Symbiosis between a Multilingual Lexicon and Translation Example Banks

Lian Tze Lim, Bali Ranaivo-Malançon and Enya Kong Tang

Faculty of Information Technology
Multimedia University, Malaysia
liantze@gmail.com, {enyakong, ranaivo}@mmu.edu.my

Abstract

We propose a symbiotic framework in which correspondences between electronic multilingual lexicons and translation example banks can be captured, so that their functions and contents may benefit and improve upon one another. Several mechanisms are used for this purpose: i.) two flexible annotation schemas, S-SSTC and SSTC+L, for supporting irregular multi-level correspondences across languages; ii.) an axis-based translation cluster structure for connecting translation equivalents; and iii.) translation profiles, for capturing contexts of translation equivalent instances in the corpora. There are two main contributions: i.) the design of SSTC+L, which allows the annotation of multi-word expressions and translation lexical gaps in the translation examples; and ii.) the overall framework facilitating the symbiotic flow of information between the multilingual lexicon and translation bank. We give illustrative examples to show how these mechanisms can be used for translation selection, addition of new language items, and verification of lexicon contents. Preliminary tests show there is potential in our approach.

Key Words – multilingual processing, lexical resources, machine translation.

1. Introduction

Multilingual lexicons are important resources for natural language processing (NLP) applications, including machine translation (MT). One important concern in multilingual lexicon design is that lexical items (LIs) in a source language (SL) can have multiple translation equivalents in the target language (TL) of distinct meanings, possibly due to:

1. *Polysemy.* An LI in a SL has multiple distinct meanings, and hence has multiple translation equivalents in a TL.
2. *Diversification.* An LI in a SL has a single meaning, but the TL has more specific LIs. For example, Malay and Chinese distinguish *cooked* «rice» («nasi» and «饭») from *uncooked* «rice» («beras» and «米»).

Another issue is that of *lexical gaps*, where a concept is not lexicalised in a specific language and can only be expressed by a gloss-like phrase. For example, English «absent» and «fuchsia» are translated as 'tidak hadir' ('not present') and 'ungu kemerahan' ('reddish purple') in Malay.

We first review how these issues are addressed by various multilingual lexicon projects. In view of the lack of mutual information exchange between lexicons and corpora or MT systems, we then propose a symbiotic framework in which a multilingual lexicon and translation example banks can mutually benefit from information collected from the other, using real examples as illustration.

2. Multilingual Lexicon Projects

Multilingual lexicons for use with NLP applications must address the two issues above. They are also expected to contain information for supporting sense disambiguation or translation selection in MT systems. Multilingual lexicon projects can be broadly categorised as having 'deep' or 'shallow' approaches, depending on their organisation paradigm.

Lexicons adopting a 'deep' approach propose language-independent formalisms to represent concepts and meanings of words. Lexicalisations in different languages are then categorised to the relevant interlingual entries. The Universal Networking Language proposes a complete interlingual system [1], while the SIMuLLDA project [2] creates a lattice of definitional attributes as a result of formal concept analysis and is capable of translating lexical gaps systematically. Elsewhere, semantic frames [3,4] and ontology frameworks [5] have been used as interlingua.

On the other hand, 'shallow' approaches often contain simple language-neutral nodes (variously called *axes* or *pivots*), serving as a convenience mechanism for declaratively linking LIs from different languages deemed as translation equivalents to express a concept. Translation equivalents expressing distinctly different concepts would be connected to separate axes. Multilingual lexicon projects adopting such a scheme include Papillon [6], PIVAX [7] and the Lexical Markup Framework (LMF) [8]. The basic principle in Pan-

Gloss [9] is similar, mining translation sets from corpora. Diversification is indicated by adding relations between the 'main' and 'diversified' axes.

'Deep' approaches to multilingual lexicon design are strongly motivated by linguistics and formal semantics and are thus better equipped at translating lexical gaps using semantic components. Expertise in these fields is therefore required to inspect and verify the lexicons, making it an expensive process. In addition, establishment of translation equivalence can be problematic in some cases: a human may accept Chinese «跳飞机» (an informal expression with negative connotations) and Indonesian «merantau» (neutral) as mutual translations, but a formal semantic-based framework may reject it due to the stylistic mismatches [10].

3. Interaction between Lexicon and Corpora

While existing electronic multilingual lexicon projects illustrate lexical meanings using various models and frameworks, few actually retain the correspondences derived from multilingual corpora. PanGloss [9] is an exception: each multilingual translation set is associated with a topic signature extracted from corpora. Also, once deployed in NLP applications, there is very little 'feedback' from the system to the lexicon itself. There are, of course, corpus concordance and collocation tools, as well as sense-tagged corpora (albeit expensive to build by hand and to verify) but these are often seen as tools pertaining to corpora rather than to lexical resources. What we are interested in is a framework in which the lexicon and corpora are 'inter-annotated' with respect to each other, to facilitate reciprocal improvement.

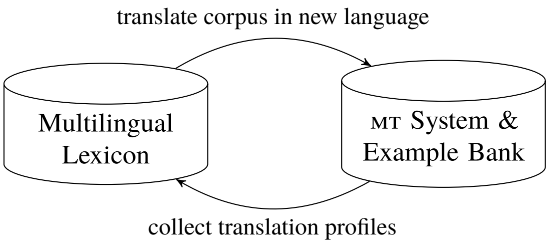

translate corpus in new language
Multilingual Lexicon
MT System & Example Bank
collect translation profiles

Figure 1. Symbiotic flow of information between a multilingual lexicon and multilingual corpora

We propose a framework for interrelating a multilingual lexicon with the translation example bank of a MT system, the interactive processing of which forms a symbiotic loop (Figure 1). By first bootstrapping from a bilingual aligned translation corpora, 'profiles' of translation equivalents are captured and associated with the respective multilingual lexicon entries. The lexicon is then used to translate new inputs in a new language, updating the translation profiles based on the new results at the same time.

In the following sections, we will describe how a translation corpus can be marked up to relate LIs to lexicon entries, including syntactically flexible multi-word expressions (MWEs) and lexical gap cases. We then describe the design of Lexicon+TX, our multilingual lexicon, and how translation profiles are bootstrapped from an aligned bilingual corpus. We show how the translation profiles can facilitate translation selection in an MT system, especially for translating from a new language and adding new LIs to Lexicon+TX as a side-effect, and how they can be used for verifying lexicon contents.

4. Synchronous String-Tree Structures

The structured string-tree correspondence (SSTC) [11] is an annotation schema for declaratively specifying (possibly irregular) correspondences between a string and its tree representation structure of arbitrary choice, at both the word (tree node) level and phrase (subtree) level. For example, the SSTC in Figure 2 captures the correspondences between the word 'John' and the *tree node* in the dependency tree using the SNODE interval `0_1`, as well as between the phrase 'the ball' and the *subtree* using the STREE interval `2_4`. The SSTC has no problem handling the discontiguous substring 'picked...up' (SNODE interval `1_2+4_5`).

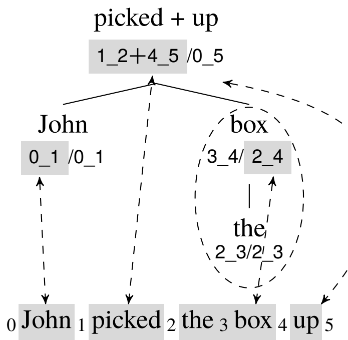
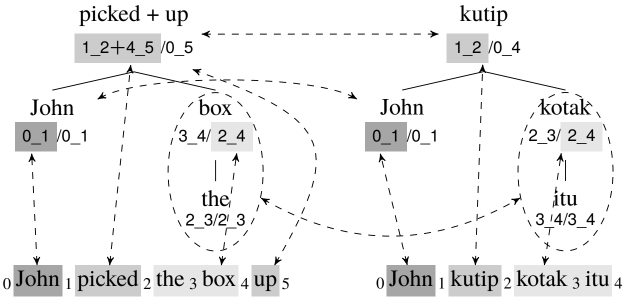


Figure 2. SSTC relating sentence and its dependency tree

The authors of [11] also proposed the synchronous structured string-tree correspondence (S-SSTC) for relating a pair of SSTCs. An S-SSTC can be used for marking up the (possibly irregular) multi-level correspondences between translation examples, as shown in Figure 3. The SNODE correspondences capture the lexical (tree node) level correspondences between the SL and TL text, while the STREE correspondences capture those on the phrase (subtree) level.

We now propose SSTC+Lexicon (SSTC+L), an extension of the SSTC, for linking substrings in a text to corresponding entries in a lexicon, also modelled as SSTCs. (The multilingual lexicon design details will be discussed in the next section.) The linking is done using SNODE intervals, as shown in Figure 4. As a simple example, 'planting' (with SNODE



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)	(3_4, 3_4)

Figure 3. S-SSTC capturing the correspondences in a translation example

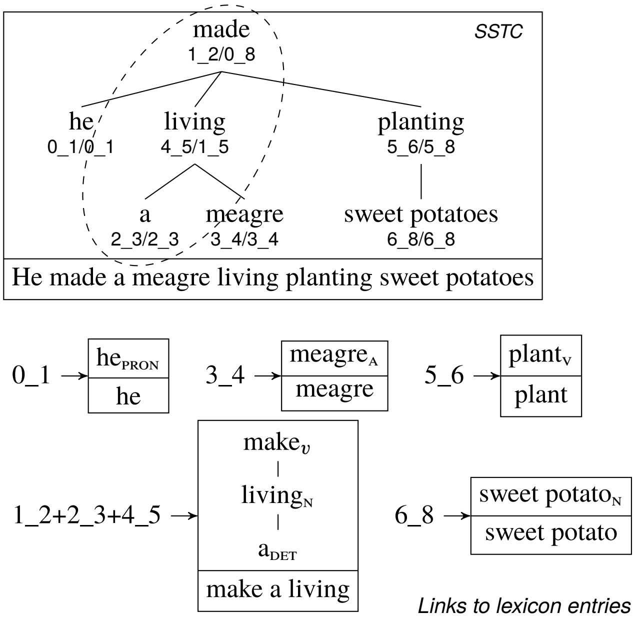


Figure 4. An SSTC+L relating LIs in a text to lexicon entries modelled as SSTCs

interval 5_6) is linked to the English verb entry for «plant», which the lexicon associates with Malay «menanam» and Chinese «种植». The schema is especially suitable for annotating syntactically flexible MWEs, such as 'make a ... living' above. Instances of MWEs where the component words are reordered, such as 'the beans are spilt', will be linked to the canonical form given in the lexicon i.e. «spill the beans».

Using the S-SSTC and SSTC+L annotation schemas in tandem, we can capture translation equivalents in an aligned corpora and their corresponding entries in a multilingual lexicon, including cases of lexical gaps such as the one in Figure 5. The English–Chinese and English–Malay S-SSTCs establish translational equivalence between «absent», «缺席» and 'tidak hadir', while the SSTC+Ls indicates that «absent», «缺席», «tidak» and «hadir» are valid LIs in their respective languages, as listed in a multilingual lexicon. The S-SSTC and SSTC+L annotations thus afford more flexibility during the MT matching phase, especially of MWEs and lexical gaps.

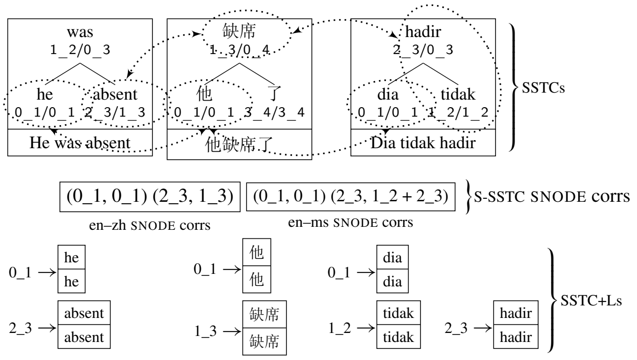


Figure 5. Annotating lexicon entries in translation examples with SSTC+L and S-SSTC when lexical gaps occur

5. Lexicon+TX

Each multilingual entry in our multilingual lexicon, Lexicon+TX, consists of 2 main parts: the translation cluster, and the translation profiles. The cluster lists translation equivalents, and the profile contains usage context information. We describe these components in the following sections.

5.1. Translation Clusters

Lexicon+TX groups translation equivalents in *translation clusters*, which are largely inspired by Papillon [6], PIVAX [7] and LMF [8]. Each cluster contains a language-independent axis, to which LIs from different languages deemed by humans as expressing the same concept are connected. To cater for MWEs, each monolingual entry is modelled as an SSTC. (Single word LIs have a trivial tree consisting of a single leaf.)

When diversification occurs, a new axis is created for connecting the more specific translations, and a link is added to the original axis. Note that the axes are not meant to constitute an actual interlingua, but rather as a convenience mechanism for linking translation equivalents.

Figure 6 shows an example translation cluster in Lexicon+TX associating English «make a living» and its translations in Malay («mencari nafkhah», «mencari rezeki») and Chinese («谋生», «找生活»). Most members here happen to be MWEs, including one with a 'variable' («make one's living»), demonstrating how they are modelled as SSTCs in Lexicon+TX. There is one diversified axis for English «scrape a living» and Chinese «糊口», as they mean '*barely* making a living'.

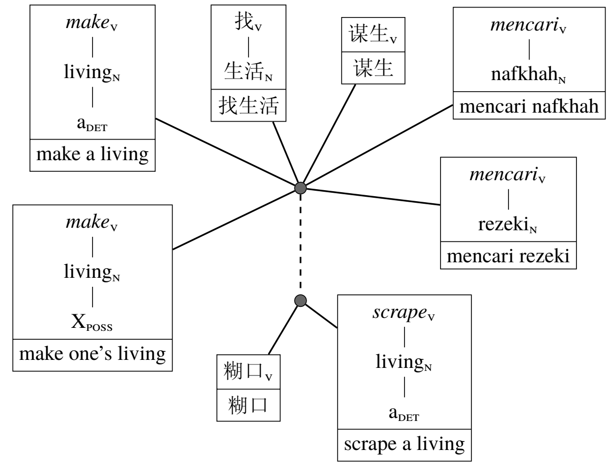


Figure 6. An example multilingual translation cluster in Lexicon+TX

5.2. Translation Profiles

The translation profile of a multilingual translation cluster captures instances of the translation equivalents in corpora using *sub-clusters* and *vectorial representations* of their usage context.

5.2.1. Sub-clusters. A translation cluster can contain multiple translation equivalent in a given language, for example «岸», «河岸» and «河畔» for English «bank» (*sloping land beside a river*) in Figure 7 (SSTCs have been simplified), but some may be used more frequently by human translators. Such information can be captured as translation 'sub'-clusters, from aligned translation corpora, and from post-editing lexical substitution actions of MT outputs.

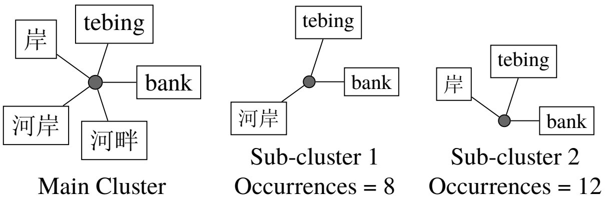


Figure 7. Translation sub-clusters

For example, Figure 7 shows two translation sub-clusters indicating that «河岸» has been used 8 times to translation «bank» and «tebing», while «岸» was used 12 times. That «河畔» does not appear in any sub-clusters simply means that it has not yet been seen in the corpus nor selected by a human post-editor.

5.2.2. Vectorial Representations. Based on the premise of distributional semantics that words that occur in the same contexts tend to have similar meanings [12], various vectorial representations have been designed to model word meanings, including the standard vector model [13] and latent semantic indexing (LSI) [14]. Each translation cluster and sub-cluster in Lexicon+TX will be associated with a vector representing the concept and/or context it occurs in.

While any vector model can be used, we describe briefly how vectors can be auto-generated from aligned corpora using LSI. We bootstrap from the translation example bank of an existing bilingual MT system. We run LSI on the translation example bank, treating each translation example as a bilingual document in the manner of [15]. Terms are extracted based on the SSTC+L annotations. A term-document matrix is constructed using the frequency of terms in each document, and singular value decomposition is then performed. We thus obtain a vector for every LI (in both languages) occurring in the translation example bank. The vector associated with each translation cluster can then be set as the term-to-term product of all available vectors of its member LIs, to emphasise the context overlaps as its 'core' meaning.

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

6. Symbiotic Actions

To illustrate how the translation profiles can be used for different purposes, we ran LSI on the small English–Malay translation example bank shown in Table 1 with 2 factors, using the EJML library[1]. We filter stop words and stem the English words, but not the Malay ones as Malay is a derivative language. The translation profiles of Lexicon+TX clusters are bootstrapped by taking the normalised term-to-term product of all member LIs' vectors.

6.1. Translation Selection for a New Language

To select an appropriate translation equivalent for «bank» in the following input:

'He stop bathing and clambered up the **bank**'

we construct a query vector by summing up the vectors of all terms in the query (skipping stop words):

$$V_Q = V(\text{«stop»}) + V(\text{«bath»}) + V(\text{«clamber»}) + V(\text{«bank»}).$$

The MT system can then select a translation equivalent for «bank» from Lexicon+TX by computing the cosine similarity

[1] http://code.google.com/p/efficient-java-matrix-library/

Table 1. English–Malay translation examples

English	Malay
I deposited my salary with the bank	Saya memasukkan wang gaji saya di bank
You should only borrow money from a bank	Anda patut meminjam wang dari bank sahaja
Money lending activities	Aktiviti meminjam wang
We lazed by the river bank	Kami berehat di tepi tebing sungai
The river bank was soon inundated by the flood water	Tebing sungai dibanjiri air bah dengan cepat nya
We bathed in the cool river water	Kami bermandi-manda di tengah air sungai yang sejuk

between V_Q and vectors of all clusters containing «bank». The cosine similarity, CSim of two vectors X, Y is

$$\text{CSim}(X,Y) = \frac{X \cdot Y}{|X||Y|} \tag{1}$$

Given translation clusters $T_1 = \{\text{«bank», «bank»}\}$ (for *financial institution*) and $T_2 = \{\text{«bank», «tebing»}\}$ (for *riverside*), we find

$$\text{CSim}(V_Q, V_{T_1}) = 0.716 \qquad \text{CSim}(V_Q, V_{T_2}) = 0.862.$$

We thus select «tebing» from T_2.

It is also possible to perform translation selection for a new language, e.g. Chinese, using Lexicon+TX (now containing English and Malay LIs) and a separate Chinese–English bilingual lexicon. Given the following text:

'银行 借贷' (*bank lending and borrowing; loans*)

The query vector is constructed by summing up the vectors of all possible translation equivalents of the Chinese termsby first consulting the Chinese–English list, and then Lexicon+TX:

$$\begin{aligned} V_Q = {} & V(\text{«bank»}_\text{E}) + V(\text{«bank»}_\text{M}) + V(\text{«tebing»}) \\ & + V(\text{«borrow»}) + V(\text{«meminjam»}) \end{aligned}$$

To translate «银行», we compare the query vector to those of possible target translation clusters, i.e. of T_1 and T_2. The results are

$$\text{CSim}(V_Q, V_{T_1}) = 0.987 \qquad \text{CSim}(V_Q, V_{T_2}) = 0.760,$$

indicating that «银行» can be added to Lexicon+TX as a new member of T_1.

6.2. Lexicon Self-Verification

In the event that a multilingual lexicon 'draft' has been generated by some automatic procedure e.g. as described in [16], chances are that the entries require further verification and cleaning up. For example, English «bank», Malay «tebing», «bank», Chinese «岸», «银行» may all be placed in the same translation cluster, i.e. both the *financial institution* and *river side* senses of English «bank» are linked to the same axis.

We now describe how a 'draft' multilingual lexicon can perform self-verification using Lexicon+TX's translation profiles. The lexicon contains English, Malay and Chinese LIs, and includes the erroneous translation cluster $T = \{\text{«bank»}_\text{E}, \text{«bank»}_\text{M}, \text{«tebing»}, \text{«银行»}, \text{«岸»}\}$. The vector for this cluster is computed based on the English–Malay translation examples in Table 1, as described in the previous section. We then use the lexicon and vectors to translate new Chinese texts, recording translation sub-clusters of T and updating their vectors in the process. Eventually, two sub-clusters with non-trivial frequencies will stand out:

$$\begin{aligned} T_a &= \{\text{«bank»}_\text{E}, \text{«bank»}_\text{M}, \text{«银行»}\} \\ T_b &= \{\text{«bank»}_\text{E}, \text{«tebing»}, \text{«岸»}\} \end{aligned}$$

We next check the angular distance between the vectors of T_a and T_b:

$$\angle(T_a, T_b) = \arccos(\text{CSim}(T_a, T_b)) = 74.78^\circ.$$

By taking a threshold value e.g. 40°, $\angle(T_a, T_b) = 74.78^\circ$ would indicate that T should be split into two distinct translation clusters T_a and T_b.

7. Preliminary Experiment and Results

An English–Malay translation example bank of 25275 example sentence pairs taken from a bilingual dictionary with 200 factors was indexed using `gensim`[2]. As an initial evaluation, we looked at how well translations of seven English test words can be discerned. Translation clusters were formed by taking the two most frequently occurring Malay translation equivalents (reflecting different meanings) for each test word. The angular distance between these translation clusters were then calculated, as shown in Table 2.

A threshold value of 40° would be able to differentiate the two translation clusters for 6 of the 7 test words. The two

[2] `http://nlp.fi.muni.cz/projekty/gensim/`. `gensim` was used instead of `EJML` in this experiment to avoid the large memory footprint.

Table 2. Angular distance between translation clusters containing polysemous English LIs

English LI, E	Malay translations		∠(E–M$_1$, E–M$_2$)
	M$_1$	M$_2$	
bank	bank	tebing	40.76°
plant	tumbuhan	loji	66.67°
letter	surat	huruf	8.88°
account	akaun	cerita	86.80°
glass	gelas	kaca	56.79°
draw	menarik	melukis	56.06°
sentence	ayat	hukuman	75.76°

clusters for «letter» are too close because the translation examples for both meanings include «read»ing and «write»ing. «bank» faces a similar problem, though to a less degree.

We also evaluated translation selection of these 7 test words in 27 English test sentences. The accuracy was 74.06 %, against a baseline of 48.15 % where the most frequent translation is always selected. The results show again that accuracy is highly dependant on the context words of the translation examples, which are very sparse and short to begin with. We are confident that better results (for both lexicon verification and translation selection) can be achieved given a denser corpora.

8. Conclusion and Future Work

We have proposed a symbiotic framework for relating a multilingual lexicon with translation example banks, such that they can benefit from and improve upon each other. Specifically, we proposed i.) SSTC+L, an annotation schema for relating text segments to lexicon LI entries including cases of MWEs and lexical gaps; ii.) Lexicon+TX, a multilingual lexicon with an axis-based mechanism for connecting translation equivalents in translation clusters; and iii.) translation profiles, in which translation sub-clusters and LSI vectorial representations of their usage context are collected from a translation example bank and associated.

Preliminary tests show potential in our approach for translation selection, adding new language LI and verifying the multilingual lexicon, although the results are hampered by the sparseness of current data. We intend to repeat our experiment using comparable corpora of greater length to overcome the sparseness.

Acknowledgements

We thank the three anonymous reviewers for their comments in improving this paper. This work was sponsored by an FRGS research grant (FRGS/1/10/TK/MMU/02/02) from the Malaysian Ministry of Higher Education.

References

[1] H. Uchida, M. Zhu, and T. D. Senta, *Universal Networking Language*. UNDL Foundation, 2005.

[2] M. Janssen, “Multilingual lexical databases, lexical gaps, and SIMuLLDA,” *International Journal of Lexicography*, vol. 17, pp. 136–154, 2004.

[3] H. Boas, “Semantic frames as interlingual representations for multilingual lexical databases,” *International Journal of Lexicography*, vol. 18, no. 4, pp. 445–478, 2005.

[4] B. J. Dorr and M. B. Olsen, “Multilingual generation: The role of telicity in lexical choice and syntactic realization,” *Machine Translation*, vol. 11, no. 1–3, pp. 37–34, 1996.

[5] P. Edmonds and G. Hirst, “Near-synonymy and lexical choice,” *Computational Linguistics*, vol. 28, no. 2, pp. 105–144, 2002.

[6] C. Boitet, M. Mangeot, and G. Sérasset, “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)*, 2002, pp. 1–3.

[7] H.-T. Nguyen, C. Boitet, and G. Sérasset, “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, 2007.

[8] International Organization for Standardization, *ISO 24613:2008 Language Resource Management – Lexical Markup Framework (LMF)*, 2008.

[9] M. Sammer and S. Soderland, “Building a sense-distinguished multilingual lexicon from monolingual corpora and bilingual lexicons,” in *Proceedings of Machine Translation Summit XI*, Copenhagen, Denmark, 2007, pp. 399–406.

[10] L. T. Lim, “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)*, Kuala Lumpur, Malaysia, 2009, pp. 732–736.

[11] M. H. Al-Adhaileh, E. K. Tang, and Y. Zaharin, “A synchronization structure of SSTC and its applications in machine translation,” in *Proceedings of COLING-2002 Workshop “Machine translation in Asia”*, Taipei, Taiwan, 2002.

[12] Z. Harris, “Distributional structure,” *Word*, vol. 10, no. 23, pp. 146–162, 1954.

[13] G. Salton, A. Wong, and C. S. Yang, “A vector space model for automatic indexing,” *Communications of the ACM*, vol. 18, no. 11, 1975.

[14] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, “Indexing by latent semantic analysis,” *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[15] S. T. Dumais, M. L. Littman, and T. K. Landauer, “Automatic cross-language retrieval using latent semantic indexing,” in *AAAI97 Spring Symposium Series: Cross Language Text and Speech Retrieval*, Stanford University, 1997, pp. 18–24.

[16] L. T. Lim, B. Ranaivo-Malançon, and E. K. Tang, “Low cost construction of a multilingual lexicon from bilingual lists,” *Polibits*, to appear.