sakit	buruk	×	×	×	×	×	0.103	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
sakit	jahat	×	×	×	×	×	0.069	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
gap	ada kelas	×	—	○	×	×	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
gap	segak	○	○	×	○	○	0.364	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×
gap	pantas	×	×	×	×	×	0.125	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
gap	bijak	×	×	×	×	×	0.118	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
gap	pintar	×	×	×	×	×	0.091	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
gap	kuat	×	×	×	×	×	0.083	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
galau	menempah	×	×	×	×	×	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
galau	menyimpan	○	○	○	○	○	0.133	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
bengkah	divisi	×	○	○	○	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
bengkah	macam	×	×	○	○	○	0.250	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×
bengkah	kelas	×	×	○	—	×	0.222	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
bengkah	golongan	○	×	—	○	○	0.167	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
bengkah	jenis	○	—	—	○	○	0.154	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
bengkah	pembahagian	○	×	○	—	○	0.133	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×

		Human decision						Decision by OTIC Filtering																																			
Iban	Malay	Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Gold Standard	OTIC score	$\alpha = 0, \beta = 0$	$\alpha = 0, \beta = 0.2$	$\alpha = 0, \beta = 0.4$	$\alpha = 0, \beta = 0.6$	$\alpha = 0, \beta = 0.8$	$\alpha = 0, \beta = 1.0$	$\alpha = 0.2, \beta = 0$	$\alpha = 0.2, \beta = 0.2$	$\alpha = 0.2, \beta = 0.4$	$\alpha = 0.2, \beta = 0.6$	$\alpha = 0.2, \beta = 0.8$	$\alpha = 0.2, \beta = 1.0$	$\alpha = 0.4, \beta = 0$	$\alpha = 0.4, \beta = 0.2$	$\alpha = 0.4, \beta = 0.4$	$\alpha = 0.4, \beta = 0.6$	$\alpha = 0.4, \beta = 0.8$	$\alpha = 0.4, \beta = 1.0$	$\alpha = 0.6, \beta = 0$	$\alpha = 0.6, \beta = 0.2$	$\alpha = 0.6, \beta = 0.4$	$\alpha = 0.6, \beta = 0.6$	$\alpha = 0.6, \beta = 0.8$	$\alpha = 0.6, \beta = 1.0$	$\alpha = 0.8, \beta = 0$	$\alpha = 0.8, \beta = 0.2$	$\alpha = 0.8, \beta = 0.4$	$\alpha = 0.8, \beta = 0.6$	$\alpha = 0.8, \beta = 0.8$	$\alpha = 0.8, \beta = 1.0$	$\alpha = 1.0, \beta = 0$	$\alpha = 1.0, \beta = 0.2$	$\alpha = 1.0, \beta = 0.4$	$\alpha = 1.0, \beta = 0.6$	$\alpha = 1.0, \beta = 0.8$	$\alpha = 1.0, \beta = 1.0$
bengkah	bahagian	○	○	○	—	○	0.095	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
bender	setia	×	—	—	—	×	0.176	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
bender	benar	×	—	—	—	×	0.111	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×
bender	jati	×	—	—	—	×	0.091	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
perumba	ras	×	—	—	—	×	1.000	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
perumba	lumba	×	—	○	—	○	1.000	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
perumba	persaingan	×	—	○	—	○	0.533	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
perumba	kaum	×	—	—	—	×	0.500	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	○	○	×	×	×	×	○	○	×	×	×	×
perumba	perlumbaan	×	—	○	—	○	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
perumba	bangsa	×	—	×	—	×	0.400	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
perumba	keturunan	×	—	×	—	×	0.095	○	○	○	○	○	○	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×	○	×	×	×	×	×
pengelikun	sekuriti	×	—	○	—	○	0.667	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
pengelikun	keselamatan	○	○	○	○	○	0.444	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	×	×	×

APPENDIX F

EVALUATION RESULTS OF 500 TRANSLATION SETS FROM LEXICON+TX

Lexicon+TX

A satisficer's multilingual lexicon

#11916

English	Bahasa Melayu	中文	
• undernourishment (LI#279241$_{[N]}$)	• kurang pemakanan (LI#68406$_{[N]}$)	• 营养不良 (LI#380266$_{[N]}$)	3

#6061

English	Bahasa Melayu	中文	
• lecture (LI#195585$_{[V]}$)	• mensyarahkan (LI#40501$_{[V]}$) • menguliahkan (LI#40502$_{[V]}$) • menceramahkan (LI#40503$_{[V]}$)	• 讲演 (LI#384838$_{[V]}$)	3

#12995

English	Bahasa Melayu	中文	
• end up (LI#155988$_{[]}$)	• berakhir (LI#96566$_{[]}$)	• 告终 (LI#313835$_{[V]}$)	3

#5699

English	Bahasa Melayu	中文	
• investor (LI#188788$_{[N]}$)	• pelabur (LI#38100$_{[N]}$) • penanam modal (LI#38101$_{[N]}$)	• 投资者 (LI#336848$_{[N]}$)	3

#10954

English	Bahasa Melayu	中文	
• sumptuous (LI#264631$_{[A]}$)	• mewah (LI#461$_{[A]}$) • mewah (LI#97629$_{[]}$)	• 奢侈 (LI#321818$_{[A]}$)	3

#415

English	Bahasa Melayu	中文	
• angle (LI#111457$_{[N]}$)	• penjuru (LI#3749$_{[N]}$) • sudut (LI#2480$_{[N]}$)	• 角 (LI#384007$_{[N]}$)	3

#174

English	Bahasa Melayu	中文	
• Advanced (LI#107088$_{[A]}$)	• lanjut (LI#1649$_{[A]}$)	• 高等 (LI#402031$_{[A]}$) • 先进 (LI#301697$_{[A]}$)	0

#12366

English	Bahasa Melayu	中文	
• voluntary (LI#284281$_{[A]}$)	• voluntari (LI#70198$_{[N]}$) • pemain organ di gereja (LI#70199$_{[N]}$) • kehendak sendiri (LI#70201$_{[A]}$)	• 自愿 (LI#377532$_{[A]}$)	2

- sukarela (LI#70196[N])
- sukarela (LI#70200[A])

#1442

English
- belfry (LI#122011[N])
- campanile (LI#129419[N])

Bahasa Melayu
- menara loceng (LI#8122[N])

中文
- 钟楼 (LI#394378[N])

3

#2936

English
- dim (LI#149052[A])

Bahasa Melayu
- samar (LI#418[A])
- kabur (LI#416[A])
- kabur (LI#102334[I])

中文
- 模糊不清 (LI#349955[N])
- 模糊不清 (LI#349956[A])

3

#3021

English
- disinfect (LI#150140[V])

Bahasa Melayu
- menyahjangkit (LI#21980[V])
- membersihkan dari bakteria (LI#21981[V])

中文
- 消毒 (LI#355692[V])

3

#9878

English
- earnest (LI#153541[N])
- serious (LI#252691[A])

Bahasa Melayu
- sungguh-sungguh (LI#23664[A])

中文
- 认真 (LI#384608[A])
- 认真 (LI#384609[N])

3

#3263

English
- duty (LI#153191[N])

Bahasa Melayu
- duti (LI#23541[N])
- kewajiban (LI#23542[N])
- cukai (LI#18593[N])

中文
- 义务 (LI#294694[N])

2

#8648

English
- accurate (LI#105570[A])
- exact (LI#159395[A])
- precise (LI#233868[A])
- punctual (LI#238199[A])

Bahasa Melayu
- tepat (LI#801[A])
- tepat (LI#97423[I])

中文
- 精确 (LI#371404[A])

2

#890

English
- barge (LI#119186[V])

Bahasa Melayu
- merempuh (LI#7457[V])

中文
- 相撞 (LI#365758[V])

3

#550

English
- apricot (LI#113662[N])

Bahasa Melayu
- aprikot (LI#4926[N])

中文
- 杏子 (LI#347159[N])

3

#8093

English
- platinum (LI#230858[N])

Bahasa Melayu
- platinum (LI#52078[N])

中文
- 白金 (LI#364630[N])
- 铂 (LI#394658[N])

3

#5994

English
- lariat (LI#194542[N])
- lasso (LI#194685[N])

Bahasa Melayu
- laso (LI#40180[N])
- tanjul (LI#13814[N])

中文
- 套索 (LI#321765[N])

3

#7796

English
- penicillin (LI#225773[N])

Bahasa Melayu
- penisilin (LI#50284[N])

中文
- 盘尼西林 (LI#365305[N])

3

#4295

English
- freight (LI#166495[N])

Bahasa Melayu
- kargo (LI#12291[N])
- muatan (LI#11069[N])
- tambang muatan (LI#29404[N])

中文
- 货运 (LI#386835[N])

2

#7975

English
- picnic (LI#229439[N])

Bahasa Melayu
- piknik (LI#51524[N])
- perkelahan (LI#26312[N])

中文
- 野餐 (LI#393757[N])

3

#11011

English
- supply (LI#265218[N])

Bahasa Melayu
- penawaran (LI#64286[N])
- pembekalan (LI#54986[N])
- bekalan (LI#25558[N])
- persediaan (LI#16069[N])

中文
- 补给 (LI#382736[N])

2

#9872

English
- sequence (LI#252561[N])
- serial (LI#252632[A])
- series (LI#252667[N])

Bahasa Melayu
- siri (LI#59870[N])

中文
- 连续 (LI#390959[N])
- 丛书 (LI#292995[N])

1

#6232

English
- hepar (LI#177497[I])
- liver (LI#198059[N])

Bahasa Melayu
- hepar (LI#33374[N])
- hati (LI#33061[N])

中文
- 肝脏 (LI#375989[N])

3

#8294

English
- practice (LI#233550[I])
- practice (LI#233551[V])

Bahasa Melayu
- praktis (LI#53139[V])
- menjalankan latihan (LI#53140[V])
- berlatih (LI#26351[V])

中文
- 实习 (LI#324520[V])
- 实践 (LI#324604[V])

1

#8632

English
- puff (LI#237878[T])
- puff (LI#237879[V])

Bahasa Melayu
- mengepulkan asap (LI#54757[V])
- berkepul (LI#54758[V])
- tercungap-cungap (LI#30249[V])

中文
- 喷出 (LI#315603[V])

2

#11818

English
- tutorial (LI#277646[N])

Bahasa Melayu

中文
- 教程 (LI#341599[N])

0

#5404

English
- incorrect (LI#186141[A])
- erroneous (LI#157973[A])

Bahasa Melayu
- khilaf (LI#27061[A])
- silap (LI#647[A])
- silap (LI#96356[T])

中文
- 不正确 (LI#292239[A])

3

#3988

English
- fickle (LI#162762[A])

Bahasa Melayu
- tidak tetap (LI#12039[A])

中文
- 薄情 (LI#381115[A])

0

#213

English
- Africa (LI#107650[N])

Bahasa Melayu
- benua Afrika (LI#1976[N])

中文
- 非洲 (LI#398907[N])

3

#10564

English
- stair (LI#260399[N])
- staircase (LI#260401[N])
- stairway (LI#260404[N])

Bahasa Melayu
- tangga (LI#22758[N])

中文
- 楼梯 (LI#349709[N])

3

#2563

English
- cypress (LI#143915[N])

Bahasa Melayu
- sipres (LI#18760[N])
- pokok saru (LI#18761[N])

中文
- 丝柏 (LI#293238[N])

3

#6274

English
- logical (LI#198426[A])

Bahasa Melayu
- logik (LI#41543[A])

中文
- 逻辑 (LI#392025[N])

3

#10218

English
- clever (LI#136030[A])
- smart (LI#256714[A])

Bahasa Melayu
- bijak (LI#6006[A])
- bijak (LI#100455[T])

中文
- 伶俐 (LI#298610[A])

3

#8263

English
- postpone (LI#233263[V])

Bahasa Melayu
- menunda (LI#19620[V])
- menunda (LI#102572[T])
- menangguhkan (LI#1457[V])
- menangguhkan (LI#99867[T])
- mengundurkan (LI#53017[V])

中文
- 缓办 (LI#373782[V])

3

#3887

English
- fancy (LI#161268[V])

Bahasa Melayu
- menggemar (LI#27134[V])
- menyukai (LI#1847[V])
- menyukai (LI#99764[T])

中文
- 想象 (LI#334467[V])
- 爱好 (LI#359893[V])

2

#10633

English
- staunch (LI#261024[A])

Bahasa Melayu
- dapat dipercayai (LI#15921[A])
- setia (LI#16528[A])
- setia (LI#96836[T])

中文
- 坚强 (LI#317858[A])

0

#8436

English
- private (LI#235341[A])

Bahasa Melayu
- pangkat terendah di dalam askar (LI#53750[N])

中文
- 私立 (LI#368673[A])

0

#2913

English
- differentiate (LI#148714[T])
- differentiate (LI#148715[V])

Bahasa Melayu
- beza balik (LI#21227[V])

中文
- 区分 (LI#308047[V])

0

#12074

English
- urban (LI#281210[A])

Bahasa Melayu
- bandar (LI#69418[A])

中文
- 都市 (LI#392836[N])

3

#11372

English
- thin (LI#271390[A])

Bahasa Melayu
- kurus (LI#9746[A])
- tipis (LI#28393[A])
- nipis (LI#30325[A])
- nipis (LI#102973[T])

中文
- 瘦 (LI#364243[A])
- 细 (LI#372736[A])

1

#5933

English
- knowledge (LI#193258[N])

Bahasa Melayu
- ilmu pengetahuan (LI#39714[N])
- kenal (LI#39715[N])
- makluman (LI#36896[N])
- cam (LI#7656[N])
- kefahaman (LI#15863[N])

中文
- 学问 (LI#323595[N])

2

#8612

English	Bahasa Melayu	中文	
• psychiatry (LI#237484[N])	• ilmu psikiatri (LI#54652[N]) • ilmu penyakit jiwa (LI#54653[N])	• 精神病学 (LI#371431[N])	3

#1294

English	Bahasa Melayu	中文	
• broccoli (LI#126941[N])	• sayur brokoli (LI#10619[N]) • kubis bunga hijau (LI#10620[N])	• 椰菜 (LI#349614[N])	3

#1240

English	Bahasa Melayu	中文	
• brand (LI#126090[N])	• jenama (LI#10233[N])	• 烙印 (LI#358895[N])	0

#9501

English	Bahasa Melayu	中文	
• rust (LI#247928[V])	• mengaratkan (LI#17205[V]) • berkarat (LI#58159[V])	• 生锈 (LI#362864[V])	3

#1844

English	Bahasa Melayu	中文	
• clip (LI#136169[N])	• klip (LI#14648[N]) • pengepit (LI#14647[N]) • sepit (LI#13452[N]) • sisipan (LI#14646[N])	• 夹 (LI#321440[N])	3

#9332

English	Bahasa Melayu	中文	
• ride (LI#246008[V])	• merentasi (LI#57541[V])	• 乘坐 (LI#295015[V])	0

#2154

English	Bahasa Melayu	中文	
• conservatism (LI#139390[N])	• faham konservatisme (LI#16420[N]) • fahaman konservatif (LI#16421[N])	• 保守主义 (LI#299858[N])	3

#9640

English	Bahasa Melayu	中文	
• scare (LI#249598[V])	• menakutkan (LI#6708[V]) • menakutkan (LI#99018[I])	• 惊吓 (LI#334202[V])	3

#287

English	Bahasa Melayu	中文	
• alkyne (LI#108951[I])	• alkuna (LI#2629[N])	• 炔 (LI#358604[N])	3

#10129

English	Bahasa Melayu	中文	
• situational (LI#255752[A])	• keadaan (LI#60941[A])	• 政局 (LI#341277[N])	1

#7309

English	Bahasa Melayu	中文	
• obscure (LI#216941[A])	• kabur (LI#416[A]) • kabur (LI#102334[I])	• 含糊 (LI#313468[A])	3

#10158

English	Bahasa Melayu	中文	
• relax (LI#243472[V]) • slacken (LI#256108[V])	• mengendurkan (LI#56658[V])	• 放松 (LI#341208[V])	3

#3020

English	Bahasa Melayu	中文	
• disillusionment (LI#150124[N])	• kebebasan dari khayalan (LI#21976[N]) • penyedaran (LI#21917[N]) • kekecewaan (LI#13124[N])	• 幻灭 (LI#329473[N])	3

#2262

English	Bahasa Melayu	中文	
• convey (LI#139995[V])	• menyampaikan (LI#15622[V]) • menyampaikan (LI#101539[I]) • membawa (LI#10546[V])	• 转达 (LI#389280[V])	2

#5449

English	Bahasa Melayu	中文	
• indulgent (LI#186575[A])	• ikut hawa nafsu (LI#36657[A]) • manja (LI#27335[A])	• 放纵 (LI#341229[N]) • 放纵 (LI#341228[A])	2

#1498

English	Bahasa Melayu	中文	
• carambola (LI#130065[N])	• belimbing besi (LI#12118[N])	• 杨桃 (LI#347429[N])	3

#7666

English	Bahasa Melayu	中文	
• pare (LI#224195[V])	• memotong (LI#3398[V]) • memotong (LI#97682[I])	• 修剪 (LI#300136[V])	3

#9437

English	Bahasa Melayu	中文	
• rotor (LI#247109[N])	• rotor (LI#57926[N])	• 转子 (LI#389157[N])	3

#1710

English	Bahasa Melayu	中文	
• bosom (LI#125509[N])	• dada (LI#9913[N])	• 胸部 (LI#376601[N])	3

- breast (LI#126389[N])
- chest (LI#133530[N])
- 胸膛 (LI#376595[N])

#4889

English
- eagle (LI#153478[N])
- hawk (LI#176052[N])

Bahasa Melayu
- helang (LI#23637[N])
- rajawali (LI#23638[N])

中文
- 鹰 (LI#403081[N])

3

#12000

English
- calm (LI#129241[A])
- equable (LI#157593[A])
- placid (LI#230406[A])
- quiet (LI#240081[N])
- quiet (LI#240082[I])
- serene (LI#252616[N])
- still (LI#261870[A])
- unperturbed (LI#280217[A])

Bahasa Melayu
- tenang (LI#11667[A])
- tenang (LI#96884[I])

中文
- 平静 (LI#329253[N])
- 平静 (LI#329254[A])

3

#7984

English
- pig (LI#229632[N])
- pig (LI#229633[I])

Bahasa Melayu
- orang tamak atau kurang sopan (LI#51583[N])

中文
- 猪 (LI#361195[N])

0

#1767

English
- chrysanthemum (LI#134939[N])

Bahasa Melayu
- bunga kekwa (LI#14033[N])
- krisantemum (LI#14034[N])

中文
- 菊花 (LI#380076[N])

3

#6922

English
- mortal (LI#209497[A])
- mortal (LI#209496[N])

Bahasa Melayu
- makhluk biasa (LI#45420[N])
- tak kekal (LI#45421[A])
- akan mati (LI#45422[A])
- akan mati (LI#99949[I])
- fana (LI#25290[A])
- manusia (LI#8106[N])
- manusia (LI#97277[I])
- maut (LI#40737[A])
- maut (LI#98476[I])

中文
- 不免一死 (LI#291664[N])
- 凡人 (LI#304626[N])

3

#11871

English
- dinkum (LI#149269[A])
- genuine (LI#169070[A])
- unalloyed (LI#278489[A])

Bahasa Melayu
- benar (LI#1213[A])
- benar (LI#99711[I])

中文
- 真正 (LI#366073[A])

2

#12399

English

Bahasa Melayu

中文

- wake (LI#284645[V])

- membangunkan (LI#5380[V])
- membangunkan (LI#98987[I])

- 激发 (LI#358042[V])

0

#2018

English
- compatibility (LI#138396[I])

Bahasa Melayu
- kesetujuan (LI#15714[N])
- bersesuaian (LI#15713[N])
- keserasian (LI#15329[N])
- kesesuaian (LI#711[N])

中文
- 兼容性 (LI#303337[N])

3

#1444

English
- camphor (LI#129434[N])

Bahasa Melayu
- kamfor (LI#11761[N])
- kapur barus (LI#11762[N])

中文
- 樟脑 (LI#349912[N])

3

#11132

English
- syndrome (LI#266374[N])

Bahasa Melayu
- sindrom (LI#64707[N])
- gejala (LI#51021[N])

中文
- 综合症 (LI#373647[N])

3

#49

English
- abyss (LI#105255[N])
- abyssal (LI#105256[A])

Bahasa Melayu
- abis (LI#492[N])
- abis (LI#493[A])

中文
- 深渊 (LI#356247[N])

3

#7377

English
- old (LI#217919[A])

Bahasa Melayu
- usang (LI#5110[A])

中文
- 旧 (LI#343798[A])

3

#6403

English
- maidan (LI#200668[N])

Bahasa Melayu
- medan (LI#5204[N])

中文
- 操场 (LI#340719[N])

3

#1342

English
- bullfinch (LI#127619[N])

Bahasa Melayu
- burung bullfinch (LI#10970[N])

中文
- 红腹灰雀 (LI#372150[N])

3

#8617

English
- puberty (LI#237759[N])

Bahasa Melayu
- remaja (LI#1551[N])
- remaja (LI#99042[I])

中文
- 青春期 (LI#398641[N])

3

#3186

English
- dress (LI#152327[N])

Bahasa Melayu
- pakaian (LI#4703[N])

中文
- 女装 (LI#322020[N])

0

#4608

English
- gong (LI#172057[N])

Bahasa Melayu
- keromong (LI#31297[N])
- gong (LI#31298[N])
- gendang (LI#23267[N])

中文
- 铜锣 (LI#394736[N])

3

#4717

English
- grit (LI#173182[N])

Bahasa Melayu
- butir pasir (LI#31845[N])
- kersik (LI#31846[N])
- kelasakan (LI#31847[N])
- kerikil (LI#31648[N])
- ketabahan (LI#16524[N])

中文
- 砂砾 (LI#367017[N])

2

#12825

English
- mission (LI#208028[N])

Bahasa Melayu
- misi (LI#86593[N])
- pusat para mubaligh (LI#86594[N])

中文
- 任务 (LI#297764[N])
- 使命 (LI#299325[N])

2

#4895

English
- head (LI#176166[N])

Bahasa Melayu
- ketua (LI#3152[N])
- ketua (LI#97081[I])

中文
- 头 (LI#321293[N])

0

#1505

English
- carbohydrate (LI#130142[N])

Bahasa Melayu
- karbohidrat (LI#12165[N])

中文
- 碳水化合物 (LI#367622[N])
- 醣 (LI#393327[N])

3

#6794

English
- mirthless (LI#207766[A])

Bahasa Melayu
- muram (LI#9212[A])
- muram (LI#97946[I])

中文
- 阴森 (LI#396436[A])

3

#1712

English
- chestnut (LI#133536[N])

Bahasa Melayu
- buah berangan (LI#13517[N])
- perang (LI#13518[A])

中文
- 栗子 (LI#348537[N])

2

#4525

English
- gilt (LI#169987[N])
- gilt (LI#169988[A])
- gilt (LI#169989[I])

Bahasa Melayu
- babi dara (LI#30750[N])
- bersepuh, bersalut emas (LI#30753[A])
- sepuhan (LI#30751[A])

中文
- 镀金 (LI#395268[N])

2

#9057

English
- devout (LI#147836[A])
- pious (LI#230040[A])
- religious (LI#243560[A])

Bahasa Melayu
- salih (LI#20846[A])

中文
- 虔诚 (LI#381415[A])

3

#7161

English
- nightmare (LI#214433[N])

Bahasa Melayu
- mimpi buruk (LI#46804[N])
- pengalaman yang menakutkan (LI#46805[N])
- igauan (LI#19874[N])

中文
- 恶梦 (LI#333820[N])
- 梦魇 (LI#349260[N])

2

#6023

English
- lawsuit (LI#195053[N])

Bahasa Melayu
- tuntutan mahkamah (LI#40373[N])

中文
- 诉讼 (LI#385145[N])

3

#3382

English
- electromagnetic (LI#154881[A])

Bahasa Melayu
- berelektromagnet (LI#24275[A])
- mempunyai kuasa magnet dan elektrik (LI#24276[A])
- elektromagnet (LI#24274[A])

中文
- 电磁 (LI#363288[A])

3

#3258

English
- dust (LI#153144[N])

Bahasa Melayu
- habuk (LI#23501[N])
- debu (LI#21659[N])
- abu (LI#5670[N])

中文
- 灰尘 (LI#358384[N])
- 粉末 (LI#371152[N])
- 尘埃 (LI#326498[N])

3

#7733

English
- patio (LI#224905[N])

Bahasa Melayu
- laman dalam (LI#49963[N])
- patio (LI#49964[N])
- halaman dalam rumah (LI#49965[N])

中文
- 天井 (LI#320691[N])

3

#8272

English
- potential (LI#233344[N])
- potential (LI#233345[A])

Bahasa Melayu
- keupayaan (LI#164[N])
- potensi (LI#53063[N])
- upaya (LI#24030[N])

中文
- 潜力 (LI#357821[N])
- 潜在 (LI#357824[N])
- 潜能 (LI#357856[N])

3

#1917

English
- coffin (LI#137071[N])

Bahasa Melayu
- larung (LI#15125[N])
- peti mayat (LI#15126[N])
- keranda (LI#15124[N])

中文
- 棺材 (LI#349544[N])

3

#2377

English	Bahasa Melayu	中文	
• courtyard (LI#141313[N])	• halaman (LI#15850[N])	• 天井 (LI#320691[N])	3

#10038

English	Bahasa Melayu	中文	
• shy (LI#254566[V])	• takut-takut (LI#60568[V])	• 惊跳 (LI#334259[V])	0

#8840

English	Bahasa Melayu	中文	
• ratio (LI#241532[N]) • proportion (LI#236205[N])	• nisbah (LI#54169[N])	• 比例 (LI#351429[N]) • 比率 (LI#351479[N])	3

#908

English	Bahasa Melayu	中文	
• barite (LI#119206[N]) • baryte (LI#119387[I])	• barit (LI#7460[N])	• 重晶石 (LI#393586[N])	3

#9096

English	Bahasa Melayu	中文	
• interpretation (LI#188261[N]) • rendering (LI#243848[N])	• pentafsiran (LI#19382[N])	• 翻译 (LI#375001[N])	2

#10857

English	Bahasa Melayu	中文	
• sublime (LI#263589[I]) • sublime (LI#263591[N])	• sublim (LI#63742[A]) • luhur (LI#63743[A]) • mulia (LI#21324[A])	• 崇高 (LI#327574[A])	3

#2351

English	Bahasa Melayu	中文	
• count (LI#141053[V])	• membilang (LI#17399[V])	• 计数 (LI#384496[V])	3

#6887

English	Bahasa Melayu	中文	
• monologue (LI#208941[N])	• monolog (LI#45183[N]) • ucapan panjang (LI#45184[N])	• 独白 (LI#360949[N])	3

#10905

English	Bahasa Melayu	中文	
• prosperity (LI#236391[N]) • success (LI#264001[N])	• kemajuan (LI#1638[N]) • kemajuan (LI#102500[I])	• 成功 (LI#335150[N])	2

#9728

English	Bahasa Melayu	中文	
• carving (LI#130864[N]) • engraving (LI#156475[N])	• ukiran (LI#12465[N])	• 雕刻 (LI#398049[A])	3

• sculpture (LI#250682[N])			

#13290

English	Bahasa Melayu	中文	
• assuming (LI#115661[A]) • bumptious (LI#127689[A]) • imperious (LI#184108[A]) • arrogant (LI#114542[A]) • brash (LI#126114[A]) • haughty (LI#175327[A]) • supercilious (LI#264802[A]) • uppish (LI#281015[A]) • uppity (LI#281016[A]) • upstage (LI#281065[A]) • hoity-toity (LI#179165[A]) • overbearing (LI#221516[A]) • pompous (LI#232670[A]) • snooty (LI#257118[A]) • swollen-headed (LI#266006[A])	• angkuh (LI#5432[A]) • angkuh (LI#101515[I]) • angkuh (LI#32921[ADV]) • sombong (LI#32922[ADV]) • sombong (LI#5431[A]) • sombong (LI#96450[I])	• 傲慢 (LI#301285[A])	3

#12247

English	Bahasa Melayu	中文	
• remnant (LI#243719[N]) • vestige (LI#283202[N]) • vestigial (LI#283204[A])	• bekas (LI#8716[N]) • bekas (LI#9506[A]) • bekas (LI#101100[I])	• 遗迹 (LI#392277[N])	0

#5004

English	Bahasa Melayu	中文	
• hibernation (LI#178473[N])	• hibernasi (LI#33702[N]) • penghibernatan (LI#33703[N])	• 冬眠 (LI#303977[N])	3

#2512

English	Bahasa Melayu	中文	
• cup (LI#142997[N])	• cawan (LI#18432[N])	• 茶杯 (LI#379458[N])	3

#2092

English	Bahasa Melayu	中文	
• case (LI#130904[N]) • circumstance (LI#135422[N]) • condition (LI#138892[N])	• keadaan (LI#12483[N])	• 情况 (LI#334111[N])	3

#9613

English	Bahasa Melayu	中文	
• savanna (LI#249236[N])	• savana (LI#58647[N])	• 大草原 (LI#320496[N])	3

#7229

English	Bahasa Melayu	中文	
• notion (LI#216257[N])	• fahaman (LI#22563[N])	• 观念 (LI#383845[N])	3

- anggapan (LI#5899[N])

#1577

English	Bahasa Melayu	中文	
• casualty (LI#131132[N])	• korban (LI#12577[N])	• 受害者 (LI#311205[N])	3

#1168

English	Bahasa Melayu	中文	
• bond (LI#125144[N])	• syer (LI#9707[N]) • bon (LI#9708[N]) • kertas tulisan mutu tinggi (LI#9709[N]) • rekatan (LI#1418[N])	• 债券 (LI#300551[N])	2

#8735

English	Bahasa Melayu	中文	
• queue (LI#240004[N])	• giliran (LI#31179[N])	• 长队 (LI#395619[N])	0

#8371

English	Bahasa Melayu	中文	
• presentable (LI#234660[A])	• elok (LI#24328[A])	• 漂亮 (LI#357608[A])	3

#2879

English	Bahasa Melayu	中文	
• diagnosis (LI#148023[N])	• diagnosis (LI#20913[N]) • pengenalan penyakit (LI#20914[N])	• 诊断 (LI#385153[N])	3

#13259

English	Bahasa Melayu	中文	
• side by side (LI#254705[I])	• berganding bahu (LI#96588[I]) • pada masa yang sama (LI#101953[I]) • beriringan (LI#102232[I])	• 并肩 (LI#329411[V]) • 并肩 (LI#329412[PREP])	2

#4086

English	Bahasa Melayu	中文	
• flatter (LI#164062[V])	• menyanjung-nyanjung (LI#1606[V]) • mengampu (LI#9835[I]) • mengampu (LI#27338[V])	• 阿谀 (LI#396859[V])	3

#10041

English	Bahasa Melayu	中文	
• sick (LI#254659[N])	• kemungkinan muntah (LI#60589[A])	• 病人 (LI#363992[N])	0

#5942

English	Bahasa Melayu	中文	
• labial (LI#193650[A])	• bibir (LI#39789[A])	• 嘴唇 (LI#315891[N])	3

#7468

English	Bahasa Melayu	中文	
• organize (LI#219984[V])	• menyelenggarakan (LI#48327[V]) • mengendalikan (LI#42566[V]) • menyusun (LI#5404[V]) • mengatur (LI#5386[V])	• 组织 (LI#372716[V])	2

#13298

English	Bahasa Melayu	中文	
• talk (LI#267647[N]) • talk (LI#267648[V])	• menjawab dengan kasar (LI#96614[I])	• 谈话 (LI#386048[N])	0

#335

English	Bahasa Melayu	中文	
• ambience (LI#109841[I])	• suasana (LI#3090[N])	• 环境 (LI#361666[N])	3

#2399

English	Bahasa Melayu	中文	
• cradle (LI#141546[N])	• endul (LI#17329[N]) • punca asal (LI#17619[N]) • penyangga (LI#9665[N]) • buaian (LI#12415[N])	• 摇篮 (LI#340303[N]) • 发祥地 (LI#311001[N])	1

#1875

English	Bahasa Melayu	中文	
• coach (LI#136550[V])	• membimbing (LI#14865[V]) • mengajar (LI#14864[V]) • mengajar (LI#97472[I])	• 辅导 (LI#389675[V])	3

#2711

English	Bahasa Melayu	中文	
• dehydrogenase (LI#146107[I])	• dehidrogenase (LI#19769[N])	• 脱氢酶 (LI#376917[N])	3

#12436

English	Bahasa Melayu	中文	
• hostile (LI#180244[A]) • warring (LI#284968[A])	• berseteru (LI#34498[A])	• 敌对 (LI#341441[A])	3

#10714

English	Bahasa Melayu	中文	
• stool (LI#262189[N]) • excrement (LI#159610[N])	• bangku (LI#8216[N]) • pucuk (LI#55861[N]) • tahi (LI#19794[N])	• 凳子 (LI#304747[N]) • 大便 (LI#319905[N])	0

#3619

English	Bahasa Melayu	中文	
• establish (LI#158322[V])	• menubuhkan (LI#16543[V]) • mendirikan (LI#10905[V]) • mengukuhkan (LI#16484[V])	• 确立 (LI#367454[V]) • 设立 (LI#384963[V]) • 建立 (LI#330294[V])	2

#8191

English	Bahasa Melayu	中文	
• polling (LI#232201[T])	• pengundian (LI#24206[N])	• 轮询 (LI#389346[N])	3

#8230

English	Bahasa Melayu	中文	
• dolphin (LI#151428[N]) • porpoise (LI#232949[N])	• ikan lumba-lumba (LI#22684[N])	• 海豚 (LI#355563[N])	3

#5970

English	Bahasa Melayu	中文	
• lamp (LI#194180[N])	• lampu (LI#40023[N]) • pelita (LI#40024[N]) • mentol (LI#10912[N])	• 灯 (LI#358356[N])	3

#1737

English	Bahasa Melayu	中文	
• chloride (LI#134009[N])	• klorida (LI#13709[N])	• 氯化物 (LI#352036[N])	3

#5794

English	Bahasa Melayu	中文	
• jerry (LI#190821[A])	• sentakan (LI#33918[N])	• 偷工减料 (LI#301119[T])	3

#2358

English	Bahasa Melayu	中文	
• countermeasure (LI#141168[N])	• langkah balas (LI#17433[N]) • tindakan penyelamat (LI#17434[N])	• 对策 (LI#325704[A])	3

#1687

English	Bahasa Melayu	中文	
• chase (LI#133111[V])	• menghambat (LI#13354[V]) • mengusir (LI#13357[V]) • mengejar (LI#13353[V]) • mengejar (LI#96740[T])	• 追逐 (LI#391272[V])	2

#4700

English	Bahasa Melayu	中文	
• greenhouse (LI#172981[N])	• rumah hijau (LI#31736[N])	• 温室 (LI#356729[N])	3

#7362

English	Bahasa Melayu	中文	
• offence (LI#217696[N])	• dosa (LI#26137[N])	• 罪过 (LI#374432[N])	3

#4199

English	Bahasa Melayu	中文	
• forceps (LI#165529[N])	• alat angkup (LI#28892[N]) • forseps (LI#28893[N])	• 钳子 (LI#394518[N])	3

#6792

English	Bahasa Melayu	中文	
• gladness (LI#170573[N]) • mirth (LI#207762[N]) • conviviality (LI#140018[N]) • jollification (LI#191111[N])	• kegembiraan (LI#5186[N]) • keriangan (LI#13439[N])	• 欢乐 (LI#350247[N])	3

#3634

English	Bahasa Melayu	中文	
• ethylene (LI#158642[N])	• etilena (LI#25914[N])	• 乙烯 (LI#295065[N])	3

#11428

English	Bahasa Melayu	中文	
• tickle (LI#272358[N])	• geletek (LI#66161[N]) • rasa geli (LI#66162[N])	• 胳肢 (LI#376507[N])	3

#8370

English	Bahasa Melayu	中文	
• present (LI#234637[V])	• menganugerahi (LI#6686[V]) • memberi hadiah (LI#53534[V])	• 赠送 (LI#387518[V])	2

#4802

English	Bahasa Melayu	中文	
• hail (LI#174244[N]) • hailstone (LI#174253[N])	• hujan batu (LI#32390[N])	• 冰雹 (LI#304080[N])	3

#6821

English	Bahasa Melayu	中文	
• mistreatment (LI#208085[N]) • maltreatment (LI#201882[N])	• penzaliman (LI#44827[N]) • penganiayaan (LI#29291[N])	• 虐待 (LI#381400[N])	3

#3399

English	Bahasa Melayu	中文	
• elf (LI#155158[N])	• awang kenik (LI#24368[N])	• 小精灵 (LI#326279[N])	3

#7681

English	Bahasa Melayu	中文	
• parsimonious (LI#224397[A])	• lokek (LI#6609[A]) • lokek (LI#101329[T])	• 节俭 (LI#378422[A])	3

#6740

English	Bahasa Melayu	中文	
• mike (LI#207124[N])	• mikrofon (LI#44289[N])	• 话筒 (LI#385396[N])	3

#7485

English	Bahasa Melayu	中文	
• ossification (LI#220629[N])	• penulangan (LI#48497[N]) • osifikasi (LI#48498[N]) • pengosan (LI#48499[N])	• 骨化 (LI#401661[N])	3

#4381

English	Bahasa Melayu	中文	
• gadget (LI#167642[N])	• perkakas (LI#4699[N])	• 小玩意 (LI#326257[N])	3

#12682

English	Bahasa Melayu	中文	
• woodman (LI#287728[N])	• penebang pokok (LI#71270[N])	• 樵夫 (LI#350071[N])	3

#1613

English	Bahasa Melayu	中文	
• celery (LI#131798[N])	• daun saderi (LI#12858[N])	• 芹菜 (LI#378850[N])	3

#4338

English	Bahasa Melayu	中文	
• fugitive (LI#167073[A])	• orang buruan (LI#29664[N]) • pelarian (LI#25713[N]) • lari (LI#29665[A]) • lari (LI#97799[I])	• 逃亡 (LI#391400[N])	2

#13016

English	Bahasa Melayu	中文	
• accord (LI#105494[I]) • accord (LI#105495[V]) • conform (LI#139076[V]) • fall in with (LI#161057[I])	• bersetuju (LI#536[V]) • bersetuju (LI#96370[I])	• 符合 (LI#370292[PREP]) • 符合 (LI#370293[V])	2

#4011

English	Bahasa Melayu	中文	
• Filipino (LI#163068[N])	• orang Filipina (LI#27808[N])	• 菲律宾人 (LI#380169[N])	3

#6213

English	Bahasa Melayu	中文	
• audience (LI#116871[N]) • listener (LI#197746[N])	• pendengar (LI#6278[N])	• 听众 (LI#313491[N])	3

#2387

English	Bahasa Melayu	中文	
• cowardliness (LI#141401[N])	• ketakutan (LI#6704[N])	• 胆小 (LI#376223[N])	3

#3107

English	Bahasa Melayu	中文	
• doddering (LI#151238[A])	• berjalan terketar-ketar (LI#22582[A])	• 老态龙锺 (LI#375147[A])	3

#3586

English	Bahasa Melayu	中文	
• age (LI#107838[N]) • epoch (LI#157553[N]) • ERA (LI#157748[N])	• zaman (LI#2056[N]) • era (LI#25587[N]) • masa (LI#25586[N])	• 时代 (LI#344008[N])	3

#6005

English	Bahasa Melayu	中文	
• latitude (LI#194837[N]) • latitudinal (LI#194838[A])	• garisan lintang (LI#40291[N]) • latitud (LI#40292[N]) • garisan lintang (LI#40293[A]) • latitud (LI#40294[A])	• 纬度 (LI#372355[N])	3

#11297

English	Bahasa Melayu	中文	
• tepid (LI#269208[A])	• agak panas (LI#65448[A]) • suam (LI#41842[A]) • suam-suam kuku (LI#41843[A])	• 温热 (LI#356772[A])	3

#11816

English	Bahasa Melayu	中文	
• tussle (LI#277625[V])	• berkelahi (LI#21694[V]) • berkelahi (LI#98347[I])	• 打斗 (LI#336044[V])	3

#11975

English	Bahasa Melayu	中文	
• united (LI#279961[A])	• bersatu (LI#3043[A]) • bersatu (LI#99599[I])	• 团结 (LI#316502[N])	3

#13034

English	Bahasa Melayu	中文	
• find out (LI#163281[I])	• menemui (LI#96430[I])	• 找出 (LI#336551[V])	3

#4270

English	Bahasa Melayu	中文	
• crisp (LI#142040[A]) • crisp (LI#142041[I]) • fragile (LI#166239[A])	• rapuh (LI#10581[A])	• 脆 (LI#376669[A])	2

#992

English
- beguile (LI#121913[V])

Bahasa Melayu
- memperdaya (LI#7217[V])
- memperdaya (LI#97932[T])

中文
- 骗 (LI#401594[V])

3

#1380

English
- butanol (LI#128095[T])

Bahasa Melayu
- butanol (LI#11234[N])

中文
- 丁醇 (LI#290510[N])

3

#8970

English
- corrective (LI#140635[A])
- rectification (LI#242523[N])
- redress (LI#242756[N])

Bahasa Melayu
- pembetulan (LI#1483[N])

中文
- 矫正 (LI#366693[N])

3

#8354

English
- preparation (LI#234439[N])
- preparatory (LI#234444[A])

Bahasa Melayu
- persiapan (LI#31931[N])
- persiapan (LI#53476[A])

中文
- 准备 (LI#304372[N])
- 预备 (LI#399756[N])

3

#5579

English
- instrument (LI#187640[N])

Bahasa Melayu
- instrumen (LI#37356[N])
- suratcara (LI#37358[N])
- perkakas (LI#4699[N])
- alatan (LI#35940[N])
- faktor penyebab (LI#37357[N])
- peralatan (LI#744[N])

中文
- 仪器 (LI#297661[N])
- 乐器 (LI#294893[N])

1

#2422

English
- cream (LI#141748[N])

Bahasa Melayu
- krim (LI#17741[N])
- kepala susu (LI#17742[N])
- berpati (LI#17743[A])
- dadih (LI#17740[N])
- berkrim (LI#17744[A])
- inti (LI#17060[N])
- sari (LI#412[N])

中文
- 奶油 (LI#322055[N])

2

#11988

English
- unmanned (LI#280095[A])

Bahasa Melayu
- tak berawak (LI#68881[A])
- tanpa pekerja (LI#68882[A])
- tanpa kakitangan (LI#68883[A])

中文
- 无人 (LI#343195[A])

2

#3666

English
- daily (LI#144382[A])
- everyday (LI#159292[A])

Bahasa Melayu
- setiap hari (LI#18878[A])
- setiap hari (LI#97979[T])

中文
- 日常 (LI#343689[A])
- 每天 (LI#351333[A])

3

#10791

English
- stripe (LI#262868[N])

Bahasa Melayu
- belang (LI#63449[N])
- pangkat (LI#4739[N])

中文
- 条纹 (LI#347335[N])

2

#3449

English
- emplacement (LI#155652[N])

Bahasa Melayu
- tempat (LI#11862[N])

中文
- 位置 (LI#298713[N])
- 安放 (LI#323987[N])

2

#10488

English
- spotlight (LI#259848[N])

Bahasa Melayu
- cahaya sorotan (LI#62448[N])
- tumpuan utama (LI#62449[N])
- lampu sorot (LI#28476[N])

中文
- 聚光灯 (LI#375742[N])

2

#850

English
- balderdash (LI#118668[N])

Bahasa Melayu
- karut (LI#7126[N])
- karut (LI#98564[T])

中文
- 梦呓 (LI#349241[N])

3

#11130

English
- syndicate (LI#266369[N])
- syndicate (LI#266370[T])

Bahasa Melayu
- sindiket (LI#64701[N])

中文
- 辛迪加 (LI#389822[N])

3

#8335

English
- prefab (LI#234103[N])

Bahasa Melayu
- bangunan pasang siap (LI#53357[N])
- pasang siap (LI#53356[N])

中文
- 活动房屋 (LI#354695[N])

3

#2171

English
- consort (LI#139463[N])

Bahasa Melayu
- suami atau isteri raja (LI#16501[N])

中文
- 配偶 (LI#392982[N])

3

#3228

English
- duel (LI#152866[N])

Bahasa Melayu
- pertarungan dua orang (LI#23365[N])

中文
- 决斗 (LI#304166[A])

3

#9948

English

Bahasa Melayu

中文

- glow (LI#170965[I])
- glow (LI#170966[N])
- radiance (LI#240520[N])
- sheen (LI#253674[N])
- sheen (LI#253675[A])

- seri (LI#20749[N])
- seri (LI#99184[I])

- 光彩 (LI#301768[N])
- 光辉 (LI#301856[N])

3

#6509

English
- mare (LI#202434[N])

Bahasa Melayu
- kuda betina (LI#42782[N])

中文
- 母马 (LI#351315[N])

3

#2904

English
- diction (LI#148532[N])

Bahasa Melayu
- diksi (LI#21138[N])
- penyebutan (LI#5534[N])

中文
- 措辞 (LI#339644[N])

3

#619

English
- arrest (LI#114505[I])
- arrest (LI#114506[N])

Bahasa Melayu
- penangkapan (LI#4868[N])
- penahanan (LI#5409[N])

中文
- 逮捕 (LI#391974[N])

3

#12722

English
- worse (LI#288036[A])

Bahasa Melayu
- lebih teruk (LI#71363[N])
- lebih buruk (LI#71364[N])

中文
- 更坏 (LI#345434[A])

3

#5290

English
- immemorial (LI#183813[A])

Bahasa Melayu
- zaman berzaman (LI#35725[A])
- sejak dahulu lagi (LI#35726[A])
- tak terjangkau oleh ingatan (LI#35727[A])

中文
- 太古 (LI#320990[N])

3

#3725

English
- exhibit (LI#159777[V])

Bahasa Melayu
- memperlihatkan (LI#26145[V])
- mempamerkan (LI#26380[V])

中文
- 陈列 (LI#397027[V])
- 展出 (LI#327008[V])
- 显出 (LI#344653[V])

3

#2968

English
- disappear (LI#149730[V])

Bahasa Melayu
- resap (LI#21705[V])
- lenyap (LI#21703[V])
- lesap (LI#21704[V])
- lesap (LI#100624[I])
- ghaib (LI#19918[V])
- hilang (LI#21702[V])
- hilang (LI#97929[I])

中文
- 消失 (LI#355678[V])
- 不见了 (LI#292471[V])

3

#11296

English
- tenuous (LI#269188[A])

Bahasa Melayu
- halus (LI#19835[A])

中文
- 纤细 (LI#372217[N])

3

#1682

English
- charge (LI#132978[N])

Bahasa Melayu
- cas (LI#13298[N])
- rekod pinjaman (LI#13299[N])

中文
- 电荷 (LI#363336[N])

2

#9930

English
- Shape (LI#253506[N])
- form (LI#165804[N])
- form (LI#165805[I])
- figure (LI#162938[N])
- Format (LI#165838[N])

Bahasa Melayu
- rupabentuk (LI#60184[N])
- agar-agar yang dibentuk (LI#60185[N])
- bentuk (LI#27761[N])
- acuan (LI#21163[N])

中文
- 形态 (LI#331594[N])
- 形状 (LI#331603[N])
- 外形 (LI#319337[N])
- 形式 (LI#331588[N])

2

#1375

English
- busy (LI#128059[A])

Bahasa Melayu
- sibuk (LI#11206[A])
- sibuk (LI#96329[I])

中文
- 热闹 (LI#359212[A])

3

#5713

English
- ire (LI#189042[N])

Bahasa Melayu
- kemarahan (LI#3731[N])

中文
- 忿怒 (LI#332894[N])

3

#10922

English
- sufferer (LI#264155[N])

Bahasa Melayu
- penghidap (LI#63953[N])

中文
- 患者 (LI#333976[N])
- 受难者 (LI#311291[N])

2

#2755

English
- denominator (LI#146623[N])

Bahasa Melayu
- penyebut (LI#20082[N])
- angka pembawah pecahan (LI#20083[N])

中文
- 分母 (LI#305335[N])

3

#5935

English
- Koran (LI#193376[N])

Bahasa Melayu
- Quran (LI#39726[N])

中文
- 可兰经 (LI#311873[N])

3

#3764

English
- experiment (LI#160081[N])

Bahasa Melayu
- eksperimen (LI#26538[N])
- ujikaji (LI#26539[N])

中文
- 试验 (LI#385310[A])
- 试验 (LI#385309[N])

3

#4582

English
- gluttonous (LI#171131[A])

Bahasa Melayu
- lahap (LI#31108[A])
- lahap (LI#99916[I])

中文
- 贪吃 (LI#386895[N])
- 贪吃 (LI#386896[A])
- 饕餮 (LI#400563[N])

3

#1214

English
- bow (LI#125746[N])
- bow (LI#125747[I])

Bahasa Melayu
- tundukan (LI#10054[N])
- lengkung sabut (LI#10055[N])

中文
- 船首 (LI#378184[N])

0

#8031

English
- pistol (LI#230211[N])

Bahasa Melayu
- pistol (LI#32110[N])

中文
- 手枪 (LI#335770[N])

3

#9147

English
- republic (LI#244230[N])

Bahasa Melayu
- republik (LI#56922[N])

中文
- 共和国 (LI#302905[N])

3

#2064

English
- computation (LI#138646[N])

Bahasa Melayu
- pengiraan (LI#15894[N])

中文
- 计数 (LI#384497[N])

3

#5821

English
- journey (LI#191177[N])

Bahasa Melayu
- perjalanan (LI#17481[N])
- perjalanan (LI#99021[I])

中文
- 历程 (LI#309888[N])
- 路程 (LI#388468[N])

3

#10398

English
- specimen (LI#258702[N])
- Sample (LI#248648[N])

Bahasa Melayu
- spesimen (LI#62064[N])
- contoh (LI#24678[N])

中文
- 样品 (LI#348604[N])
- 标本 (LI#348396[N])

3

#7991

English
- pilfer (LI#229730[I])
- pilfer (LI#229731[V])

Bahasa Melayu
- mencuri (LI#11094[V])
- mencuri (LI#98321[I])

中文
- 盗 (LI#365276[V])
- 窃 (LI#369685[V])

3

#2922

English
- digest (LI#148796[V])

Bahasa Melayu
- menghadam (LI#21287[V])
- mencerna (LI#21288[V])

中文
- 消化 (LI#355664[V])

3

#7005

English
- bounteous (LI#125704[A])
- generous (LI#168927[A])
- munificent (LI#210582[A])

Bahasa Melayu
- dermawan (LI#10026[A])

中文
- 慷慨 (LI#334914[A])

3

#1160

English
- bole (LI#125028[N])

Bahasa Melayu
- batang pokok (LI#9646[N])

中文
- 树干 (LI#348494[N])

3

#5825

English
- jocund (LI#191014[A])
- jolly (LI#191115[A])
- jovial (LI#191185[A])
- cheerful (LI#133329[A])
- chirpy (LI#133855[A])

Bahasa Melayu
- ria (LI#10511[A])
- riang (LI#9316[A])

中文
- 高兴 (LI#401804[A])
- 快活 (LI#332801[A])
- 快活 (LI#332802[N])

3

#7623

English
- pantaloon (LI#223411[N])
- pants (LI#223467[N])

Bahasa Melayu
- seluar (LI#9395[N])

中文
- 裤子 (LI#383219[N])

3

#542

English
- apprentice (LI#113605[N])

Bahasa Melayu
- murid tukang (LI#4876[N])
- magang (LI#4875[N])
- pelatih (LI#4873[N])
- perantis (LI#4874[N])

中文
- 学徒 (LI#323558[N])
- 徒弟 (LI#331966[N])
- 见习生 (LI#383764[N])

3

#11116

English
- symbolize (LI#266127[V])

Bahasa Melayu
- menyimbolkan (LI#64619[V])
- menggunakan lambang (LI#64620[V])

中文
- 象征 (LI#386440[V])

2

#11349

English
- theological (LI#270890[A])

Bahasa Melayu
- teologis (LI#65709[A])
- teologikal (LI#65710[A])
- berkenaan teologi (LI#65711[A])

中文
- 神学 (LI#368150[N])

3

#3509

English
- dilate (LI#149011[V])
- enlarge (LI#156525[V])

Bahasa Melayu
- membesar (LI#21377[V])

中文
- 扩大 (LI#336312[V])

3

#9071

English

Bahasa Melayu

中文

- cure (LI#143102[N])
- remedial (LI#243654[A])

- pemulihan (LI#16810[N])
- pemulihan (LI#56720[A])

- 治疗 (LI#353612[N])

3

#13030

English
- figure of speech (LI#162944[I])
- figurative (LI#162935[A])

Bahasa Melayu
- metafora (LI#98872[I])
- simili (LI#98873[I])
- kiasan (LI#27757[A])
- kiasan (LI#98874[I])

中文
- 比喻 (LI#351441[N])

3

#4952

English
- helm (LI#176813[N])

Bahasa Melayu
- helm (LI#33233[N])
- kincir kemudi (LI#33235[N])

中文
- 舵 (LI#378117[N])

3

#5660

English
- interview (LI#188396[V])

Bahasa Melayu
- berwawancara (LI#37846[V])
- temuduga (LI#37847[V])
- mewawancara (LI#37848[V])
- menemuramah (LI#37849[V])
- menemuduga (LI#37850[V])
- menginterviu (LI#37851[V])

中文
- 接见 (LI#339468[V])

2

#1065

English
- bind (LI#123353[V])

Bahasa Melayu
- menjilid (LI#8734[V])
- ikat (LI#8731[V])
- tambat (LI#8732[V])

中文
- 包扎 (LI#307614[V])
- 包扎 (LI#307615[I])
- 装订 (LI#383164[V])

3

#10665

English
- sterilize (LI#261594[V])

Bahasa Melayu
- mensteril (LI#63009[V])
- memandulkan (LI#62029[V])
- majir (LI#63010[V])

中文
- 消毒 (LI#355692[V])

2

#3241

English
- dump (LI#152945[V])

Bahasa Melayu
- menghempuk (LI#23426[V])
- melambakkan (LI#23427[V])

中文
- 倾倒 (LI#300595[V])

3

#3820

English
- extravagant (LI#160476[A])

Bahasa Melayu
- boros (LI#26778[A])
- boros (LI#98167[I])
- membazir (LI#26779[A])
- membazir (LI#97643[I])

中文
- 奢侈 (LI#321818[A])

3

#12007

English
- unreasonable (LI#280314[A])
- inordinate (LI#187303[A])
- intemperance (LI#187796[N])
- intemperate (LI#187797[A])

Bahasa Melayu
- keterlaluan (LI#32764[A])
- keterlaluan (LI#26790[N])

中文
- 不合理 (LI#291855[A])
- 过度 (LI#390213[N])

3

#9177

English
- resident (LI#244433[A])

Bahasa Melayu
- kediaman (LI#221[N])

中文
- 居留 (LI#326926[N])

3

#12154

English
- Variable (LI#282337[N])
- Variable (LI#282338[A])
- mutable (LI#210786[A])
- inconstant (LI#186112[A])

Bahasa Melayu
- variabel (LI#69599[N])
- boleh berubah (LI#69600[A])
- pembolehubah (LI#69598[N])
- berubah (LI#13233[A])
- berubah-ubah (LI#12040[A])

中文
- 变量 (LI#311405[N])
- 可变 (LI#311886[N])
- 易变 (LI#344379[N])

2

#2430

English
- credulity (LI#141837[N])

Bahasa Melayu
- mudah percaya (LI#17796[N])

中文
- 轻信 (LI#389482[A])

3

#419

English
- anhydride (LI#111557[N])

Bahasa Melayu
- anhidrida (LI#3819[N])

中文
- 酐 (LI#393023[N])

3

#10814

English
- student (LI#263096[N])

Bahasa Melayu
- pelajar (LI#40454[N])
- penuntut (LI#14325[N])
- mahasiswa (LI#40456[N])

中文
- 学生 (LI#323577[N])

2

#3644

English
- Europe (LI#159023[N])
- European (LI#159024[A])

Bahasa Melayu
- Eropah (LI#26023[N])

中文
- 欧洲 (LI#350326[N])
- 欧洲人 (LI#350329[N])

3

#8450

English
- proceeding (LI#235495[N])

Bahasa Melayu
- tindakan (LI#1125[N])

中文
- 行动 (LI#382469[N])

3

#6619

English
- medulla (LI#203674[N])

Bahasa Melayu
- tulang bahagian tengah (LI#43404[N])

中文
- 延髓 (LI#330237[N])

3

#5707

English	Bahasa Melayu	中文	
• iodine (LI#188893[N])	• iodin (LI#38174[N])	• 碘 (LI#367527[N])	3

#931

English	Bahasa Melayu	中文	
• battalion (LI#119704[N])	• pasukan (LI#7747[N])	• 集团 (LI#397933[N])	3

#6313

English	Bahasa Melayu	中文	
• cute (LI#143477[A]) • lovely (LI#199030[A])	• jelita (LI#3728[A])	• 可爱 (LI#311967[A])	3

#2765

English	Bahasa Melayu	中文	
• denunciation (LI#146697[N])	• kecaman (LI#6157[N]) • kutukan (LI#141[N])	• 谴责 (LI#386301[N])	3

#8922

English	Bahasa Melayu	中文	
• estimate (LI#158378[V]) • reckon (LI#242279[V])	• mengagak (LI#8402[V]) • menganggap (LI#729[V])	• 估计 (LI#298579[V])	3

#9235

English	Bahasa Melayu	中文	
• retaliation (LI#244822[N])	• tindakan balas (LI#17417[N])	• 报复 (LI#337181[N])	3

#9770

English	Bahasa Melayu	中文	
• section (LI#251139[N]) • department (LI#146748[N]) • division (LI#150892[N])	• seksyen (LI#59362[N]) • bahagian (LI#2791[N]) • keratan (LI#14659[N]) • belahan (LI#14532[N])	• 剖面 (LI#306535[N]) • 地段 (LI#317532[N]) • 章节 (LI#369993[N]) • 部门 (LI#392790[N]) • 部分 (LI#392773[N])	0

#10082

English	Bahasa Melayu	中文	
• silly (LI#255155[A])	• dekat dengan pemukul dalam permainan kriket (LI#60745[A]) • si pandir (LI#60744[N]) • si bodoh (LI#5771[N]) • si bodoh (LI#96636[T])	• 糊涂 (LI#371495[A]) • 愚蠢 (LI#334636[A])	2

#1719

English	Bahasa Melayu	中文	
• children (LI#133683[N])	• kanak-kanak (LI#13583[N]) • anak-anak (LI#13584[N])	• 孩子们 (LI#323602[N])	3

#4053

English	Bahasa Melayu	中文	
• first-class (LI#163569[A])	• kelas satu (LI#28008[A]) • kelas satu (LI#102766[T])	• 第一流 (LI#370340[A])	3

#10623

English	Bahasa Melayu	中文	
• statistician (LI#260980[N])	• pakar statistik (LI#62834[N]) • pakar perangkaan (LI#62835[N])	• 统计员 (LI#373348[A])	3

#12835

English	Bahasa Melayu	中文	
• several (LI#253232[PRON]) • several (LI#253231[A])	• beberapa (LI#60046[PRON])	• 几个 (LI#304562[A]) • 几 (LI#304559[A])	3

#2062

English	Bahasa Melayu	中文	
• compressor (LI#138610[N])	• pemampat (LI#15875[N]) • kompresor (LI#15876[N]) • pemadat (LI#15877[N])	• 压缩机 (LI#309939[N])	3

#791

English	Bahasa Melayu	中文	
• aeronautics (LI#107402[N]) • aviation (LI#117658[N])	• penerbangan (LI#1792[N])	• 航空 (LI#378059[N]) • 航空术 (LI#378067[N])	3

#8565

English	Bahasa Melayu	中文	
• prosecute (LI#236298[V])	• mendakwa (LI#814[V])	• 起诉 (LI#387904[V])	3

#1751

English	Bahasa Melayu	中文	
• chop (LI#134513[N])	• cap (LI#13862[N])	• 戳记 (LI#335499[N])	3

#5743

English	Bahasa Melayu	中文	
• isle (LI#189433[N]) • islet (LI#189435[N])	• pulau kecil (LI#38443[N])	• 小岛 (LI#326156[N])	3

#885

English	Bahasa Melayu	中文	
• barbecue (LI#119102[N])	• dapur panggang, biasanya di tempat terbuka (LI#7402[N]) • panggang (LI#7400[N])	• 烤肉 (LI#358994[N])	2

- barbeku (LI#7401[N])

#2816

English
- designate (LI#147292[V])

Bahasa Melayu
- menggelarkan (LI#20530[V])
- menandakan (LI#9585[V])
- melantik (LI#4814[V])

中文
- 指明 (LI#338451[V])
- 指定 (LI#338416[V])

2

#3067

English
- distance (LI#150552[N])

Bahasa Melayu
- kejauhan (LI#22274[N])
- jauhnya (LI#22275[N])

中文
- 距离 (LI#388340[N])

3

#4126

English
- fluent (LI#164624[A])

Bahasa Melayu
- lancar (LI#28558[A])

中文
- 流利 (LI#354816[A])

3

#6635

English
- membrana (LI#204178[I])

Bahasa Melayu
- membran (LI#43538[N])
- selaput (LI#14925[N])

中文
- 膜 (LI#377280[N])

3

#1734

English
- chisel (LI#133862[I])
- chisel (LI#133863[V])
- chip (LI#133797[V])

Bahasa Melayu
- memahat (LI#13670[V])
- mengukir (LI#13642[V])

中文
- 刻 (LI#306237[V])
- 凿 (LI#305152[V])
- 凿 (LI#305155[I])

3

#13302

English
- taste (LI#268119[I])
- taste (LI#268120[V])

Bahasa Melayu
- merasa dengan lidah (LI#65085[V])
- merasa (LI#27407[V])
- mengecap (LI#13868[V])
- sesedap rasa (LI#102596[I])
- mengikut selera atau kegemaran seseorang (LI#102597[I])

中文
- 尝 (LI#326533[V])

2

#4163

English
- follow (LI#165086[V])

Bahasa Melayu
- mengerti (LI#28762[V])

中文
- 听得懂 (LI#313518[V])

3

#380

English
- anaesthetist (LI#110874[N])

Bahasa Melayu
- pakar anestesia (LI#3487[N])
- anestetis (LI#3489[N])
- pakar pengebasan (LI#3490[N])

中文
- 麻醉师 (LI#403278[N])

3

#7806

English
- impecunious (LI#184043[A])
- needy (LI#212555[A])
- penurious (LI#226122[A])

Bahasa Melayu
- miskin (LI#35853[A])
- miskin (LI#97627[I])

中文
- 贫穷 (LI#386965[A])

3

#8830

English
- rapt (LI#241393[A])
- rapt (LI#241394[I])

Bahasa Melayu
- ralit (LI#55797[A])

中文
- 全神贯注 (LI#302292[I])

3

#9542

English
- salad (LI#248343[N])

Bahasa Melayu
- sayur mentah (LI#58322[N])
- ulam (LI#40752[N])
- salad (LI#15229[N])

中文
- 色拉 (LI#378309[N])

3

#173

English
- advance (LI#107077[V])

Bahasa Melayu
- mendahulukan (LI#1645[V])
- memajukan (LI#1643[V])
- meminjamkan (LI#1646[V])
- meningkatkan (LI#1644[V])

中文
- 促进 (LI#299738[V])
- 前进 (LI#306466[V])

0

#9760

English
- secede (LI#251012[V])

Bahasa Melayu
- berpisah (LI#22500[V])
- berpisah (LI#97428[I])

中文
- 分离 (LI#305359[V])

3

#2514

English
- cupboard (LI#143006[N])

Bahasa Melayu
- gerobok (LI#18435[N])
- almari (LI#11412[N])

中文
- 碗柜 (LI#367526[A])

3

#2938

English
- dimension (LI#149074[N])

Bahasa Melayu
- matra (LI#21412[N])

中文
- 尺寸 (LI#326654[N])

3

#5340

English
- impostor (LI#184254[N])

Bahasa Melayu
- penyamar (LI#35906[N])

中文
- 冒牌 (LI#303651[N])

3

#11749

English
- truism (LI#276842[N])

Bahasa Melayu
- kebenaran hakiki (LI#67564[N])

中文
- 陈词滥调 (LI#397075[N])

0

#12707

English
- work-load (LI#287893[N])
- workload (LI#287942[N])

Bahasa Melayu
- beban kerja (LI#71312[N])

中文
- 工作量 (LI#327844[N])

3

#2082

English
- conclusion (LI#138806[N])

Bahasa Melayu
- pembentukan triti (LI#15995[N])
- kesimpulan (LI#14705[N])
- penutup (LI#11024[N])

中文
- 结论 (LI#373135[N])

2

#128

English
- adapt (LI#106493[V])

Bahasa Melayu
- membiasakan (LI#664[V])

中文
- 改编 (LI#341080[V])
- 适应 (LI#391372[V])

2

#3420

English
- embark (LI#155364[V])

Bahasa Melayu
- memuat (LI#11072[V])
- menaiki (LI#5604[V])
- menaiki (LI#97320[I])
- naik (LI#9537[V])

中文
- 乘船 (LI#295032[V])

2

#12704

English
- worker (LI#287913[N])
- labourer (LI#193730[N])

Bahasa Melayu
- petugas (LI#17861[N])
- pekerja (LI#22539[N])

中文
- 工作人员 (LI#327825[N])
- 工人 (LI#327817[N])

1

#12962

English
- by chance (LI#128310[I])

Bahasa Melayu
- menurut apa yang berlaku atau terjadi (LI#97702[I])
- tak dirancang (LI#97703[I])

中文
- 偶然 (LI#301092[PREP])

2

#5481

English
- hospital (LI#180219[N])
- infirmary (LI#186833[N])

Bahasa Melayu
- rumah sakit (LI#34478[N])
- hospital (LI#34477[N])

中文
- 医院 (LI#308116[N])

3

#155

English
- admixture (LI#106922[N])

Bahasa Melayu
- campuran (LI#1328[N])

中文
- 混合物 (LI#356313[N])

3

#8187

English
- political (LI#232162[A])
- politics (LI#232174[N])

Bahasa Melayu
- politik (LI#52530[A])
- ilmu politik (LI#52538[N])

中文
- 政治 (LI#341301[A])
- 政治 (LI#341300[N])

2

- politik (LI#52528[N])
- muslihat (LI#5545[N])

中文
- 政治学 (LI#341308[N])

#7232

English
- notoriety (LI#216275[N])

Bahasa Melayu
- keburukan (LI#7024[N])

中文
- 恶名 (LI#333786[N])

3

#5898

English
- kindred (LI#192718[N])

Bahasa Melayu
- keluarga (LI#9367[N])

中文
- 家族 (LI#324955[N])

3

#10048

English
- sideline (LI#254748[N])

Bahasa Melayu
- kerja sampingan (LI#60631[N])
- garis tepi (LI#60630[N])

中文
- 副业 (LI#306623[N])

2

#8966

English
- redeem (LI#242670[V])

Bahasa Melayu
- bertaubat (LI#29109[V])
- menunaikan (LI#12501[V])

中文
- 赎回 (LI#387361[V])
- 挽回 (LI#338831[V])

1

#13196

English
- peel off (LI#225482[I])

Bahasa Melayu
- mengupas (LI#101307[I])

中文
- 剥掉 (LI#306552[N])

3

#3343

English
- eggshell (LI#154355[N])

Bahasa Melayu
- kulit telur (LI#24057[N])

中文
- 蛋壳 (LI#381723[N])

3

#10699

English
- stitch (LI#261988[N])
- stitch (LI#261989[I])

Bahasa Melayu
- rasa mencucuk (LI#63137[N])
- jahitan (LI#59271[N])
- gulungan (LI#15183[N])

中文
- 针法 (LI#394283[N])

2

#13158

English
- bear in mind (LI#121447[I])
- remember (LI#243665[V])
- keep in mind (LI#191955[I])

Bahasa Melayu
- ingat (LI#7857[V])
- ingat (LI#96900[I])
- mengenang (LI#56723[V])
- mengingat (LI#25133[V])

中文
- 记住 (LI#384778[V])
- 牢记 (LI#360300[V])

3

#8793

English
- rain (LI#240948[N])

Bahasa Melayu
- yang melimpah-limpah (LI#55679[N])

中文
- 雨 (LI#398076[N])

3

- hujan (LI#55678[N])

#1521

English
- careful (LI#130451[A])

Bahasa Melayu
- waspada (LI#6677[A])
- waspada (LI#101747[I])

中文
- 仔细 (LI#297287[A])

3

#197

English
- affect (LI#107539[V])

Bahasa Melayu
- mengafek (LI#1849[V])

中文
- 影响 (LI#331720[V])

3

#11714

English
- trim (LI#275993[V])
- crop (LI#142165[V])
- prune (LI#237029[V])

Bahasa Melayu
- menghiasai (LI#67327[V])
- melangsingkan (LI#67328[V])
- menyesuaikan muatan kapal (LI#67329[V])
- memangkas (LI#18001[V])

中文
- 修剪 (LI#300136[V])

2

#9328

English
- riddle (LI#246002[V])

Bahasa Melayu
- melubangi (LI#50490[V])

中文
- 解谜 (LI#384168[V])

0

#9026

English
- regulate (LI#243274[V])

Bahasa Melayu
- melaraskan (LI#56593[V])
- mengatur (LI#5386[V])

中文
- 调节 (LI#385983[V])

2

#7915

English
- petulant (LI#227314[A])

Bahasa Melayu
- bengis (LI#6130[A])

中文
- 暴躁 (LI#345306[N])

3

#5444

English
- inducement (LI#186547[N])

Bahasa Melayu
- pendorongan (LI#36644[N])
- penggalakan (LI#20724[N])
- pemujukan (LI#4734[N])
- galakan (LI#24763[N])
- perangsangan (LI#30046[N])
- pujukan (LI#4733[N])
- rangsangan (LI#24764[N])

中文
- 诱因 (LI#385631[N])

3

#10430

English
- spicule (LI#259114[N])

Bahasa Melayu
- spikul (LI#62205[N])

中文
- 骨针 (LI#401709[N])

3

#12438

English
- warship (LI#284972[N])

Bahasa Melayu
- kapal perang (LI#7774[N])

中文
- 战船 (LI#335411[N])

3

#10785

English
- string (LI#262832[N])
- cord (LI#140362[N])

Bahasa Melayu
- untaian (LI#63433[N])
- untaian kata (LI#63434[N])
- rentetan (LI#58107[N])
- tali (LI#13888[N])

中文
- 细绳 (LI#372774[N])

2

#8838

English
- ratify (LI#241526[V])

Bahasa Melayu
- meratifikasi (LI#55838[V])
- menguatkan (LI#29128[V])

中文
- 批准 (LI#336493[V])

2

#11805

English
- turn (LI#277421[V])

Bahasa Melayu
- memutarkan (LI#57370[V])
- beredar (LI#40481[V])
- memusingkan (LI#60991[V])
- berbelok (LI#67756[V])
- membelokkan (LI#67759[V])
- berpaling (LI#67757[V])

中文
- 转动 (LI#389112[V])

2

#12867

English
- accept (LI#105381[V])
- agree (LI#108026[V])
- agree to (LI#108032[I])
- assent (LI#115530[V])
- comply (LI#138531[V])

Bahasa Melayu
- menyetujui (LI#584[V])
- menyetujui (LI#96517[I])

中文
- 同意 (LI#312848[V])

3

#8965

English
- redbreast (LI#242655[N])

Bahasa Melayu
- burung kelicap (LI#56309[N])

中文
- 知更鸟 (LI#366641[N])

3

#1884

English
- coastline (LI#136648[N])

Bahasa Melayu
- garisan pantai (LI#14920[N])

中文
- 海岸线 (LI#355422[N])

3

#12231

English
- adept (LI#106704[A])
- adroit (LI#107019[A])
- expert (LI#160095[A])
- proficient (LI#235707[A])
- skillful (LI#255926[A])
- sleight (LI#256315[N])

Bahasa Melayu
- mahir (LI#1597[A])
- mahir (LI#5544[N])
- mahir (LI#99376[I])

中文
- 熟练 (LI#359631[A])

3

- versed (LI#283051[A])

#8587

English	Bahasa Melayu	中文	
• prepare (LI#234447[V]) • provide (LI#236944[V])	• menyediakan (LI#1921[V]) • menyediakan (LI#98986[I])	• 预备 (LI#399755[V])	3

#7365

English	Bahasa Melayu	中文	
• offer (LI#217708[V])	• menghulur (LI#32630[V]) • menguntukkan (LI#2736[V])	• 提供 (LI#339718[V]) • 献祭 (LI#361246[V])	2

#12386

English	Bahasa Melayu	中文	
• vulnerable (LI#284449[A])	• mudah diserang (LI#70252[A]) • rentan (LI#64414[A]) • mudah terpengaruh (LI#36013[A])	• 易受攻击 (LI#344378[A])	3

#3432

English	Bahasa Melayu	中文	
• embryo (LI#155457[N])	• mudigah (LI#24541[N]) • embrio (LI#24542[N]) • janin (LI#24543[N]) • lembaga (LI#6419[N])	• 胚胎 (LI#376340[N])	2

#1257

English	Bahasa Melayu	中文	
• breast (LI#126389[N])	• payu dara (LI#10348[N]) • buah dada (LI#10349[N]) • tetek (LI#10346[N]) • susu (LI#10347[N])	• 胸怀 (LI#376574[N])	2

#4943

English	Bahasa Melayu	中文	
• hegemony (LI#176606[N])	• hegemoni (LI#33176[N]) • kekuasaan sesebuah negara ke atas negara lain (LI#33177[N])	• 霸权 (LI#398564[N])	3

#12502

English	Bahasa Melayu	中文	
• wedge (LI#285560[N])	• baji (LI#18428[N]) • pasak (LI#7749[N])	• 楔形物 (LI#349635[N]) • 楔子 (LI#349631[N]) • 楔子 (LI#349632[A])	3

#11305

English	Bahasa Melayu	中文	
• terminus (LI#269315[N])	• perhentian akhir (LI#65481[N])	• 终点站 (LI#372870[N])	3

	• penghabisan (LI#15793[N]) • tamat (LI#26557[N]) • tamat (LI#96568[I])		

#568

English	Bahasa Melayu	中文	
• arboretum (LI#113893[N])	• aboretum (LI#5067[N]) • tempat semaian (LI#5068[N])	• 植物园 (LI#349570[N])	3

#12358

English	Bahasa Melayu	中文	
• fluent (LI#164624[A]) • voluble (LI#284257[A])	• fasih (LI#28557[A]) • petah (LI#5526[A])	• 流利 (LI#354816[A])	3

#1308

English	Bahasa Melayu	中文	
• browse (LI#127193[I]) • browse (LI#127194[V])	• membaca-baca (LI#10734[V]) • memakan rumput (LI#10733[V]) • melihat-lihat (LI#10735[V])	• 浏览 (LI#355077[V])	2

#11266

English	Bahasa Melayu	中文	
• temper (LI#268939[V])	• membajai (LI#65350[V]) • mencampuri (LI#43344[V]) • mencampuri (LI#102558[I]) • melembutkan (LI#45021[V]) • melembutkan (LI#102782[I])	• 锻炼 (LI#395259[V])	0

#12175

English	Bahasa Melayu	中文	
• vegetate (LI#282644[V])	• tumbuh (LI#30626[V])	• 长大 (LI#395464[V])	0

#10100

English	Bahasa Melayu	中文	
• sin (LI#255333[N]) • sin (LI#255334[I])	• dosa (LI#26137[N])	• 罪 (LI#374414[N])	3

#12787

English	Bahasa Melayu	中文	
• zoo (LI#289716[N])	• zoo (LI#71592[N])	• 动物园 (LI#307128[N])	3

#12372

English	Bahasa Melayu	中文	
• vortex (LI#284327[N])	• pusar (LI#70215[N]) • vorteks (LI#70216[N])	• 旋涡 (LI#343071[N])	3

#8474

English	Bahasa Melayu	中文	
• productivity (LI#235659[N])	• daya pengeluaran (LI#53876[N]) • produktiviti (LI#53877[N]) • pengeluaran (LI#23059[N]) • penghasilan (LI#29611[N])	• 生产力 (LI#362599[N])	3

#8775

English	Bahasa Melayu	中文	
• radicalism (LI#240546[N])	• faham radikalisme (LI#55563[N])	• 激进主义 (LI#358066[N])	3

#8056

English	Bahasa Melayu	中文	
• plaid (LI#230476[N])	• kain berpetak-petak (LI#51965[N]) • kain bulu kambing bercorak genggang (LI#51966[N]) • kain bulu (LI#7080[N])	• 格子花呢 (LI#348808[N])	3

#3049

English	Bahasa Melayu	中文	
• dispute (LI#150410[V]) • conflict (LI#139065[V])	• memperbalah (LI#22184[V]) • membatah (LI#22186[V]) • mempertikai (LI#16797[V]) • berbahas (LI#19207[V]) • berbalah (LI#16180[V]) • bertengkar (LI#5240[V]) • bertengkar (LI#101002[I])	• 争执 (LI#295511[V])	3

#4934

English	Bahasa Melayu	中文	
• Hebrew (LI#176505[A])		• 希伯来人 (LI#328551[N])	0

#161

English	Bahasa Melayu	中文	
• adoration (LI#106979[N])	• pemujaan (LI#1573[N])	• 崇拜 (LI#327558[N])	3

#138

English	Bahasa Melayu	中文	
• adept (LI#106705[N]) • adept (LI#106704[A])	• pakar (LI#16347[N]) • pakar (LI#20881[A]) • pakar (LI#99934[I])	• 行家 (LI#382494[N]) • 擅长 (LI#340697[A])	3

#4032

English	Bahasa Melayu	中文	
• fine (LI#163309[N])	• dendaan (LI#27895[N])	• 罚款 (LI#374376[N]) • 罚款 (LI#374375[A])	3

#10946

English	Bahasa Melayu	中文	
• sum (LI#264568[N])	• hasil tambah (LI#64081[N]) • matematik (LI#27762[N])	• 总数 (LI#333417[N])	2

#8976

English	Bahasa Melayu	中文	
• reed (LI#242824[N])	• rumput mensiang (LI#56367[N])	• 芦苇 (LI#378588[A]) • 簧 (LI#371013[N])	2

#2497

English	Bahasa Melayu	中文	
• cube (LI#142793[N]) • cubic (LI#142802[A])	• kubus (LI#18294[N]) • kuasa tiga (LI#18296[N]) • kuasa tiga (LI#18301[A]) • kubus (LI#18303[A]) • berbentuk kiub (LI#18300[A]) • kubik (LI#18302[A]) • berkubus (LI#18304[A])	• 立方体 (LI#369867[N]) • 立方体 (LI#369868[A])	2

#10472

English	Bahasa Melayu	中文	
• sponge (LI#259682[N])	• span (LI#62345[N])	• 海绵 (LI#355542[N])	3

#4599

English	Bahasa Melayu	中文	
• godchild (LI#171896[N])	• anak pembatisan (LI#31224[N]) • anak angkat (LI#29174[N])	• 教子 (LI#341563[N])	3

#9363

English	Bahasa Melayu	中文	
• ripe (LI#246334[A])	• masak (LI#18353[A])	• 熟 (LI#359601[I]) • 熟 (LI#359603[N])	3

#13186

English	Bahasa Melayu	中文	
• out-and-out (LI#221148[A])	• seluruhnya (LI#98345[I])	• 彻头彻尾 (LI#331756[I])	3

#11746

English	Bahasa Melayu	中文	
• correct (LI#140628[A]) • exact (LI#159395[A]) • punctual (LI#238199[A]) • right (LI#246109[A]) • right (LI#246110[I]) • true (LI#276817[A])	• betul (LI#802[A]) • betul (LI#100795[I])	• 正确 (LI#350658[A])	2

#8408

English	Bahasa Melayu	中文	
• prick (LI#235054[V])	• tertikam (LI#53645[V])	• 刺 (LI#306180[V])	3

#3277

English	Bahasa Melayu	中文	
• eardrum (LI#153495[N])	• gegendang telinga (LI#23648[N]) • membran timpanum (LI#23649[N])	• 鼓膜 (LI#403841[N])	3

#13003

English	Bahasa Melayu	中文	
• lose face (LI#198843[I])	• dihinakan (LI#98642[I]) • hilang maruah (LI#98643[I]) • dihina (LI#98644[I]) • terhina (LI#98641[I])	• 失面子 (LI#321287[V])	3

#10

English	Bahasa Melayu	中文	
• abet (LI#104831[V])	• bersekongkol (LI#124[V]) • bersubahat (LI#121[V]) • bersubahat (LI#99196[I]) • menggalakkan (LI#120[V]) • menggalakkan (LI#98421[I])	• 教唆 (LI#341556[V])	2

#2439

English	Bahasa Melayu	中文	
• chalk (LI#132651[N]) • cretaceous (LI#141922[A])	• kapur (LI#13160[N]) • kapur (LI#17851[A])	• 白垩 (LI#364431[N])	3

#10968

English	Bahasa Melayu	中文	
• sunshine (LI#264715[N])	• sinar matahari (LI#64131[N]) • cahaya suria (LI#64132[N]) • cahaya matahari (LI#64121[N])	• 阳光 (LI#396351[N])	3

#577

English	Bahasa Melayu	中文	
• architect (LI#114013[N])	• arkitek (LI#5154[N]) • jurubina (LI#5155[N])	• 建筑师 (LI#330303[N])	2

#1456

English	Bahasa Melayu	中文	
• candlestick (LI#129593[N])	• batang lilin (LI#11834[N]) • kaki dian (LI#11835[N]) • kaki lilin (LI#11818[N])	• 烛台 (LI#358901[N])	2

#12275

English	Bahasa Melayu	中文	
• foul (LI#166089[A]) • nasty (LI#212087[A]) • vile (LI#283546[A])	• keji (LI#172[A]) • dahsyat (LI#4695[A]) • dahsyat (LI#99931[I])	• 恶劣 (LI#333778[N]) • 恶劣 (LI#333779[A])	3

#11843

English	Bahasa Melayu	中文	
• tyrannical (LI#278034[A]) • tyrannous (LI#278039[A])	• zalim (LI#10774[A]) • zalim (LI#97232[I])	• 暴虐 (LI#345300[A]) • 暴虐 (LI#345301[N])	3

#291

English	Bahasa Melayu	中文	
• allegorical (LI#109167[A])	• ibarat (LI#2670[A])	• 寓意 (LI#325446[N])	3

#9035

English	Bahasa Melayu	中文	
• reign (LI#243325[V])	• menyelubungi (LI#8364[V]) • menyelubungi (LI#97818[I])	• 占优势 (LI#309301[V])	0

#11290

English	Bahasa Melayu	中文	
• tense (LI#269127[A]) • tense (LI#269128[I]) • strained (LI#262443[A]) • taut (LI#268170[A])	• kala (LI#50571[N]) • tegang (LI#1245[A]) • tegang (LI#102626[I])	• 紧张 (LI#371694[A])	3

#12662

English	Bahasa Melayu	中文	
• tact (LI#266817[N]) • wit (LI#287091[N]) • wittiness (LI#287572[N])	• kebijaksanaan (LI#6011[N])	• 机智 (LI#346793[N])	3

#4168

English	Bahasa Melayu	中文	
• following (LI#165150[A])	• umat (LI#28767[N]) • berikut (LI#28768[A]) • penurut (LI#16200[N])	• 下列 (LI#291368[N])	2

#6723

English	Bahasa Melayu	中文	
• microscope (LI#206729[N])	• mikroskop (LI#44313[N])	• 显微镜 (LI#344663[N])	3

#7216

English	Bahasa Melayu	中文	
• nose (LI#215879[N])	• hidung (LI#16317[N])	• 鼻子 (LI#403896[N])	3

#11658

English
- traverse (LI#274972[V])

Bahasa Melayu
- berjalan menyeberangi (LI#67086[V])
- mengedari (LI#14182[V])
- menjelajahi (LI#26596[V])

中文
- 横过 (LI#350053[V])

3

#7538

English
- exceed (LI#159467[V])
- overrun (LI#221836[V])

Bahasa Melayu
- melebihi (LI#10316[V])
- melebihi (LI#96350[]])

中文
- 超过 (LI#388026[V])

3

#7354

English
- odious (LI#217301[A])

Bahasa Melayu
- jelik (LI#31223[A])

中文
- 丑恶 (LI#292690[A])

3

#2350

English
- adviser (LI#107188[N])
- advisory (LI#107191[]])
- advisory (LI#107192[A])
- consultant (LI#139576[N])
- counsellor (LI#141049[N])

Bahasa Melayu
- penasihat (LI#1720[A])
- penasihat (LI#1719[N])
- perunding (LI#16590[N])

中文
- 顾问 (LI#399697[N])

3

#5144

English
- humane (LI#180594[A])

Bahasa Melayu
- berperikemanusiaan (LI#34639[A])

中文
- 人道 (LI#296976[N])

3

#2989

English
- discontinuous (LI#149903[A])

Bahasa Melayu
- terputus (LI#21832[A])
- tak bersinambung (LI#21833[A])
- tak selanjar (LI#21834[A])
- tidak bersambung (LI#21835[A])

中文
- 不连续 (LI#292570[N])

3

#4846

English
- delight (LI#146233[N])
- gay (LI#168696[A])
- happy (LI#174923[A])

Bahasa Melayu
- sukacita (LI#13437[A])
- sukacita (LI#19847[N])

中文
- 愉快 (LI#334539[N])
- 愉快 (LI#334541[A])

3

#2275

English
- cool (LI#140079[A])
- coolness (LI#140099[N])
- cool (LI#140078[N])

Bahasa Melayu
- kenyamanan (LI#3178[N])

中文
- 冷静 (LI#304295[A])
- 凉爽 (LI#304404[A])

3

#7588

English
- painfulness (LI#222762[N])

Bahasa Melayu
- kepedihan (LI#1062[N])
- kesakitan (LI#11115[N])

中文
- 痛苦 (LI#364145[N])
- 痛 (LI#364119[N])

3

#9227

English
- consequence (LI#139377[N])
- consequent (LI#139378[N])
- aftermath (LI#107735[N])
- outcome (LI#221183[N])
- Result (LI#244758[N])
- resultant (LI#244764[A])
- resultant (LI#244765[N])

Bahasa Melayu
- akibat (LI#2011[N])
- akibat (LI#16404[A])
- hasil (LI#2013[N])
- hasil (LI#5710[A])

中文
- 结果 (LI#373099[N])

3

#11638

English
- transmission (LI#274733[N])

Bahasa Melayu
- transmisi (LI#66991[N])
- pembawa kuasa dari enjin ke roda belakang (LI#66992[N])
- penularan (LI#16615[N])
- penyiaran (LI#10601[N])

中文
- 传动 (LI#298289[N])

2

#7433

English
- optics (LI#219715[N])
- optic (LI#219699[A])
- optical (LI#219701[A])

Bahasa Melayu
- ilmu optik (LI#48207[N])
- optik (LI#48200[A])
- optik (LI#48206[N])

中文
- 光学 (LI#301751[N])
- 光学 (LI#301752[A])

3

#7631

English
- Paper (LI#223495[N])

Bahasa Melayu
- esei (LI#25767[N])
- kertas (LI#28746[N])
- kertas kerja (LI#49421[N])
- kertas ujian (LI#49422[N])
- makalah (LI#5513[N])
- disertasi (LI#21924[N])
- dokumen (LI#22566[N])
- akhbar (LI#46727[N])

中文
- 纸 (LI#372549[N])

2

#6704

English
- meter (LI#205484[N])
- metre (LI#205767[N])

Bahasa Melayu
- meter (LI#44033[N])

中文
- 公尺 (LI#302598[N])

3

#10711

English
- stoma (LI#262102[N])

Bahasa Melayu
- stoma (LI#63187[N])
- mulut (LI#39619[N])

中文
- 气孔 (LI#351862[N])

3

#7294

English
- objection (LI#216864[N])
- dissidence (LI#150488[N])

Bahasa Melayu
- pembangkang (LI#22231[N])
- bangkangan (LI#22230[N])
- pembantah (LI#15765[N])
- penolak (LI#21782[N])
- bantahan (LI#13168[N])
- penolakan (LI#19409[N])

中文
- 异议 (LI#330746[N])

3

#6713

English
- mica (LI#205934[N])

Bahasa Melayu
- mika (LI#44162[N])

中文
- 云母 (LI#295883[N])

3

#13149

English
- meet with (LI#203743[T])

Bahasa Melayu
- menemui (LI#96430[T])
- menjumpai (LI#97841[T])

中文
- 遇到 (LI#392084[V])
- 碰到 (LI#367568[V])
- 会见 (LI#298205[V])

3

#4921

English
- heartbeat (LI#176364[N])

Bahasa Melayu
- denyutan jantung (LI#33082[N])

中文
- 心跳 (LI#332532[N])

3

#7290

English
- object (LI#216846[N])

Bahasa Melayu
- objek (LI#47560[N])

中文
- 对象 (LI#325734[N])
- 客体 (LI#324684[N])
- 宾语 (LI#325130[N])

3

#1265

English
- bribe (LI#126531[V])

Bahasa Melayu
- menumbuk rusuk (LI#10440[V])
- merasuah (LI#10438[V])
- merasuah (LI#101183[T])
- menyuap (LI#10441[V])
- memberi rasuah (LI#10439[V])
- memberi rasuah (LI#99428[T])

中文
- 贿赂 (LI#387242[V])
- 收买 (LI#340929[V])
- 行贿 (LI#382566[V])

3

#3856

English
- fade (LI#160812[V])

Bahasa Melayu
- melunturkan (LI#9205[V])
- luntur (LI#26980[V])
- beransur hilang (LI#26982[V])
- memudarkan (LI#23396[V])
- memudarkan (LI#102178[T])
- melayukan (LI#26981[V])
- pudar (LI#26979[V])

中文
- 褪色 (LI#383313[V])
- 褪色 (LI#383314[T])

2

#5793

English
- imperil (LI#184107[V])
- jeopardize (LI#190794[V])

Bahasa Melayu
- membahayakan (LI#24793[V])
- membahayakan (LI#101832[T])

中文
- 危害 (LI#309647[V])

3

#1555

English
- case (LI#130904[N])

Bahasa Melayu
- kes (LI#12481[N])
- kasus (LI#12482[N])
- lapis luar (LI#12480[N])

中文
- 案例 (LI#348970[N])

2

#12891

English
- not bad (LI#215979[T])

Bahasa Melayu
- agak baik (LI#96791[T])
- boleh tahan (LI#96790[T])

中文
- 不错 (LI#292613[A])

3

#3742

English
- exorcism (LI#159944[N])

Bahasa Melayu
- menghalau hantu (LI#26456[N])

中文
- 驱邪 (LI#401372[N])

3

#876

English
- banner (LI#119003[N])

Bahasa Melayu
- kain rentang (LI#7340[N])
- tetunggul (LI#7342[N])
- ranggi panji (LI#7343[N])
- sepanduk (LI#7339[N])

中文
- 横幅 (LI#349990[N])

3

#3719

English
- exempt (LI#159702[V])

Bahasa Melayu
- melepaskan (LI#1036[V])

中文
- 使免除 (LI#299317[V])

3

#10466

English
- Scatter (LI#249667[N])
- Split (LI#259577[A])

Bahasa Melayu
- pecahan (LI#29260[N])

中文
- 分散 (LI#305291[N])

2

#8254

English
- postal (LI#233131[A])

Bahasa Melayu
- pos (LI#52974[A])

中文
- 邮局 (LI#392583[N])

3

#10936

English
- suitcase (LI#264246[N])

Bahasa Melayu
- beg baju (LI#64000[N])

中文
- 手提箱 (LI#335761[N])

3

#8783

English

Bahasa Melayu

中文

- ragged (LI#240878[A])

- bergerigi (LI#38775[A])
- tidak sama (LI#21881[A])

- 参差 (LI#310250[A])

3

#2452

English
- criterion (LI#142065[N])

Bahasa Melayu
- kriteria (LI#17940[N])
- batu uji (LI#17943[N])
- batu uji (LI#96423[T])

中文
- 轨范 (LI#389072[N])

3

#4949

English
- helix (LI#176769[N])

Bahasa Melayu
- lingkar (LI#16886[N])
- spiral (LI#33220[N])
- heliks (LI#33221[N])
- ulir (LI#33222[N])

中文
- 螺旋 (LI#382137[N])

3

#7595

English
- pair (LI#222802[V])

Bahasa Melayu
- berpasang (LI#49198[V])
- menjadikan sepasang (LI#49200[V])
- memasangkan (LI#49197[V])
- terpasang (LI#49199[V])

中文
- 配对 (LI#392993[V])

3

#7938

English
- Phoenician (LI#228068[A])

Bahasa Melayu

中文
- 腓尼基 (LI#377101[N])

0

#2638

English
- debug (LI#145247[V])

Bahasa Melayu
- mengawas silap (LI#19238[V])
- menyahkan pijat (LI#19239[V])
- membetulkan program (LI#19240[V])

中文
- 调试 (LI#385989[V])

3

#9871

English
- consecutive (LI#139365[A])
- serial (LI#252632[A])

Bahasa Melayu
- berturutan (LI#16393[A])
- cerita bersiri (LI#59911[N])
- bersiri (LI#13289[A])
- satu lepas satu (LI#59912[A])
- satu lepas satu (LI#102846[T])
- berangkai (LI#57871[A])

中文
- 连续 (LI#390957[A])
- 连续 (LI#390959[N])

2

#2187

English
- constructive (LI#139547[A])

Bahasa Melayu
- konstruktif (LI#16573[A])

中文
- 建设性 (LI#330312[A])

3

#11864

English
- referee (LI#242933[T])
- referee (LI#242934[V])
- umpire (LI#278424[V])

Bahasa Melayu
- mengadili (LI#5049[V])

中文
- 裁判 (LI#383072[V])

3

#8168

English
- poke (LI#232023[V])

Bahasa Melayu
- menebuk (LI#23153[V])

中文
- 刺 (LI#306180[V])

3

#9620

English
- savoury (LI#249325[A])

Bahasa Melayu
- lazat (LI#4767[A])
- lazat (LI#100625[T])

中文
- 可口 (LI#311889[A])

3

#6155

English
- ligament (LI#196881[N])
- ligamentum (LI#196888[T])

Bahasa Melayu
- ligamen (LI#40978[N])

中文
- 韧带 (LI#399255[N])

3

#1531

English
- carnival (LI#130563[N])

Bahasa Melayu
- keramaian (LI#12337[N])
- pesta (LI#12336[N])

中文
- 狂欢节 (LI#360769[N])

3

#6990

English
- multimeter (LI#210381[N])

Bahasa Melayu
- multimeter (LI#45770[N])

中文
- 万用表 (LI#290698[N])

3

#109

English
- acrobatics (LI#106107[N])

Bahasa Melayu
- akrobat-akrobat (LI#1067[N])
- para akrobat (LI#1068[N])

中文
- 杂技 (LI#346940[N])

2

#8245

English
- luxurious (LI#199601[A])
- plush (LI#231605[A])
- posh (LI#233047[A])

Bahasa Melayu
- mewah (LI#461[A])
- mewah (LI#97629[T])

中文
- 豪华 (LI#386466[A])

3

#6798

English
- miscellany (LI#207813[N])

Bahasa Melayu
- rampaian (LI#44683[N])
- macam-macam (LI#44684[N])
- beraneka (LI#44682[N])

中文
- 杂集 (LI#346983[A])
- 杂录 (LI#346934[A])
- 杂集 (LI#346982[N])
- 杂录 (LI#346933[N])

3

#5030

English
- historian (LI#178928[N])

Bahasa Melayu
- sejarawan (LI#33889[N])
- ahli sejarah (LI#33890[N])

中文
- 历史学家 (LI#309859[N])
- 史家 (LI#312115[N])

3

#2195

English
- consumer (LI#139587[N])

Bahasa Melayu
- konsumer (LI#16603[N])
- pengguna (LI#16602[N])

中文
- 消费者 (LI#355727[N])

3

#8062

English
- plaintive (LI#230503[A])

Bahasa Melayu
- sedih (LI#9455[A])
- sedih (LI#9947[I])

中文
- 哀伤 (LI#314455[A])

3

#8887

English
- reassure (LI#242030[V])

Bahasa Melayu
- meyakinkan semula (LI#56022[V])

中文
- 再保证 (LI#303578[V])

3

#7987

English
- pike (LI#229692[N])

Bahasa Melayu
- ikan pike (LI#51614[N])

中文
- 矛 (LI#366598[N])

0

#9618

English
- saving (LI#249295[N])

Bahasa Melayu
- simpanan (LI#18590[N])

中文
- 存款 (LI#323322[N])

3

#8437

English
- private (LI#235341[A])
- personal (LI#227005[A])

Bahasa Melayu
- pangkat terendah di dalam askar (LI#53750[N])
- bukan milik awam (LI#53751[A])
- swasta (LI#53753[A])
- persendirian (LI#36591[A])
- persendirian (LI#53749[N])
- sendirian (LI#53752[A])
- sulit (LI#14356[A])

中文
- 私人 (LI#368643[A])

2

#11730

English
- tropic (LI#276682[N])

Bahasa Melayu
- kawasan tropika (LI#67493[N])
- khatulistiwa (LI#25521[N])

中文
- 回归线 (LI#316322[A])

0

#10493

English
- spray (LI#259871[N])

Bahasa Melayu
- percikan (LI#62308[N])
- tangkai (LI#18175[N])

中文
- 水花 (LI#352274[N])
- 喷雾器 (LI#315640[N])

1

#12155

English
- variant (LI#282347[N])

Bahasa Melayu
- varian (LI#69605[N])

中文
- 异体 (LI#330684[N])

3

#8855

English
- RAW (LI#241641[N])

Bahasa Melayu
- melecet (LI#55892[A])

中文
- 生肉 (LI#362829[N])

0

#5406

English
- increase (LI#186161[V])
- augment (LI#116930[I])
- augment (LI#116931[V])

Bahasa Melayu
- bertambah (LI#763[V])
- menaikkan (LI#25702[V])
- meningkat (LI#33785[V])
- menaikkan (LI#100151[I])
- menokok (LI#4743[V])

中文
- 增加 (LI#318727[V])

3

#9750

English
- seascape (LI#250933[N])

Bahasa Melayu
- pandangan laut (LI#59290[N])

中文
- 海景 (LI#355465[N])

3

#4297

English
- freighter (LI#166503[N])

Bahasa Melayu
- kapal pengangkut (LI#29408[N])
- penghantar muatan (LI#29409[N])

中文
- 货船 (LI#386831[N])
- 货轮 (LI#386833[N])

3

#7857

English
- permutation (LI#226874[N])

Bahasa Melayu
- permustasi (LI#50696[N])
- pilih atur (LI#50698[N])

中文
- 排列 (LI#339215[N])

3

#1359

English
- bunker (LI#127739[N])

Bahasa Melayu
- tempat simpanan bahan bakar (LI#11042[N])
- lubang pasir (LI#11043[N])
- bunker (LI#11044[N])
- kubu bawah tanah (LI#11045[N])

中文
- 暗堡 (LI#345181[N])

2

#2088

English
- concurrent (LI#138850[A])

Bahasa Melayu
- sama (LI#16031[N])

中文
- 一致 (LI#290356[A])

3

#2112

English
- confidence (LI#139019[N])

Bahasa Melayu
- keyakinan (LI#4608[N])

中文
- 自信 (LI#377404[N])

3

#9297

English
- revolution (LI#245378[N])
- revolutionary (LI#245381[A])

Bahasa Melayu
- revolusi (LI#57360[N])

中文
- 革命 (LI#399110[N])

3

#4726

English
- Field (LI#162815[A])
- ground (LI#173266[N])

Bahasa Melayu
- kawasan (LI#5196[N])

中文
- 场地 (LI#317637[N])

3

#11982

English
- unknown (LI#280026[N])

Bahasa Melayu
- anu (LI#68835[A])

中文
- 未知数 (LI#346479[N])

3

#11483

English
- today (LI#273243[N])

Bahasa Melayu
- sekarang (LI#47352[N])
- sekarang (LI#47354[ADV])
- sekarang (LI#101481[I])

中文
- 当今 (LI#331439[N])
- 当今 (LI#331440[A])

3

#3249

English
- duplicate (LI#153049[N])

Bahasa Melayu
- salinan (LI#17012[N])

中文
- 复本 (LI#319130[N])

3

#6808

English
- miss (LI#207989[V])

Bahasa Melayu
- tak mengena (LI#44787[V])
- tidak faham (LI#44789[V])
- tidak kena (LI#44788[V])
- terlepas (LI#24440[V])

中文
- 遗漏 (LI#392257[V])

2

#7762

English
- peck (LI#225308[V])

Bahasa Melayu
- memagut (LI#50116[V])
- mematuk (LI#50117[V])
- mencium dengan cepat (LI#50118[V])

中文
- 啄 (LI#315140[V])

3

#9405

English
- roe (LI#246672[N])

Bahasa Melayu
- kijang betina (LI#57763[N])
- telur ikan (LI#57764[N])

中文
- 鱼子 (LI#402457[N])

2

#8339

English
- preference (LI#234128[N])
- predilection (LI#234023[N])

Bahasa Melayu
- yang lebih diutamakan (LI#53376[N])
- kecenderungan (LI#4955[N])
- keutamaan (LI#4084[N])

中文
- 偏爱 (LI#300847[N])

3

#1412

English
- cage (LI#128836[N])

Bahasa Melayu
- peti huruf (LI#11498[N])
- kurungan (LI#10148[N])
- kandang (LI#11497[N])

中文
- 笼 (LI#370429[N])
- 槛 (LI#349859[N])

2

#9513

English
- sacrament (LI#248114[N])

Bahasa Melayu
- sakramen (LI#58224[N])

中文
- 圣事 (LI#317190[N])
- 圣礼 (LI#317277[N])

3

#3243

English
- dune (LI#152968[N])

Bahasa Melayu
- dun (LI#23441[N])
- gumuk (LI#23443[N])
- bukit pasir (LI#23442[N])

中文
- 沙丘 (LI#353132[N])

3

#8211

English
- bridge (LI#126565[N])
- pons (LI#232710[N])

Bahasa Melayu
- jambatan (LI#10471[N])

中文
- 桥 (LI#349087[N])

3

#1203

English
- boundless (LI#125698[A])

Bahasa Melayu
- tak bertepi (LI#10022[A])
- tak terbatas (LI#10024[A])
- tak terkawal (LI#10023[A])
- tak terkawal (LI#96717[I])

中文
- 无边无际 (LI#343588[N])

2

#5437

English
- individual (LI#186479[A])
- individual (LI#186480[N])

Bahasa Melayu
- perorangan (LI#36592[A])
- individu (LI#36593[A])
- persendirian (LI#36591[A])
- tersendiri (LI#36590[A])
- sendiri (LI#36589[A])

中文
- 个体 (LI#293524[N])

3

#2932

English
- dilate (LI#149011[V])

Bahasa Melayu
- mengembang (LI#21375[V])

中文
- 扩大 (LI#336312[V])
- 膨胀 (LI#377294[V])

3

#10089

English

Bahasa Melayu

中文

- silverfish (LI#255207[N])
- gegat (LI#45471[N])
- 蠹虫 (LI#382281[N])

3

#84

English
- accumulator (LI#105568[N])

Bahasa Melayu
- anak bateri (LI#792[N])
- akumulator (LI#793[N])
- penumpuk (LI#794[N])
- penimbun (LI#796[N])
- pengumpul (LI#795[N])

中文
- 蓄电池 (LI#380809[N])

3

#12549

English
- pier (LI#229564[N])
- quay (LI#239905[N])
- wharf (LI#285948[N])

Bahasa Melayu
- dermaga (LI#22551[N])

中文
- 码头 (LI#367001[N])

3

#5185

English
- hyacinth (LI#180892[N])

Bahasa Melayu
- kemeling (LI#34817[N])
- keladi bunting (LI#34819[N])

中文
- 风信子 (LI#400126[N])

3

#11813

English
- tusk (LI#277613[N])

Bahasa Melayu
- gading (LI#38721[N])

中文
- 牙 (LI#360126[N])

3

#9702

English
- scowl (LI#250370[V])

Bahasa Melayu
- bermasam muka (LI#59047[V])
- bermasam muka (LI#98631[I])
- meradang (LI#3946[V])

中文
- 皱眉头 (LI#365056[V])

3

#3999

English
- fifteenth (LI#162875[A])

Bahasa Melayu

中文
- 第十五 (LI#370402[N])

0

#7327

English
- obverse (LI#217039[N])
- obverse (LI#217040[A])

Bahasa Melayu
- depan (LI#28906[N])
- bahagian depan (LI#47666[N])
- bahagian hadapan (LI#47667[N])
- muka syiling gambar kepala orang (LI#47668[N])

中文
- 正面 (LI#350696[N])

2

#409

English
- anaesthetist (LI#110874[N])
- anesthetist (LI#111315[N])

Bahasa Melayu
- pakar bius (LI#3488[N])

中文
- 麻醉师 (LI#403278[N])

3

#11624

English
- ephemeral (LI#157183[A])
- transient (LI#274670[A])

Bahasa Melayu
- seketika (LI#25291[A])

中文
- 短暂 (LI#366726[A])

3

#9624

English
- scab (LI#249419[N])

Bahasa Melayu
- kuping kudis (LI#58681[N])
- keruping (LI#58679[N])
- kuping (LI#58680[N])

中文
- 痂 (LI#363987[N])

3

#372

English
- ampoule (LI#110539[N])

Bahasa Melayu
- ampul (LI#3394[N])
- bekas kecil pengisi cecair suntikan (LI#3395[N])

中文
- 安瓿 (LI#324012[N])

3

#11735

English
- trouble (LI#276736[N])

Bahasa Melayu
- kekacauan.kk.t./i (LI#67520[N])
- kesulitan (LI#32704[N])
- kesulitan (LI#99620[I])
- kesukaran (LI#8730[N])
- kesukaran (LI#99619[I])

中文
- 麻烦 (LI#403247[N])
- 辛苦 (LI#389817[N])

2

#11064

English
- swab (LI#265601[V])

Bahasa Melayu
- mengesat (LI#18839[V])
- menyapu (LI#11263[V])

中文
- 拭抹 (LI#338134[V])
- 擦洗 (LI#340787[V])

3

#5097

English
- horsemanship (LI#180174[N])

Bahasa Melayu
- keahlian menunggang (LI#34446[N])

中文
- 马术 (LI#401207[N])

3

#10376

English
- sparkle (LI#258507[V])

Bahasa Melayu
- menunjukkan kepintaran atau kecergasan (LI#61994[V])
- mengerlap (LI#60319[V])
- bersinar (LI#30909[V])
- mengerlip (LI#30936[V])
- bercahaya (LI#28480[V])
- berkilau (LI#28197[V])

中文
- 闪耀 (LI#395782[V])

2

#9207

English
- reprieve (LI#244172[N])

Bahasa Melayu
- penangguhan (LI#1459[N])

中文
- 缓刑 (LI#373781[N])

3

- respite (LI#244622[N])

#8347

English	Bahasa Melayu	中文	
• premature (LI#234332[A])	• pramasa (LI#53433[A]) • pramatang (LI#53434[A]) • tak dijangka (LI#53435[A]) • belum tiba waktunya (LI#53437[A]) • lebih awal (LI#53436[A])	• 过早 (LI#390239[A])	2

#3844

English	Bahasa Melayu	中文	
• effortless (LI#154305[A]) • facile (LI#160745[A])	• senang (LI#16825[A]) • senang (LI#97630[I])	• 容易 (LI#325007[A])	3

#3628

English	Bahasa Melayu	中文	
• alcohol (LI#108578[N]) • ethanol (LI#158474[N])	• etanol (LI#2465[N])	• 酒精 (LI#393083[N]) • 乙醇 (LI#295083[N])	1

#4410

English	Bahasa Melayu	中文	
• foreman (LI#165656[N]) • gaffer (LI#167662[N]) • ganger (LI#168128[N])	• mandur (LI#28957[N])	• 领班 (LI#399868[N])	2

#12183

English	Bahasa Melayu	中文	
• grassland (LI#172750[N]) • veld (LI#282684[N])	• padang rumput (LI#31613[N])	• 草原 (LI#379535[A])	3

#5040

English	Bahasa Melayu	中文	
• hobby (LI#179081[N])	• hobi (LI#33949[N])	• 嗜好 (LI#315700[N])	3

#582

English	Bahasa Melayu	中文	
• arduous (LI#114085[A])	• perlu ketekunan (LI#5191[A]) • payah (LI#5188[A]) • sukar (LI#5189[A]) • sukar (LI#97628[I])	• 险峻 (LI#397264[N])	3

#2085

English	Bahasa Melayu	中文	
• concoct (LI#138810[V])	• mereka (LI#15997[V]) • membancuh (LI#15998[V])	• 编造 (LI#373870[V])	2

#13250

English	Bahasa Melayu	中文	
• set out (LI#253040[I])	• bertolak (LI#101032[I])	• 动身 (LI#307160[V])	3

#1785

English	Bahasa Melayu	中文	
• circuit (LI#135341[N]) • circuitry (LI#135349[N])	• peredaran (LI#14165[N]) • peredaran (LI#14169[A])	• 电路 (LI#363378[N])	3

#7521

English	Bahasa Melayu	中文	
• outstanding (LI#221364[A])	• terkemuka (LI#33753[A])	• 杰出 (LI#347479[A])	3

#2842

English	Bahasa Melayu	中文	
• detective (LI#147573[N]) • detective (LI#147574[A])	• detektif (LI#20693[N])	• 侦探 (LI#299576[N])	3

#11873

English	Bahasa Melayu	中文	
• unarmed (LI#278520[A])	• tak bersenjata (LI#68114[A])	• 未武装 (LI#346474[A]) • 徒手 (LI#331968[A])	3

#7604

English	Bahasa Melayu	中文	
• palm (LI#223040[N])	• telapak tangan (LI#49279[N]) • pokok palma (LI#49280[N]) • pokok kelapa (LI#49281[N]) • tapak tangan (LI#49278[N])	• 手掌 (LI#335754[N])	2

#2051

English	Bahasa Melayu	中文	
• composition (LI#138562[N])	• komposisi (LI#15841[N])	• 作文 (LI#299176[N])	0

#3717

English	Bahasa Melayu	中文	
• exempt (LI#159701[A])	• terkecuali (LI#26217[A])	• 免除 (LI#301999[A])	3

#6077

English	Bahasa Melayu	中文	
• legend (LI#195709[N])	• legenda (LI#40567[N]) • dongeng sejarah (LI#40568[N]) • dongengan (LI#26885[N])	• 传说 (LI#298394[N])	3

#11752

English	Bahasa Melayu	中文

- bugle (LI#127506[N])
- bugle (LI#127507[I])
- trump (LI#276846[N])
- trumpet (LI#276853[N])

- trompet (LI#10899[N])

- 喇叭 (LI#315418[N])

3

#7716

English
- pastoral (LI#224743[N])
- pastoral (LI#224744[A])

Bahasa Melayu
- pastoral (LI#49876[A])
- berkenaan hidup di desa (LI#49879[A])
- kedesaan (LI#49878[A])

中文
- 牧歌 (LI#360323[N])
- 牧人 (LI#360314[N])

2

#64

English
- acceptance (LI#105400[N])

Bahasa Melayu
- persetujuan (LI#597[N])

中文
- 认可 (LI#384590[N])

3

#178

English
- adventurer (LI#107123[N])

Bahasa Melayu
- pekelana (LI#1676[N])
- petualang (LI#1673[N])
- pengembara (LI#1677[N])

中文
- 冒险家 (LI#303663[N])

3

#8126

English
- plight (LI#231363[V])

Bahasa Melayu
- berjanji (LI#29110[V])

中文
- 约定 (LI#372256[V])

3

#11688

English
- trend (LI#275147[N])

Bahasa Melayu
- arah alir (LI#67145[N])
- trend (LI#67146[N])
- hala (LI#5725[N])
- haluan (LI#10056[N])
- aliran (LI#16085[N])

中文
- 趋势 (LI#388111[N])
- 倾向 (LI#300601[N])

2

#9802

English
- seize (LI#251491[V])

Bahasa Melayu
- menyambar (LI#31444[V])
- menawan (LI#2832[V])

中文
- 抓住 (LI#336748[V])

3

#9893

English
- sesame (LI#252892[N])

Bahasa Melayu
- bijan (LI#59994[N])

中文
- 芝麻 (LI#378544[N])

3

#12059

English
- uplift (LI#280965[N])

Bahasa Melayu
- meninggikan (LI#69357[N])

中文
- 高扬 (LI#401932[N])

3

#8196

English
- polyhedral (LI#232393[A])
- polyhedron (LI#232394[N])

Bahasa Melayu
- polihedron (LI#52654[A])
- polihedron (LI#52655[N])

中文
- 多面体 (LI#319730[N])

3

#7615

English
- panda (LI#223253[N])

Bahasa Melayu
- beruang panda (LI#49346[N])

中文
- 熊猫 (LI#359542[N])

3

#3816

English
- alien (LI#108812[A])
- extraneous (LI#160447[A])

Bahasa Melayu
- asing (LI#2555[A])

中文
- 外来 (LI#319353[A])

3

#5231

English
- icicle (LI#183211[N])

Bahasa Melayu
- isikel (LI#35387[N])
- jurai air batu (LI#35388[N])
- tiruk ais (LI#35389[N])

中文
- 冰柱 (LI#304028[N])

3

#3482

English
- endocrine (LI#156096[N])
- endocrine (LI#156095[A])

Bahasa Melayu
- endokrinologi (LI#24841[N])

中文
- 内分泌腺 (LI#303393[N])
- 内分泌 (LI#303392[N])

3

#12162

English
- vary (LI#282436[V])

Bahasa Melayu
- berubah (LI#49806[V])

中文
- 变化 (LI#311314[V])

3

#9076

English
- commemorate (LI#138126[V])
- remember (LI#243665[V])

Bahasa Melayu
- mengenang (LI#56723[V])
- memperingati (LI#15521[V])
- ingat (LI#7857[V])
- ingat (LI#96900[I])
- mengingat (LI#25133[V])

中文
- 纪念 (LI#372337[V])
- 记得 (LI#384788[V])

3

#7110

English
- careless (LI#130455[A])
- negligent (LI#212598[A])

Bahasa Melayu
- cuai (LI#12274[A])

中文
- 粗心 (LI#371201[A])

3

#3231

English
- duke (LI#152891[N])

Bahasa Melayu
- duke (LI#23383[N])
- bangsawan Inggeris (LI#23384[N])

中文
- 公爵 (LI#302667[A])

3

#7917

English	Bahasa Melayu	中文	
• penis (LI#225788[N]) • phallus (LI#227460[N])	• zakar (LI#50290[N])	• 阴茎 (LI#396444[N])	3

#12150

English	Bahasa Melayu	中文	
• vantage (LI#282269[N])	• faedah (LI#1650[N])	• 优势 (LI#298096[N])	3

#3813

English	Bahasa Melayu	中文	
• extra (LI#160397[A]) • excess (LI#159508[A])	• pelakon tambahan (LI#26705[N]) • ekstra (LI#26708[A]) • lebih (LI#26707[A]) • lebihan (LI#1330[N]) • tambahan (LI#630[N]) • tambahan (LI#1314[A]) • tambahan (LI#96353[I])	• 外加 (LI#319274[A]) • 额外 (LI#400028[A])	3

#10002

English	Bahasa Melayu	中文	
• shortcoming (LI#254195[N]) • failing (LI#160887[N])	• kesalahan (LI#25666[N]) • kekurangan (LI#19181[N]) • kekurangan (LI#101270[I])	• 缺点 (LI#374060[N])	3

#12060

English	Bahasa Melayu	中文	
• Top (LI#273617[A]) • topmost (LI#273701[A]) • uppermost (LI#281009[A])	• tertinggi (LI#13563[A])	• 最高 (LI#345661[A])	3

#5597

English	Bahasa Melayu	中文	
• integrative (LI#187760[A])	• integrasi (LI#37447[A])	• 一体化 (LI#290011[N])	3

#4182

English	Bahasa Melayu	中文	
• foot (LI#165227[V])	• membayar (LI#12502[V]) • membayar (LI#97146[I])	• 结算 (LI#373117[V])	3

#1561

English	Bahasa Melayu	中文	
• cashew (LI#130969[N])	• gajus (LI#12504[N]) • janggus (LI#12505[N]) • ketereh (LI#12506[N])	• 腰果 (LI#377163[N])	3

#4130

English	Bahasa Melayu	中文	
• flunk (LI#164671[V])	• menggagalkan (LI#248[V])	• 失败 (LI#321266[V])	3

#11787

English	Bahasa Melayu	中文	
• tun (LI#277227[N])	• tong (LI#7517[N])	• 大桶 (LI#320294[N])	3

#3611

English	Bahasa Melayu	中文	
• espionage (LI#158283[N])	• espionaj (LI#25756[N]) • perisikan (LI#25754[N]) • penyuluhan (LI#25755[N]) • pengintipan (LI#25753[N])	• 间谍活动 (LI#395942[N])	3

#7847

English	Bahasa Melayu	中文	
• peripheral (LI#226635[A])	• periferi (LI#50605[N])	• 周边 (LI#314014[N])	3

#4524

English	Bahasa Melayu	中文	
• gill (LI#169976[N])	• gil (LI#30747[N]) • insang (LI#10229[N]) • sesuku (LI#30746[N]) • pial (LI#30748[N]) • seperempat (LI#29231[N])	• 峡谷 (LI#327482[N])	0

#2023

English	Bahasa Melayu	中文	
• compensation (LI#138423[N])	• kompensasi (LI#15728[N]) • ganti rugi (LI#15727[N]) • pampasan (LI#15729[N]) • imbuhan (LI#1899[N])	• 赔偿金 (LI#387398[N])	2

#2223

English	Bahasa Melayu	中文	
• contraception (LI#139765[N])	• pencegahan hamil (LI#16715[N]) • kontrasepsi (LI#16716[N])	• 避孕 (LI#392390[N])	3

#4588

English	Bahasa Melayu	中文	
• glycoprotein (LI#171206[I])	• glikoprotein (LI#31138[N])	• 糖蛋白 (LI#371530[N])	3

#2200

English	Bahasa Melayu	中文	
• contain (LI#139623[V])	• mengawal (LI#16621[V]) • mengawal (LI#100334[I]) • menahan (LI#4862[V])	• 遏制 (LI#392110[V])	3

APPENDIX G

VECTOR COSINE SIMILARITY FOR WORDSIM-353 WORD PAIRS

Word 1	Word 2	Human score	CSim score with different no. of factors					
			300	400	500	600	700	800
love	sex	6.77	1.36	0.88	0.78	1.08	0.98	0.93
tiger	cat	7.35	6.22	6.58	4.43	2.43	0.44	0.28
tiger	tiger	10.00	10.00	10.00	10.00	10.00	10.00	10.00
book	paper	7.46	4.17	3.77	3.04	2.81	2.41	2.03
computer	keyboard	7.62	5.08	5.27	4.15	3.53	3.27	2.70
computer	internet	7.58	1.73	1.53	1.68	1.90	1.87	1.94
plane	car	5.77	0.45	0.69	0.57	0.63	0.54	0.54
train	car	6.31	2.06	1.72	1.25	0.95	0.75	0.71
telephone	communication	7.50	6.91	6.91	6.56	6.23	5.80	5.33
television	radio	6.77	3.81	2.07	1.52	1.32	1.38	1.21
media	radio	7.42	4.48	3.90	2.84	2.48	2.46	2.44
drug	abuse	6.85	4.96	4.51	4.34	3.78	3.99	3.53
bread	butter	6.19	9.07	8.34	7.82	7.72	7.37	7.10
cucumber	potato	5.92	7.85	6.47	5.51	4.66	4.10	3.59
doctor	nurse	7.00	6.76	6.12	5.32	4.97	4.81	5.19
professor	doctor	6.62	6.35	5.49	4.58	4.24	4.02	3.75
student	professor	6.81	6.72	6.31	5.81	5.24	4.44	3.87
smart	student	4.62	1.95	1.43	1.21	1.07	0.73	0.29
smart	stupid	5.81	3.29	2.45	1.48	1.20	1.17	1.17
company	stock	7.08	6.83	5.39	4.60	3.92	3.28	3.06
stock	market	8.08	7.80	7.47	6.94	6.63	6.14	5.63
stock	phone	1.62	0.17	0.08	0.06	0.23	0.17	0.24
stock	jaguar	0.92	0.44	0.17	0.01	0.03	0.19	0.01
stock	egg	1.81	0.29	0.34	0.40	0.16	0.07	0.16
fertility	egg	6.69	1.45	1.39	0.97	0.80	0.96	1.30
stock	live	3.73	0.51	0.54	0.07	0.05	0.14	0.06
stock	life	0.92	0.58	0.30	0.54	0.41	0.04	0.27
book	library	7.46	3.74	4.15	4.29	4.39	4.56	4.73
bank	money	8.12	5.96	5.70	5.19	4.97	4.63	4.22
wood	forest	7.73	4.44	3.64	3.63	3.68	3.64	3.27
money	cash	9.15	9.06	8.88	8.51	8.14	7.76	7.49
professor	cucumber	0.31	0.06	0.14	0.12	0.16	0.08	0.06
king	cabbage	0.23	0.04	0.01	0.01	0.00	0.05	0.03
king	queen	8.58	5.69	3.79	2.98	2.70	2.23	1.96
king	rook	5.92	2.81	2.64	2.33	2.07	1.79	1.63
bishop	rabbi	6.69	0.06	0.21	0.38	0.49	0.43	0.48
Jerusalem	Israel	8.46	7.06	6.04	4.61	3.88	3.49	3.24
Jerusalem	Palestinian	7.65	5.40	4.65	3.48	3.30	2.19	1.54
holy	sex	1.62	0.50	0.55	0.36	0.32	0.17	0.05
fuck	sex	9.44	4.28	3.81	3.21	3.05	2.71	2.22
football	soccer	9.03	8.64	8.19	8.07	7.98	7.83	7.74
football	basketball	6.81	4.95	4.73	4.56	4.55	4.53	4.45
football	tennis	6.63	2.60	1.53	1.35	1.67	1.86	2.07

Word 1	Word 2	Human score	CSim score with different no. of factors					
			300	400	500	600	700	800
tennis	racket	7.56	6.30	5.77	5.31	5.02	4.52	4.35
Arafat	peace	6.73	4.92	3.92	3.64	3.07	2.58	2.04
Arafat	terror	7.65	3.78	2.89	2.52	2.05	1.96	1.77
Arafat	Jackson	2.50	0.10	0.01	0.06	0.02	0.27	0.16
law	lawyer	8.38	6.87	6.98	6.15	5.37	4.81	4.42
movie	star	7.38	3.41	2.23	2.17	2.07	2.02	1.83
movie	popcorn	6.19	1.34	1.72	1.79	1.46	1.24	1.30
movie	critic	6.73	5.24	4.31	3.85	3.69	3.50	3.18
movie	theater	7.92	4.57	3.76	3.39	3.05	2.86	2.55
physics	proton	8.12	6.48	4.73	3.56	2.76	2.15	1.27
physics	chemistry	7.35	2.54	2.63	3.03	3.32	3.42	3.50
space	chemistry	4.88	0.49	0.17	0.16	0.18	0.17	0.07
alcohol	chemistry	5.54	3.37	3.63	2.82	2.45	2.22	2.03
vodka	gin	8.46	0.26	0.12	0.15	0.65	0.96	0.42
vodka	brandy	8.13	1.76	0.91	1.04	1.00	0.85	0.51
drink	car	3.04	0.57	0.50	0.11	0.04	0.13	0.11
drink	ear	1.31	0.71	1.22	0.90	0.52	0.24	0.06
drink	mouth	5.96	1.44	0.71	0.08	0.45	0.56	0.49
drink	eat	6.87	7.28	6.40	5.26	3.55	3.19	2.80
baby	mother	7.85	7.52	7.06	6.68	6.04	5.55	5.28
drink	mother	2.65	1.58	1.37	0.70	0.48	0.46	0.50
car	automobile	8.94	8.13	8.13	7.93	7.68	7.42	7.28
gem	jewel	8.96	2.24	1.64	1.34	1.27	1.26	0.99
journey	voyage	9.29	3.92	3.73	3.88	3.54	3.12	2.90
boy	lad	8.83	2.95	2.80	1.84	1.73	1.39	1.09
coast	shore	9.10	6.77	6.35	6.01	5.48	4.71	4.39
asylum	madhouse	8.87	0.65	0.39	0.46	0.49	0.26	0.06
magician	wizard	9.02	3.01	1.76	0.99	0.97	1.08	1.16
midday	noon	9.29	5.55	5.97	6.28	6.47	6.63	6.37
furnace	stove	8.79	2.02	2.47	2.11	1.87	1.60	1.06
food	fruit	7.52	2.42	2.42	1.93	1.40	1.06	0.78
bird	cock	7.10	1.03	1.03	1.09	1.21	1.22	0.89
bird	crane	7.38	5.60	4.53	3.96	3.71	3.88	4.21
tool	implement	6.46	6.10	4.85	4.30	3.93	3.72	3.23
brother	monk	6.27	1.37	0.60	0.30	0.23	0.34	0.39
crane	implement	2.69	0.31	0.12	0.13	0.19	0.14	0.25
lad	brother	4.46	1.88	1.07	0.92	0.95	1.13	0.99
journey	car	5.85	1.50	0.81	0.47	0.35	0.34	0.40
monk	oracle	5.00	0.98	0.78	0.33	0.37	0.61	0.51
cemetery	woodland	2.08	1.57	1.13	0.45	0.43	0.17	0.10
food	rooster	4.42	3.36	1.16	0.29	0.44	0.29	0.27
coast	hill	4.38	1.51	0.40	0.81	0.51	0.50	0.55
forest	graveyard	1.85	1.24	1.12	0.48	0.35	0.39	0.11
shore	woodland	3.08	0.44	0.10	0.06	0.08	0.09	0.03
monk	slave	0.92	0.24	0.42	0.34	0.77	0.30	0.46
coast	forest	3.15	0.23	0.01	0.07	0.26	0.14	0.27
lad	wizard	0.92	0.31	0.18	0.27	0.37	0.41	0.43
chord	smile	0.54	0.70	0.01	0.30	0.65	0.72	0.61
glass	magician	2.08	1.02	0.31	0.50	0.26	0.32	0.16
noon	string	0.54	0.41	0.01	0.05	0.05	0.34	0.22
rooster	voyage	0.62	0.07	0.59	0.39	0.44	0.34	0.24
money	dollar	8.42	7.23	6.45	5.90	5.51	5.05	4.47
money	cash	9.08	9.06	8.88	8.51	8.14	7.76	7.49

Word 1	Word 2	Human score	CSim score with different no. of factors					
			300	400	500	600	700	800
money	currency	9.04	6.77	6.19	5.90	5.63	5.28	5.05
money	wealth	8.27	4.71	4.87	4.43	4.03	3.89	3.73
money	property	7.57	1.63	1.75	1.69	1.67	1.71	1.58
money	possession	7.29	1.25	1.18	1.17	1.35	1.27	1.28
money	bank	8.50	5.96	5.70	5.19	4.97	4.63	4.22
money	deposit	7.73	3.69	3.59	3.46	3.40	3.53	3.60
money	withdrawal	6.88	1.69	1.54	1.24	1.29	1.37	1.28
money	laundering	5.65	7.26	6.88	6.49	6.22	5.90	6.04
money	operation	3.31	0.92	0.74	0.66	0.64	0.53	0.42
tiger	jaguar	8.00	6.17	6.50	6.42	4.12	2.78	2.26
tiger	feline	8.00	5.69	6.23	5.39	3.76	1.86	1.63
tiger	carnivore	7.08	5.40	5.07	2.44	2.04	1.97	1.86
tiger	mammal	6.85	6.15	3.44	0.23	0.22	0.09	0.23
tiger	animal	7.00	6.13	5.48	1.99	0.46	0.36	0.35
tiger	organism	4.77	0.41	0.17	0.47	0.10	0.16	0.17
tiger	fauna	5.62	3.48	1.87	1.33	0.53	0.33	0.26
tiger	zoo	5.87	7.28	6.95	5.95	4.79	4.43	4.12
psychology	psychiatry	8.08	5.49	5.45	5.10	4.85	4.26	3.75
psychology	anxiety	7.00	4.07	3.92	3.82	3.64	3.58	3.21
psychology	fear	6.85	2.24	2.20	2.22	2.03	2.16	1.95
psychology	depression	7.42	2.08	1.68	1.69	1.68	1.77	1.47
psychology	clinic	6.58	2.21	1.39	1.07	0.66	0.28	0.12
psychology	doctor	6.42	3.53	2.71	2.27	1.98	1.33	0.92
psychology	Freud	8.21	7.93	8.06	7.94	7.77	7.60	7.24
psychology	mind	7.69	7.00	7.01	6.61	6.14	5.62	4.90
psychology	health	7.23	4.23	3.60	3.14	2.45	1.57	1.16
psychology	science	6.71	7.76	7.41	6.44	5.76	5.00	4.53
psychology	discipline	5.58	8.49	8.13	7.29	6.43	5.65	5.28
psychology	cognition	7.48	8.36	8.42	8.16	8.13	8.11	7.88
planet	star	8.45	6.67	4.63	3.69	2.57	2.29	1.93
planet	constellation	8.06	4.04	1.00	0.12	0.38	0.37	0.47
planet	moon	8.08	8.10	5.79	2.72	1.22	1.01	0.92
planet	sun	8.02	8.74	7.78	6.02	4.22	3.13	2.07
planet	galaxy	8.11	7.33	3.99	2.41	0.77	1.05	0.84
planet	space	7.92	3.04	1.22	0.85	0.37	0.34	0.18
planet	astronomer	7.94	7.67	6.89	6.03	5.39	5.06	4.82
precedent	example	5.85	1.79	1.22	0.97	0.94	0.82	0.74
precedent	information	3.85	0.64	0.02	0.00	0.14	0.13	0.11
precedent	cognition	2.81	0.41	0.31	0.23	0.01	0.07	0.00
precedent	law	6.65	9.44	9.25	8.69	8.02	7.31	6.99
precedent	collection	2.50	0.47	0.16	0.40	0.44	0.60	0.52
precedent	group	1.77	0.31	0.38	0.26	0.08	0.13	0.03
precedent	antecedent	6.04	2.29	1.91	1.89	1.90	1.80	1.63
cup	coffee	6.58	0.59	0.73	0.82	0.84	0.81	0.98
cup	article	2.40	0.00	0.02	0.06	0.07	0.11	0.11
cup	artifact	2.92	0.14	0.01	0.36	0.37	0.14	0.00
cup	object	3.69	0.11	0.06	0.04	0.18	0.04	0.01
cup	entity	2.15	0.10	0.04	0.08	0.08	0.07	0.06
cup	drink	7.25	0.34	0.33	0.39	0.39	0.41	0.41
cup	food	5.00	0.18	0.00	0.11	0.05	0.05	0.06
cup	substance	1.92	0.07	0.06	0.12	0.15	0.11	0.01
cup	liquid	5.90	0.18	0.23	0.13	0.11	0.18	0.13
jaguar	cat	7.42	4.94	5.85	5.27	5.85	5.95	5.65

Word 1	Word 2	Human score	CSim score with different no. of factors					
			300	400	500	600	700	800
jaguar	car	7.27	3.24	2.42	1.68	1.82	1.54	1.67
energy	secretary	1.81	0.52	0.20	0.51	0.61	0.38	0.25
secretary	senate	5.06	4.58	3.36	2.87	2.44	2.33	2.29
energy	laboratory	5.09	2.93	2.21	1.64	1.22	0.62	0.41
computer	laboratory	6.78	1.53	1.52	1.53	1.36	1.03	0.96
weapon	secret	6.06	1.87	1.72	1.43	1.64	1.72	1.57
FBI	fingerprint	6.94	4.94	4.74	4.42	3.71	3.45	3.33
FBI	investigation	8.31	7.53	7.33	7.09	6.71	6.26	6.06
investigation	effort	4.59	3.82	2.33	2.11	1.84	1.81	1.63
Mars	water	2.94	1.11	0.63	0.85	0.75	0.73	0.65
Mars	scientist	5.63	4.05	3.22	2.70	2.12	2.19	2.25
news	report	8.16	3.81	4.06	3.81	3.55	3.28	3.07
canyon	landscape	7.53	4.43	3.76	3.06	2.73	2.49	2.09
image	surface	4.56	1.51	1.05	0.92	0.90	0.94	0.85
discovery	space	6.34	4.38	3.47	2.51	2.13	1.85	1.38
sign	recess	2.38	0.64	0.23	0.06	0.25	0.02	0.01
Wednesday	news	2.22	1.03	0.84	0.91	0.80	0.88	0.67
mile	kilometer	8.66	2.75	2.63	3.36	2.50	1.87	1.32
computer	news	4.47	0.80	0.15	0.12	0.02	0.26	0.34
territory	surface	5.34	0.39	0.58	0.25	0.24	0.11	0.01
atmosphere	landscape	3.69	1.63	1.22	0.96	0.60	0.57	0.67
president	medal	3.00	0.25	0.19	0.21	0.14	0.17	0.14
war	troops	8.13	5.85	5.02	3.77	3.28	2.97	2.71
record	number	6.31	0.99	0.85	0.72	0.58	0.46	0.42
skin	eye	6.22	6.02	5.31	3.84	3.14	2.23	1.48
Japanese	American	6.50	1.66	1.46	1.40	1.34	1.27	1.22
theater	history	3.91	1.64	1.46	1.15	0.98	1.25	1.04
volunteer	motto	2.56	1.87	1.27	1.01	0.90	1.01	0.76
prejudice	recognition	3.00	4.25	3.98	3.54	2.90	2.50	2.36
decoration	valor	5.63	1.33	1.35	1.73	1.52	1.34	1.18
century	year	7.59	1.69	1.16	0.90	0.75	0.66	0.63
century	nation	3.16	0.62	0.52	0.44	0.36	0.35	0.30
delay	racism	1.19	0.19	0.09	0.01	0.08	0.18	0.28
delay	news	3.31	1.93	1.71	0.97	0.86	0.75	0.68
minister	party	6.63	1.12	1.04	1.05	1.01	1.00	0.98
peace	plan	4.75	1.91	1.87	1.59	1.16	0.61	0.48
minority	peace	3.69	0.75	0.41	0.05	0.02	0.17	0.08
attempt	peace	4.25	5.08	4.12	3.41	2.74	2.17	1.71
government	crisis	6.56	6.05	5.18	4.94	4.64	4.34	4.10
deployment	departure	4.25	0.24	0.31	0.24	0.28	0.15	0.15
deployment	withdrawal	5.88	1.91	1.64	1.86	1.90	1.63	1.64
energy	crisis	5.94	0.99	0.56	0.49	0.46	0.19	0.30
announcement	news	7.56	5.45	4.91	4.49	3.85	3.17	2.87
announcement	effort	2.75	3.67	3.21	2.88	2.61	2.33	2.27
stroke	hospital	7.03	1.76	1.16	1.22	0.95	0.92	0.25
disability	death	5.47	1.93	1.50	0.92	0.09	0.04	0.02
victim	emergency	6.47	6.26	5.43	4.58	3.78	3.11	2.58
treatment	recovery	7.91	3.78	2.92	2.62	2.25	1.99	1.96
journal	association	4.97	3.06	2.63	2.53	2.03	1.47	1.27
doctor	personnel	5.00	1.00	1.36	0.93	0.87	0.61	0.35
doctor	liability	5.19	0.05	0.04	0.04	0.06	0.30	0.12
liability	insurance	7.03	8.32	7.75	7.52	7.16	6.93	6.78
school	center	3.44	0.83	0.80	0.65	0.54	0.49	0.50

Word 1	Word 2	Human score	CSim score with different no. of factors					
			300	400	500	600	700	800
reason	hypertension	2.31	1.51	1.44	1.01	0.89	0.84	0.47
reason	criterion	5.91	4.82	4.09	3.70	3.27	3.04	2.86
hundred	percent	7.38	2.76	2.32	1.98	1.77	1.46	1.28
Harvard	Yale	8.13	5.55	5.02	4.10	3.72	3.08	2.81
hospital	infrastructure	4.63	2.52	2.47	2.41	2.13	1.92	1.29
death	row	5.25	0.29	0.17	0.29	0.26	0.53	0.26
death	inmate	5.03	4.41	3.92	3.63	2.69	1.64	1.61
lawyer	evidence	6.69	3.36	3.33	2.87	2.39	2.02	1.84
life	death	7.88	5.11	4.48	4.17	4.12	3.63	3.30
life	term	4.50	2.22	1.00	0.75	0.62	0.75	0.58
word	similarity	4.75	5.89	3.70	2.60	2.22	1.94	1.92
board	recommendation	4.47	3.62	3.27	3.32	3.03	2.84	2.67
governor	interview	3.25	0.00	0.23	0.19	0.41	0.15	0.29
OPEC	country	5.63	4.61	2.98	2.40	1.64	1.24	1.14
peace	atmosphere	3.69	0.23	0.37	0.09	0.10	0.21	0.22
peace	insurance	2.94	0.49	0.60	0.45	0.59	0.54	0.30
territory	kilometer	5.28	1.81	1.37	0.72	0.78	0.99	0.89
travel	activity	5.00	2.31	1.84	1.63	1.85	1.59	1.43
competition	price	6.44	1.27	1.50	1.66	1.42	1.19	1.18
consumer	confidence	4.13	1.84	1.83	1.21	0.77	0.82	0.78
consumer	energy	4.75	0.89	0.22	0.19	0.43	0.50	0.44
problem	airport	2.38	0.13	0.11	0.08	0.06	0.04	0.04
car	flight	4.94	0.13	0.22	0.24	0.03	0.01	0.12
credit	card	8.06	2.38	2.32	2.39	2.86	2.98	2.79
credit	information	5.31	2.25	2.01	1.86	1.58	1.41	1.19
hotel	reservation	8.03	1.01	0.61	0.70	0.55	0.12	0.10
grocery	money	5.94	1.56	1.02	0.82	0.66	0.40	0.48
registration	arrangement	6.00	0.15	0.05	0.12	0.12	0.13	0.15
arrangement	accommodation	5.41	0.17	0.10	0.01	0.21	0.14	0.11
month	hotel	1.81	0.43	0.25	0.26	0.18	0.15	0.25
type	kind	8.97	4.59	4.14	3.62	3.38	3.09	2.95
arrival	hotel	6.00	0.99	0.47	0.36	0.62	0.67	0.38
closet	clothes	8.00	4.39	4.04	2.87	2.37	1.96	1.53
situation	conclusion	4.81	6.39	5.78	5.42	4.89	4.51	4.38
situation	isolation	3.88	3.54	3.24	2.92	2.45	2.24	2.11
impartiality	interest	5.16	1.41	1.31	1.10	0.56	0.71	0.88
direction	combination	2.25	3.48	2.90	2.33	1.89	1.66	1.47
street	place	6.44	1.71	1.84	1.52	1.38	1.15	0.84
street	avenue	8.88	7.72	7.71	7.40	7.24	6.83	6.78
street	block	6.88	2.22	1.86	1.91	1.74	1.40	1.28
street	children	4.94	1.24	0.42	0.45	0.70	0.17	0.24
listing	proximity	2.56	0.02	0.16	0.02	0.13	0.12	0.02
listing	category	6.38	0.34	0.13	0.08	0.05	0.18	0.28
cell	phone	7.81	0.27	0.51	0.58	0.53	0.52	0.46
production	hike	1.75	0.18	0.53	0.30	0.08	0.09	0.15
benchmark	index	4.25	6.69	5.44	5.24	4.81	4.71	4.63
media	trading	3.88	0.33	0.23	0.48	0.46	0.66	0.69
media	gain	2.88	2.65	2.14	2.02	1.91	1.76	1.62
dividend	payment	7.63	7.13	5.91	5.49	4.87	4.54	4.53
dividend	calculation	6.48	1.63	1.54	1.52	1.36	1.65	1.75
calculation	computation	8.44	7.08	6.71	6.52	6.42	6.15	5.85
currency	market	7.50	1.93	1.79	1.60	1.53	1.33	1.18
OPEC	oil	8.59	7.71	8.50	9.19	9.34	9.32	9.24

Word 1	Word 2	Human score	CSim score with different no. of factors					
			300	400	500	600	700	800
oil	stock	6.34	2.90	0.86	0.09	0.15	0.37	0.39
announcement	production	3.38	1.00	1.07	0.92	0.88	0.68	0.70
announcement	warning	6.00	2.33	1.49	1.22	1.05	0.96	0.92
profit	warning	3.88	0.49	0.20	0.07	0.19	0.19	0.15
profit	loss	7.63	3.28	3.53	3.38	3.27	3.11	2.91
dollar	yen	7.78	6.47	6.18	6.24	6.00	5.69	5.53
dollar	buck	9.22	1.51	1.26	1.39	1.43	1.33	1.24
dollar	profit	7.38	2.66	2.42	1.76	1.52	1.39	1.19
dollar	loss	6.09	1.06	0.96	0.57	0.64	0.64	0.50
computer	software	8.50	7.70	5.53	3.86	2.87	2.41	2.30
network	hardware	8.31	1.14	0.56	0.85	0.79	0.73	0.65
phone	equipment	7.13	0.04	0.30	0.31	0.44	0.35	0.38
equipment	maker	5.91	2.07	1.39	1.14	1.11	1.06	1.03
luxury	car	6.47	4.98	4.37	3.94	3.77	3.42	3.28
report	gain	3.63	1.14	0.71	0.29	0.05	0.16	0.17
investor	earning	7.13	3.51	2.40	1.88	1.21	1.05	1.02
liquid	water	7.89	6.22	4.80	4.13	3.68	3.37	3.09
baseball	season	5.97	0.14	0.15	0.19	0.23	0.20	0.17
game	victory	7.03	0.88	0.90	0.86	0.87	0.86	0.79
game	team	7.69	1.21	0.78	0.82	0.81	0.79	0.77
marathon	sprint	7.47	7.34	6.80	6.46	6.16	5.99	5.76
game	series	6.19	1.85	1.52	1.29	1.20	1.20	1.11
game	defeat	6.97	0.01	0.03	0.05	0.21	0.23	0.22
seven	series	3.56	4.16	3.41	2.66	2.24	2.14	1.77
seafood	sea	7.47	2.13	1.89	1.70	1.54	1.32	1.24
seafood	food	8.34	7.40	6.15	4.83	4.30	3.63	3.28
seafood	lobster	8.70	8.12	7.03	6.22	5.70	5.49	5.26
lobster	food	7.81	7.16	5.81	4.65	3.48	2.80	2.40
lobster	wine	5.70	0.05	0.11	0.02	0.01	0.06	0.03
food	preparation	6.22	6.77	6.28	5.54	5.01	4.49	3.99
video	archive	6.34	1.77	1.58	1.07	0.82	0.46	0.30
start	year	4.06	4.77	4.31	4.32	4.03	3.71	3.44
start	match	4.47	2.05	1.79	1.58	1.42	1.29	1.19
game	round	5.97	0.02	0.24	0.23	0.22	0.22	0.18
boxing	round	7.61	0.52	0.23	0.48	0.44	0.16	0.06
championship	tournament	8.36	4.00	4.26	3.69	3.56	3.52	3.33
line	insurance	2.69	0.34	0.38	0.18	0.02	0.19	0.15
day	summer	3.94	2.39	1.36	1.04	0.93	0.90	0.84
summer	drought	7.16	3.33	2.75	2.59	2.20	2.11	2.04
summer	nature	5.63	0.91	0.95	0.78	0.83	0.68	0.62
day	dawn	7.53	4.68	4.09	2.66	2.48	2.33	2.28
nature	environment	8.31	5.47	5.26	4.76	4.49	4.02	3.68
environment	ecology	8.81	5.86	5.90	5.70	5.77	5.81	5.71
nature	man	6.25	1.75	1.66	1.34	1.27	1.28	1.01
man	woman	8.30	8.10	8.02	7.74	7.33	6.52	5.62
man	governor	5.25	1.00	0.97	0.78	0.80	0.91	0.89
murder	manslaughter	8.53	6.18	5.35	5.15	5.08	5.16	5.49
soap	opera	7.94	4.02	3.71	3.34	3.14	3.15	3.15
opera	performance	6.88	5.66	5.16	4.75	4.47	3.89	3.64
life	lesson	5.94	3.86	3.47	3.29	3.03	2.67	2.59
focus	life	4.06	3.89	3.45	2.67	2.53	2.36	1.93
production	crew	6.25	0.74	0.41	0.47	0.30	0.43	0.60
television	film	7.72	0.08	0.03	0.09	0.11	0.10	0.10

Word 1	Word 2	Human score	CSim score with different no. of factors					
			300	400	500	600	700	800
lover	quarrel	6.19	3.27	2.53	2.11	1.92	1.64	1.49
viewer	serial	2.97	2.51	1.86	1.23	1.05	1.15	1.14
possibility	girl	1.94	0.21	0.50	0.04	0.04	0.42	0.30
population	development	3.75	0.34	0.29	0.20	0.15	0.09	0.11
morality	importance	3.31	3.85	2.87	2.59	2.32	2.16	2.04
morality	marriage	3.69	2.28	1.48	1.24	1.05	0.73	0.61
Mexico	Brazil	7.44	4.18	1.06	0.31	0.15	0.07	0.00
gender	equality	6.41	4.13	4.10	4.11	3.84	3.79	3.57
change	attitude	5.44	3.19	3.17	2.60	2.36	2.14	2.06
family	planning	6.25	0.34	0.26	0.40	0.38	0.31	0.30
opera	industry	2.63	0.07	0.05	0.01	0.01	0.02	0.02
sugar	approach	0.88	0.83	0.49	0.11	0.15	0.39	0.23
practice	institution	3.19	3.36	3.18	2.82	2.24	1.88	1.67
ministry	culture	4.69	0.70	0.79	0.72	0.75	0.77	0.54
problem	challenge	6.75	4.08	3.77	3.34	3.05	2.87	2.45
size	prominence	5.31	1.37	1.39	0.99	0.65	0.56	0.35
country	citizen	7.31	4.50	4.25	3.62	3.42	3.14	2.99
planet	people	5.75	0.33	0.09	0.08	0.04	0.05	0.06
development	issue	3.97	0.84	1.09	0.94	0.96	1.15	1.09
experience	music	3.47	1.47	1.04	0.79	0.67	0.54	0.46
music	project	3.63	0.14	0.12	0.00	0.11	0.10	0.04
glass	metal	5.56	4.38	3.96	3.74	3.28	3.11	2.91
aluminum	metal	7.83	4.90	4.65	4.38	4.23	3.89	3.48
chance	credibility	3.88	2.23	2.08	1.76	1.34	1.26	1.18
exhibit	memorabilia	5.31	3.75	2.92	2.51	2.25	2.29	2.20
concert	virtuoso	6.81	3.44	3.10	2.21	1.89	1.65	1.72
rock	jazz	7.59	4.05	2.94	2.24	1.28	1.27	1.12
museum	theater	7.19	3.90	2.99	2.11	1.26	1.01	0.81
observation	architecture	4.38	1.47	0.85	0.60	0.32	0.19	0.07
space	world	6.53	0.59	0.38	0.62	0.61	0.64	0.43
preservation	world	6.19	2.37	2.00	1.33	0.99	0.92	0.91
admission	ticket	7.69	0.64	1.07	1.39	1.32	1.20	1.26
shower	thunderstorm	6.31	6.51	5.91	4.64	4.18	3.95	3.90
shower	flood	6.03	3.09	2.88	2.48	2.06	1.54	1.12
weather	forecast	8.34	6.60	6.55	6.58	6.37	6.31	6.21
disaster	area	6.25	2.02	1.70	1.46	1.27	1.17	1.09
governor	office	6.34	2.79	2.96	2.89	2.68	2.26	2.26
architecture	century	3.78	2.41	1.70	1.56	1.35	1.29	1.29

APPENDIX H

CONTEXT-DEPENDENT LEXICAL LOOKUP RESULTS

Test sentence	Rank of correct translation set using strategy		
	wiki-lsi	base-freq	goog-tr
The reprocessing **plant** was apparently operating, whilst the delegation were there.	1	2	1
Plant species from around the world.	2	1	1
Honda were losing over 1,000 workers from their manufacturing **plant** in yokohama.	1	2	1
We have four of them next to the veg plots and we use them to grow tender **plants** or to extend the growing season.	1	1	1
Defective viral dna ameliorates symptoms of geminivirus infection in transgenic **plants**.	1	1	1
Power **plants** range from 49cc 2 stroke, to 110cc four stroke engines.	1	2	1
Each time, he sprayed the sunflower crossbreeds and backcrossed the most robust specimens with the original cultivated parent **plants**.	2	1	1
Aston, b.c. (1923) the poisonous, suspected and medicinal **plants** of new zealand.	1	2	1
Plant pathology topics shown at scientific meetings in recent years.	2	1	1
At newton abbott, a recent contract called for the removal of a pen stock wall in a sewage treatment **plant**.	1	2	1
The uk government should seek to support replacement of magnox by new nuclear **plant**.	1	2	1
Plant breeders, and no need for lawyers.	1	1	1
These were intended to put in place the key components of seed supply such as processing **plants**, stores and quality control facilities.	1	2	1
The areas surrounding the water have been allowed to develop naturally with trees, shrubbery and wild **plants**.	1	1	1
This time of year is a perfect time to be lifting and splitting herbaceous **plants** before they put on too much vegetative growth.	2	1	1
A taiwan police officer tried to rob a **bank** with a toy gun.	3	1	1
The swiss national bank is another central **bank** to have greatly reduced its dollar ratio in recent years.	3	1	1
Transfer the money to our **bank** account via bank account via bank transfer from your bank account or by visiting any branch of barclays **bank**.	1	1	1
A few of the old mill cottages still survive but the ones on the river **bank** have long gone.	1	2	1
To get back home was a task in itself up a nice steep **bank**.	2	3	3
Bank holiday is already planned to mark the queen's golden jubilee.	3	1	1

Test sentence	Rank of correct translation set using strategy		
	wiki-lsi	base-freq	goog-tr
Please note that the college only accepts checks drawn on a uk clearing **bank**, sterling drafts or postal orders.	1	1	1
Situated on south **bank** very near the dam wall.	1	2	1
Bank statement showing our final balance.	1	1	1
He works for an investment **bank** in london at canary wharf.	1	1	1
Cross to the north **bank** at shaw's bridge.	1	2	1
Bank robbers and psychotics like hamilton will take no notice.	3	1	1
Thumby planned for his family a fairly large three-storey home to be some 200 feet from the east **bank** of the river.	1	2	1
Bank balance at 31 december 1999 was £3.0 million.	3	1	1
The river seems to have flowed close to the shore but was separated from it by great shingle **banks**.	2	3	–
Bank loans in shanghai were in real estate.	3	1	1
High street **banks** say many students have to supplement their state loans with commercial ones in order to meet their living costs.	3	1	1
Kabinet Kerajaan akan bermesyuarat pada bulan hadapan bagi membincangkan pelbagai isu terkini.	1	2	1
almari **kabinet** ini dihiasi ukiran cantik.	1	1	1
Dapur lebih kemas dengan gaya moden, nampak menawan dengan **kabinet** 3G Arkrilik gaya baru.	2	1	1
Saidin berdiri mengambil fail-fail dari dalam **kabinet**	2	1	1
Draf pindaan Akta DBP 1959 sedang disemak dan dijangka dibawa ke **kabinet** sebelum dibentangkan di parlimen pada tahun depan.	1	2	1
Pakar dalam membuat **kabinet** dapur dan almari .	1	1	1
Kami menjual pelbagai barangan perabot seperti meja, **kabinet**, almari, katil	1	1	1
Kabinet Dapur dan perabot terus dari kilang	2	1	1
Saya ada menjual perabot rumah terpakai seperti, Meja makan, Almari dapur, **kabinet** memasak bersama dapur memasak	1	1	1
BAGI ruang sempit, penggunaan rak atau **kabinet** amat penting kerana ia menjimatkan dan menampilkan suasana kemas, bersih serta menarik.	2	1	1
Kabinet kayu dengan rak tv	1	1	1
Kabinet telah bersetuju untuk memberi kuasa kepada DBP untuk mendenda pencemar bahasa di negara ini.	1	2	1
Kabinet memutuskan untuk menangguhkan pelaksanaan sistem Saraan Baru Perkhidmatan Awam	1	2	1
jawatankuasa **Kabinet** itu telah ditafsirkan oleh jawatankuasa Pegawai-Pegawai seperti berikut	1	2	1
Laman Web Rasmi Bahagian **Kabinet**, Perlembagaan dan Perhubungan Antara Kerajaan.	1	2	1
Mangga merupakan satu genus tumbuhan yang terdiri daripada 35 spesies pokok buah tropika dalam famili Anacardiaceae.	1	1	1
Mangga ini sangat sesuai ditanam di kawasan utara Semenanjung Malaysia.	2	1	1

Test sentence	Rank of correct translation set using strategy		
	wiki-lsi	base-freq	goog-tr
Mangga boleh dibuka dengan anak kunci	2	2	–
Mangga atau ibu kunci adalah sejenis kunci yang mudah alih untuk melindungi sesuatu daripada dicuri, laku musnah, sabotaj, pengintipan, penggunaan tanpa kebenaran dan rosak.	1	2	–
Mangga (Mangifera indica) adalah buah tropika	1	1	1
Tanaman **Mangga** bisa tumbuh dengan baik di daerah dataran rendah dan berhawa panas.	2	1	1
Satu buah **mangga** mengandung tujuh gram serat yang dapat membantu sistem pencernaan.	1	1	1
Ribuan **mangga** kunci yang diletakkan pada pagar sebuah jambatan kereta api di Rhine, Cologne kelmarin.	1	2	2
kunci keselamatan, memotong **mangga**, perkhidmatan kunci pintu dan banyak lagi	2	2	4
Mangga Pagar Sentiasa Rosak Akibat Terdedah Kepada Hujan	1	2	–
kerabu **Mangga** ala Thai Mudah dibuat dan sesuai untuk pembuka selera,	1	1	1
Selain menggunakan tambahan peralatan keselamatan seperti **mangga** kunci dan rantai besi	1	2	1
Semalam ita buat sambal pelam **mangga** muda.	2	1	1
河流纵**谷**里的平凡与幸福	1	2	–
亚玛力人和迦南人住在**谷**中，明天你们要转回，从红海的路往旷野去。	2	2	1
这两天的价钱是糙米六块,**谷**三块。	1	1	–
这里的土壤肥沃，种出来的**谷**又大又美。	1	1	–
大河流过**谷**底，水流很急。	1	2	1
当余之从师也,负箧曳屣,行深山巨**谷**中。	2	2	1
在**谷**底发现一名妇女的尸体。	1	2	–
他说他不得不和**谷**商做生意	1	1	1
下一季度，现有**谷**租不会改变。	1	1	–
过去，**谷**底被用来种小麦。	2	2	–
那一段时间，**谷**里非常热闹，不比现在这样冷冷清清的!	1	2	1
据科学家们的调查，该**谷**中发现的各种死于非命的飞禽走兽、大小动物的尸骸已超过4000只	2	2	1
多吃**谷**蔬限红肉可控Ⅱ型糖尿病	1	1	–
牛耕田，马吃**谷**	1	1	–
Udah ujan nya ngetu terbubuh, matahari enggau **emperaja** lalu ayan ba langit.	1	1	–
Lelaki nya tikah enggau **emperaja** iya, siko dayang ke ligung.	1	2	–
Sida nurun niti **emperaja** lalu terengkah ba kampung alaisida betemu enggau.	2	1	–
Iya ngena baju burung ke bechura baka **emperaja**	1	1	–
Pengerindu aku ti tuchi enggau **emperaja** enda nemu luya	1	2	–
Terima kasih ngagai **emperaja** laban udah ngibun serta nyaga aku selama tok	2	2	–

REFERENCES

[1] Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., & Soroa, A. (2009). A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT'09)* (pp. 19–27). Boulder, Colorado.

[2] Agirre, E., & Rigau, G. (1996). Word Sense Disambiguation Using Conceptual Density. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 96).* Copenhagen, Denmark.

[3] Al-Adhaileh, M. H., Tang, E. K., & Zaharin, Y. (2002). A synchronization structure of SSTC and its applications in machine translation. In *Proceedings of COLING-2002 Workshop "Machine translation in Asia".* Taipei, Taiwan.

[4] Banerjee, S., & Pedersen, T. (2003). Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence* (pp. 805–810).

[5] Berment, V. (2004). *Méthods pour informatiser les langues et les groupes de langues 'peu dotées'* (Unpublished doctoral dissertation). Université Joseph Fourier, Grenoble, France.

[6] Boitet, C. (2007, January 15). *Terminology relative to NLP for (African) "pi-languages" and "pi-pairs", towards more oral systems.* [Electronic mailing list message]. Retrieved from `http://www.mail-archive.com/mt-list@eamt.org/msg01004.html` (25 May 2009)

[7] Boitet, C., Mangeot, M., & Sérasset, G. (2002). The PAPILLON project: Cooperatively building a multilingual lexical database to derive open source dictionaries & lexicons. In *Proceedings of the 2nd Workshop on NLP and XML (NLPXML'02)* (pp. 1–3). Taipei, Taiwan.

[8] Boitet, C., & Zaharin, Y. (1988, 8). Representation trees and string-tree correspondences. In *Proceedings of the 12th International Conference on Computational Linguistics* (Vol. 1, pp. 59–64). Budapest, Hungary.

[9] Boitet, C., Zaharin, Y., & Tang, E. K. (2011). Learning-to-translate based on the S-SSTC annotation schema. In *Proceedings of 25th Pacific Asia Conference on Language, Information and Computation (PACLIC 2011).* Singapore.

[10] Bond, F., & Ogura, K. (2008). Combining linguistic resources to create a machine-tractable Japanese–Malay dictionary. *Language Resources and Evaluation*, *42*, 127–136.

[11] Bond, F., & Paik, K. (2012). A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)* (pp. 64–71). Matsue, Japan.

[12] Bond, F., Ruhaida, b. S., Yamazaki, T., & Ogura, K. (2001). Design and construction of a machine-tractable Japanese-Malay dictionary. In *Proceedings of MT Summit VIII* (pp. 53–58). Santiago de Compostela, Spain.

[13] Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., & Mercer, R. L. (1991). Word-sense disambiguation using statistical methods. In *Proceedings of the 29th Annual Meeting on Association for Computational Linguistics* (pp. 264–270). Berkeley, California.

[14] Bugoslavsky, I. (2005). Some controversial issues of UNL: Linguistic aspects. *Research on Computing Science*, *12*, 77–100.

[15] Cardeñosa, J., Gelbukh, A., & Tovar, E. (Eds.). (2005). Universal Networking Language: Advances in theory and applications. *Research on Computing Science*, *12*.

[16] Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, *31*(5), 489–509.

[17] Dagan, I., & Itai, A. (1994). Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, *20*(4), 563–596.

[18] Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407.

[19] de Melo, G., & Weikum, G. (2009). Towards a universal wordnet by learning from combined evidence. In *Proceedings of the 18th acm conference on information and knowledge management (cikm 2009)* (pp. 513–522). New York, NY, USA: ACM. doi: http://doi.acm.org/10.1145/1645953.1646020

[20] Diab, M., & Finch, S. (2000). A statistical word-level translation model for comparable corpora. In *Proceedings of the Conference on Content-Based Multimedia Information Access (RIAO).*

[21] Dong, Z., & Dong, Q. (2006). *HowNet and the computation of meaning*. Singapore: World Scientific.

[22] Dorow, B., Laws, F., Michelbacher, L., Scheible, C., & Utt, J. (2009). A graph-theoretic algorithm for automatic extension of translation lexicons. In *Proceedings of the EACL 2009 Workshop on*

GEMS: GEometical Models of Natural Language Semantics (pp. 91–95). Athens, Greece.

[23] Dumais, S. T., Littman, M. L., & Landauer, T. K. (1997). Automatic cross-language retrieval using latent semantic indexing. In *AAAI97 Spring Symposium Series: Cross Language Text and Speech Retrieval* (pp. 18–24). Stanford University.

[24] Edmonds, P., & Hirst, G. (2002). Near-synonymy and lexical choice. *Computational Linguistics*, *28*(2), 105–144.

[25] Ethnologue. (2012, April). *Ethnologue report for language code: iba.* Retrieved from `http://www.ethnologue.com/show_language.asp?code=iba` (12 September 2012)

[26] Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database*. Cambridge, Massachusetts: MIT Press.

[27] Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, *20*(1), 116–141.

[28] Fontenelle, T. (Ed.). (2003). *International Journal of Lexicography: Special issue on FrameNet and frame semantics* (Vol. 16) (No. 3).

[29] Forcada, M. L. (2009, February 1). *Cheeseburgery hamburgers and the problem of computerised translations.* [Online forum comment]. Retrieved from `http://blogs.ft.com/brusselsblog/2009/01/cheeseburgery-hamburgers-and-the-problem-of-computerised-translations#comment-226812` (6 March 2009)

[30] Francopoulo, G., Bel, N., George, M., Calzolari, N., Monachini, M., Pet, M., & Soria, C. (2009). Multilingual resources for NLP in the lexical markup framework (LMF). *Language Resources and Evaluation*, *43*(1), 57–70. doi: 10.1007/s10579-008-9077-5

[31] Fung, P., & Lo, Y. Y. (1998). An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL 98)* (pp. 414–420). Montreal, Quebec, Canada.

[32] Harris, Z. (1954). Distributional structure. *Word*, *10*(23), 146–162.

[33] Hartmann, R., & Stork, F. (1972). *Dictionary of language and linguistics*. London: Applied Science.

[34] Hong, K. S., Tan, T. P., & Tang, E. K. (2012). Using dependency parse tree structure level and type information to improve Malay large vocabulary automatic speech recognition systems. In *Proceedings of the 6th International Workshop of Malay and Indonesian Language Engineering (MALINDO 2012).* Kota Samarahan, Malaysia.

[35] Hovy, E. (1999). Toward finely differentiated evaluation metrics for machine translation. In *Proceedings of the EAGLES Workshop on Standards and Evaluation.* Pisa, Italy.

[36] Hurford, J. R. (1990). Nativist and functional explanations in language acquisition. In I. M. Roca (Ed.), *Logical issues in language acquisition.* Dordrecht, the Netherlands: Foris Publications.

[37] Hutchins, J. (1999, 6). The development and use of machine translation systems and computer-based translation tools. In *Proceedings of the International Symposium on Machine Translation and Computer Language Information Processing* (pp. 1–16). Beijing, China.

[38] Hutchins, J. (2007). *Machine translation: problems and issues.* Presentation. (Chelyabinsk, Russia. 18 slides.)

[39] Ide, N., Erjavec, T., & Tufiş, D. (2002). Sense discrimination with parallel corpora. In *Proceedings of the SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions* (pp. 54–60). Philadelphia, USA.

[40] Ide, N., & Véronis, J. (1998). Word Sense Disambiguation: The State of the Art. *Computational Linguistics*, *24*(1), 1–41.

[41] Inkpen, D., & Hirst, G. (2006). Building and using a lexical knowledge base of near-synonymy differences. *Computational Linguistics*, *32*(2), 223–262.

[42] International Organization for Standardization. (2008). *ISO 24613:2008 Language resource management – Lexical Markup Framework (LMF).*

[43] Jalabert, F., & Lafourcade, M. (2002, 7). From sense naming to vocabulary augmentation in Papillon. In *Proceedings of PAPILLON-2003 Workshop.* Sapporo, Japan.

[44] Janssen, M. (2003, 10). Lexical translation and conceptual hierarchies. In *Proceedings of the 5th International Symposium on Language, Logic and Computation.* Tbilisi, Georgia.

[45] Janssen, M. (2004). Multilingual lexical databases, lexical gaps, and SIMuLLDA. *International Journal of Lexicography*, *17*, 136–154.

[46] Janssen, M., Verkuyl, H., & Jansen, F. (2003). The codification of usage by labels. In van Sterkenburg (Ed.), *A practical guide to lexicography.* Amsterdam: John Benjamins Publishing Company.

[47] Johns, A. H. (Ed.). (2000). *Kamus Inggeris Melayu Dewan: An English-Malay Dictionary*. Kuala Lumpur, Malaysia: Dewan Bahasa dan Pustaka.

[48] Kilgariff, A. (1996). *BNC database and word frequency lists.* Retrieved from `http://www.kilgarriff.co.uk/bnc-readme.html`

[49] Kipper, K., Korhonen, A., Ryant, N., & Palmer, M. (2008). A large-scale classification of English verbs. *Language Resources and Evaluation Journal*, *42*(1), 21–40.

[50] Lafourcade, M. (2002, 8). Automatically populating acception lexical database through bilingual dictionaries and conceptual vectors. In *Proceedings of PAPILLON-2002.* Tokyo, Japan.

[51] Leacock, C., & Chodorow, M. (1998). Combining local context and wordnet similarity for word sense identification. In C. Felbaum (Ed.), *Wordnet: An electronic lexical database* (pp. 265–283). Cambridge, Massachusetts, USA: MIT Press.

[52] Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the ACM-SIGDOC conference* (pp. 24–26). Toronto, Canada.

[53] Li, H., & Li, C. (2004). Word translation disambiguation using bilingual bootstrapping. *Computational Linguistics*, *30*(1), 1–22. doi: http://dx.doi.org/10.1162/089120104773633367

[54] Lim, L. T., Ranaivo-Malançon, B., & Tang, E. K. (2011a). Low cost construction of a multilingual lexicon from bilingual lists. *Polibits*, *43*, 45–51.

[55] Lim, L. T., Ranaivo-Malançon, B., & Tang, E. K. (2011b). Symbiosis between a multilingual lexicon and translation example banks. *Procedia: Social and Behavioral Sciences*, *27*, 61–69. (Tier 4; Scopus SNIP impact factor: 0.162.)

[56] Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the International Conference on Machine Learning.* Madison, Wisconsin.

[57] Magnini, B., & Cavaglià, G. (2000, 5). Integrating subject field codes into WordNet. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000)* (pp. 1413–1418). Athens, Greece.

[58] Magnini, B., Strapparava, C., Pezzulo, G., & Gliozzo, A. (2001). Using domain information for word sense disambiguation. In *Proceedings of the 2nd International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2)* (pp. 111–114). Toulouse, France.

[59] Markó, K., Schulz, S., & Hahn, U. (2005). Multilingual lexical acquisition by bootstrapping cognate seed lexicons. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP) 2005.* Borovets, Bulgaria.

[60] Mausam, Soderland, S., Etzioni, O., Weld, D., Skinner, M., & Bilmes, J. (2009). Compiling a massive, multilingual dictionary via probabilistic inference. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (pp. 262–270). Suntec, Singapore.

[61] McCarthy, D. (2011). Measuring similarity of word meaning in context with lexical substitutes and translations. In *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2011)* (Vol. 6608, pp. 238–252). Tokyo, Japan: Springer.

[62] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography (special issue)*, *3*(4), 235–312.

[63] Ng, H. T., Wang, B., & Chan, Y. S. (2003). Exploiting parallel texts for word sense disambiguation: An empirical study. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (pp. 455–462). Sapporo, Japan.

[64] Nguyen, H.-T., Boitet, C., & Sérasset, G. (2007). PIVAX, an online contributive lexical database for heterogenous MT systems using a lexical pivot. In *Proceedings of the 7th International Symposium on Natural Language Processing (SNLP 2007).* Bangkok, Thailand.

[65] Niles, I., & Pease, A. (2001). Towards a standard upper ontology. In *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001).* Ogunquit, Maine.

[66] Niles, I., & Pease, A. (2003). Linking lexicons and ontologies: Mapping wordnet to the suggested upper merged ontology. In *Proceedings of the 2003 International Conference on Information and Knowledge Engineering (IKE '03).* Las Vegas, Nevada.

[67] O'Hara, T., Bruce, R., Donner, J., & Wiebe, J. (2004, July). Class-based collocations for word sense disambiguation. In *Proceedings of the 3rd International Workshop on the Evaluation of*

Systems for the Semantic Analysis of Text (SENSEVAL-3) (pp. 199–202). Barcelona, Spain.

[68] Otero, P. G., Campos, J. R. P., Ramom, J., Campos, P., & Compostela, S. D. (2005). An approach to acquire word translations from non-parallel texts. In *Proceedings of the 12th Portuguese Conference on Artificial Intelligence (EPIA-05)* (p. 600–610).

[69] Pease, A., Fellbaum, C., & Vossen, P. (2008). Building the Global WordNet Grid. In *Proceedings of the 18th International Congress of Linguists (CIL18).* Seoul, Republic of Korea.

[70] *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004).* (2004). Geneva, Switzerland.

[71] Rapp, R. (1999). Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics* (p. 519—526).

[72] Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, *11*, 95–130.

[73] Resnik, P., & Yarowsky, D. (1997). A perspective on word sense disambiguation methods and their evaluation. In *Proceedings of the SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?* (pp. 79–86). Washington, D.C., USA.

[74] Rymer, R. (2012, July). Vanishing Voices. *National Geographic*, *222*(1), 60–93.

[75] Sabrina, T., Rosni, A., & Tang, E. K. (2011). Subword unit concatenation for Malay speech synthesis. *International Journal of Computer Science Issues*, *8*(2), 68–74.

[76] Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing & Computational Linguistics (CICLING-2002)* (pp. 1–15). Mexico City, Mexico.

[77] Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11).

[78] Sammer, M., & Soderland, S. (2007). Building a sense-distinguished multilingual lexicon from monolingual corpora and bilingual lexicons. In *Proceedings of Machine Translation Summit XI* (pp. 399–406). Copenhagen, Denmark.

[79] Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, *24*(1), 97–123.

[80] Shi, L., & Mihalcea, R. (2005). Putting pieces together: Combining FrameNet, VerbNet and WordNet for robust semantic parsing. In *Proceedings of Computational Linguistics and Intelligent Text Processing (CICLing 2005)* (Vol. 3406). Mexico: Springer.

[81] Shirai, K., & Yagi, T. (2004). Learning a robust word sense disambiguation model using hypernyms in definition sentences. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)* (pp. 917–923). Geneva, Switzerland.

[82] Shirai, S., & Yamamoto, K. (2001). Linking English words in two bilingual dictionaries to generate another language pair dictionary. In *Proceedings of the 19th International Conference on Computer Processing of Oriental Languages (ICCPOL-2001)* (pp. 174–179S). Seoul, Korea.

[83] Simon, H. A. (1947). *Administrative behavior: a study of decision-making processes in administrative organization* (1st ed.). New York: Macmillan.

[84] Song, S., Cheah, Y.-N., Tang, E. K., & Ranaivo-Malançon, B. (2008). Rule extraction for automatic question answering based on structural clustering. *IJCSNS International Journal of Computer Science and Network Security*, *8*(3), 208–215.

[85] Stevenson, M., & Wilks, Y. (2001). The interaction of knowledge sources in word sense disambiguation. *Computational Linguistics*, *27*(3), 321–349. doi: http://dx.doi.org/10.1162/089120101317066104

[86] Sutlive, V., & Sutlive, J. (1992). *Handy reference dictionary of Iban and English*. Tun Jugah Foundation Association.

[87] Tanaka, K., Umemura, K., & Iwasaki, H. (1998). Construction of a bilingual dictionary intermediated by a third language. *Transactions of the Information Processing Society of Japan*, *39*(6), 1915–1924. (In Japanese)

[88] Tanner, A. (2007, March 28). *Google seeks world of instant translations.* Retrieved from `http://www.reuters.com/article/2007/03/28/us-google-translate-idUSN1921881520070328` (5 October 2012)

[89] Tufiş, D., Ion, R., & Ide, N. (2004). Fine-grained word sense disambiguation based on parallel corpora, word alignment, word clustering and aligned wordnets. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)* (pp. 1312–1318). Geneva,

Switzerland.

[90] Tufiş, D., Cristea, D., & Stamou, S. (2004). BalkaNet: Aims, methods, results and perspectives – a general overview. *Romanian Journal of Information Science and Technology Special Issue*, *7*(1), 9–43.

[91] Uchida, H., Zhu, M., & Senta, T. D. (2005). *Universal networking language*. UNDL Foundation.

[92] United Nations. (n.d.). *UN official languages.* Retrieved from `http://www.un.org/en/aboutun/languages.shtml` (5 October 2012)

[93] UNL Center. (2004). *The Universal Networking Language (UNL) specifications version 3 edition 3* (Specifications). UNL Foundation.

[94] Varga, I., Yokoyama, S., & Hashimoto, C. (2009). Dictionary generation for less-frequent language pairs using WordNet. *Literary and Linguistic Computing*, *24*(4), 449–466.

[95] Verma, N., & Bhattacharyya, P. (2003). Automatic generation of multilingual lexicon by using WordNet. In *Proceedings of International Conference on Convergence of Knowledge, Culture, Language and Information Technology.* Library of Alexandria, Egypt.

[96] Vossen, P. (1997). EuroWordNet: a multilingual database for information retrieval. In *In Proceedings of the DELOS Workshop on Cross-language Information Retrieval* (pp. 5–7).

[97] Vossen, P. (2004). EuroWordNet: A multilingual database of autonomous and language-specific wordnets connected via an Inter-Lingual-Index. *Special Issue on Multilingual Databases, International Journal of Linguistics*, *17*(2).

[98] W3Techs. (2012). *Usage statistics of content languages for websites, September 2012.* Retrieved from `http://w3techs.com/technologies/overview/content_language/all` (6 September 2012)

[99] Wiktionary. (2012, March 3). *Wiktionary:statistics.* Retrieved from `http://en.wiktionary.org/wiki/Wiktionary:Statistics` (3 March 2012)

[100] Wilks, Y., Fass, D., Guo, C.-M., McDonald, J., Plate, T., & Slator, B. (1993). Providing machine tractable dictionary tools. In J. Pustejovsky (Ed.), *Semantics and the lexicon* (pp. 341–401). Dordrecht, the Netherlands: Kluwer Academic Publishers.

[101] Wilks, Y., & Stevenson, M. (1998). Word sense disambiguation using optimised combinations

of knowledges sources. In *Proceedings of the 17th International Conference on Computational Linguistics* (pp. 1398–1402).

[102] Wu, Z., & Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics (COLING-94)* (pp. 133–138). Las Cruces, New Mexico.

[103] Yeo, A. W., Suhaila, S., & Wilfred, J. (2008). Preservation of Sarawak indigenous languages. *Multilingual*, *19*(8), 36–39.

[104] Zajac, R. (1996). Structuring a multilingual multipurpose lexical database using a simple interlingual approach. In *Proceedings of AMTA-96 Workshop on Interlinguas.* Montreal, Canada.

[105] Zhong, Z., & Ng, H. T. (2009). Word sense disambiguation for all words without hard labor. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI 2009)* (pp. 1616–1621). Pasadena, California.

[106] Zhou, M., Ding, Y., & Huang, C. (2001). Improviging translation selection with a new translation model trained by independent monolingual corpora. *Computational Linguistics and Chinese language Processing*, *6*(1), 1–26.

GLOSSARY

aligned corpus A collection of multilingual documents, in which texts in one language are paired with their translations in another language. The texts are often aligned at the sentence or sub-sentence level (phrase, word, etc).

Creative Commons Attribution-ShareAlike License (CC-BY-SA) A copyright license that allow the distribution of copyrighted works, allowing users to copy, distribute and transmit the work; to adapt the work; and to make commercial use of the work, with conditions of attribution and share-alike (`http://creativecommons.org/licenses/by-sa/3.0/us/`).

cognates Words that have a common etymological origin, e.g. English «shirt», «skirt», German «Schürze» and Dutch «schort» are all derived from Proto-Germanic **skurtjōn-*.

comparable corpus A collection of multilingual documents, in which texts in one language are paired with a similar text in another language, that need not be exact translations of each other. The texts are therefore not aligned (except at the document level).

crowdsourcing The act of outsourcing tasks, traditionally performed by an employee or contractor, to a large group of people or community (a crowd), through an open call, usually without substantial monetary compensation.

F_1 or F-measure A measure of a test's accuracy, a weighted average of the precision and recall.

GNU Free Documentation License (GFDL) A copyleft license for free documentation. It is similar to the GNU General Public License, giving readers the rights to copy, redistribute, and modify a work and requires all copies and derivatives to be available under the same license (`http://www.gnu.org/copyleft/fdl.html`).

homonym LIs having the same spelling but different meanings and origins, e.g. «bat» (a nocturnal, flying mammal) and «bat» (a club for hitting the ball in sports).

Inter-Lingual Index (ILI) The language-independent index for linking multilingual LIs in EuroWordNet.

interlingua A formal system describing the underlying semantics of natural language text, but independent of any real-world natural language.

information retrieval (IR) Area of study concerned with searching for documents, for information within documents, and for metadata about documents.

lexical gap The situation in which no single word exists in a language to denote a particular concept. Also known as *lacuna*.

lexical item (LI) A unit of the vocabulary of a language such as a word, phrase or term as listed in a dictionary. It usually has a pronounceable or graphic form, fulfils a grammatical role in a sentence, and carries semantic meaning. For convenience's sake, sometimes used interchangeably with 'word' in the early part of this thesis.

lexicalisation The process of making a word to express a concept.

Lexical Markup Framework (LMF) The ISO International Organization for Standardization ISO/TC37 standard for NLP and MRD lexicons (ISO 24613:2008).

latent semantic indexing (LSI) An indexing and retrieval method that uses a mathematical technique called SVD to identify patterns in the relationships between the terms and concepts contained in an unstructured collection of text.

macrostructure The organisation of the lexical entries in the body of a dictionary into lists, tree structures or networks.

microstructure The consistent organisation of lexical information in lexical dictionary entries.

machine-readable dictionary (MRD) A dictionary stored as machine (computer) data instead of being printed on paper, i.e. an electronic dictionary and lexical database. Used interchangably with 'lexicon' in this thesis.

mean reciprocal rank (MRR) A statistic measure for evaluating any process that produces a list of possible responses to a query, ordered by probability of correctness. It is the average of the reciprocal ranks of the correct response for a sample of queries.

machine translation (MT) The use of computers to translate from one human language to another.

multi-word expression (MWE) A lexical item that contains multiple word units.

neologism A newly coined term, word, or phrase, that may be in the process of entering common use, but has not yet been accepted into mainstream language.

natural language processing (NLP) The application of computational linguistics principles to problems.

one-time inverse consultation (OTIC) A procedure proposed by Tanaka et al. (1998) for generating a bilingual dictionary for a new language pair L_1–L_3 from bilingual dictionaries of existing language pairs L_1–L_2, L_2–L_3 and L_3–L_2, using L_2 as an intermediate language.

polyseme An LIs with different, but related senses, e.g. «man» can mean the human species; or an adult male of the human species..

part of speech (POS) The linguistic category of lexical items, e.g. noun, verb, adjective, etc. Also called word class, lexical class, or lexical category.

precision The number of true positives divided by the total number of elements labelled as belonging to the positive class.

recall The number of true positives divided by the total number of elements that actually belong to the positive class.

Synchronous SSTC (S-SSTC) A flexible annotation schema that declaratively captures correspondences between a pair of SSTCs.

sense In linguistics, one of the meanings of a word.

source language (SL) The original language of a text to be translated or to be looked up.

Structured String-Tree Correspondence (SSTC) A flexible annotation schema, especially suitable for capturing irregular correspondences between a string and its arbitrary tree representation.

SSTC+Lexicon (SSTC+L) A flexible annotation schema, based on the SSTC for associating (possibly discontiguous) phrases in a text to LIs entries from a lexicon.

Suggested Upper Merged Ontology (SUMO) A formal upper ontology intended as a foundation ontology for a variety of computer information processing systems (Niles & Pease, 2001).

singular value decomposition (SVD) A factorisation of a real or complex matrix $M = U\Sigma V^T$, where U is a $m \times r$ matrix; Σ is an $r \times r$ rectangular diagonal matrix, and V^T is an $r \times n$ matrix. The eigenvectors of MM^T make up the columns of U; the eigenvectors of $M^T M$ make up those of V; and the values on the diagonal of Σ are the square roots of the eigenvalues from MM^T or $M^T M$..

synset The basic organisation unit in a wordnet system, where all member LIs convey the same sense, e.g. (car, auto, automobile, machine, motorcar).

term-document matrix A mathematical matrix that describes the frequency of terms that occur in a collection of documents. In a document-term matrix, rows correspond to documents in the collection and columns correspond to terms.

term frequencey–inverse document frequency (TF-IDF) A numerical statistic which reflects how important a word is to a document in a collection or corpus.

target language (TL) The desired language of a text to be translated into or to be looked up.

translation equivalent A corresponding word or expression in another language.

translation selection A process to select the most appropriate translation word from a set of TL words corresponding to a SL word, reflecting its sense in a particular context.

translation set A multilingual entry in Lexicon+TX comprising lexical items from different languages, depicting the same concept or meaning of coarse granularity.

under-resourced language Human languages having limited NLP resources. Alternative terms include *π-languages*, *less-equipped language*.

Universal Networking Language (UNL) A declarative formal language specifically designed to represent semantic data extracted from natural language texts.

Universal Word (UW) Words of the UNL, constituting the UNL vocabulary. They are labels for concepts, syntactic and semantic units to form UNL Expressions.

wordnet Any lexical database using the same scheme as that of the Princeton WordNet (Miller et al., 1990).

word sense disambiguation (WSD) The problem of identifying which sense of a word is used in a sentence.

PUBLICATION LIST

Journal Articles

[1] Lim, L. T., Ranaivo-Malançon, B., & Tang, E. K. (2011a). Low cost construction of a multilingual lexicon from bilingual lists. *Polibits*, *43*, 45–51.

[2] Lim, L. T., Ranaivo-Malançon, B., & Tang, E. K. (2011b). Symbiosis between a multilingual lexicon and translation example banks. *Procedia: Social and Behavioral Sciences*, *27*, 61–69. (Tier 4; Scopus SNIP impact factor: 0.162.)

[3] Lim, L. T., Soon, L.-K., Lim, T. Y., Tang, E. K., & Ranaivo-Malançon, B. (Accepted). Lexicon+TX: Rapid construction of a multilingual lexicon with under-resourced languages. *Language Resources and Evaluation*. (Tier 2; ISI impact factor: 0.659.)

Conference Proceedings

[1] Lim, L. T. (2009). Multilingual lexicons for machine translation. In *Proceedings of the 11th International Conference on Information Integration and Web-based Applications & Services (iiWAS2009) Master and Doctoral Colloquium (MDC)* (pp. 732–736). Kuala Lumpur, Malaysia.

[2] Lim, L. T., Ranaivo-Malançon, B., & Tang, E. K. (2011). Symbiosis between a multilingual lexicon and translation example banks. In *Proceedings of the 12th Conference of the Pacific Association for Computational Linguistics (PACLING 2011).* Kuala Lumpur, Malaysia. (Tier B)

[3] Lim, L. T., Soon, L.-K., Lim, T. Y., Tang, E. K., & Ranaivo-Malançon, B. (2013). Context-dependent multilingual lexical lookup for under-resourced languages. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013).* Sofia, Bulgaria. (Tier A)