Chapter 8

LMF and Its Implementation in Some Asian Languages[1]

8.1 Introduction

Corpus-based approaches and statistical approaches have been the main stream of natural language processing research for the past two decades. One of the advantages of these approaches is that the techniques are less language specific than classical rule-based approaches where a human analyses the behaviour of target languages and constructs rules manually. The language resources play a key role in such approaches. There is a long history of creating a standard for Western language resources. The Human Language Technology (HLT) society in Europe has been particularly zealous for its standardization, making a series of attempts such as EAGLES[2], PAROLE/SIMPLE (Lenci et al., 2000), ISLE/MILE (Calzolari et al., 2003) and LIRICS[3]. These continuous efforts have been crystallized as activities in ISO-TC37/SC4 which aims at making an international standard for language resources.

However, due to the great diversity of languages themselves and the level of current development of technology for each language, it is still unclear if corpus-based techniques developed for well-computerised languages are applicable to all Asian languages. In particular, language resources play a key role in such approaches, but there is an insufficient amount of language resources in many Asian languages. In such situation, creating a common standard for Asian language resources that is compatible with an international standard has at least three strong advantages: (i) to increase the competitive edge of Asian countries, (ii) to bring Asian countries to closer to their Western counterparts, (iii) and to bring more cohesion among Asian countries.

This paper aims at creating a common standard for Asian language resources that is compatible with an international standard - Lexical Markup Framework (LMF; ISO24613). In particular, it focuses on four issues: i) lexical specification and data categories relevant for building multilingual lexical resources for Asian languages (Section 2); ii) a core upper-layer ontology needed for ensuring

---

[1] Chapter written by Takenobu Tokunaga, Sophia Y. M. Lee, Virach Sornlertlamvanich, Kiyoaki Shirai, Shu-Kai Hsieh, Chu-Ren Huang

[2] http://www.ilc.cnr.it/Eagles96/home.html

[3] http://lirics.loria.fr/documents.html

Preprint version of Chapter 8 (pp. 119-132) in "LMF Lexical Markup Framework", Gil Francopoulo Ed. ISTE Ltd and John Wiley&Sons, Inc., 2013.

multilingual interoperability (Section 3) and iii) the evaluation platform used to test the entire architectural framework (Section 4).

8.2. Lexical Specification and Data Categories

8.2.1 Lexical Specification

The lexical specification used in this paper is based on and compliant with the Lexical Mark-up Framework (LMF) (Francopoulo et al., 2006), the high-level conceptual model developed within both the European e-Content Project LIRICS and ISO TC37/SC43[4]. LMF is a structural data model expressed by a set of UML packages, each of which contains lexical classes. It is comprised of a core package and a set of extensions. Each class is described by an UML specification for linking with other classes and can be adorned by a set of attribute-value pairs taken from a data category registry. Lexical classes and data categories provide the main building blocks for a common shared representation of lexical objects that allows the encoding of rich linguistic information.

We have contributed to ISO TC37/SC4 activities by testing and ensuring the portability and applicability of LMF to the development of a description framework for NLP lexicons for Asian languages. A major achievement has been the proposal of necessary extensions of the framework with respect to requirements and characteristics of Asian languages. This activity culminated in the modeling of additional packages concerning the characteristics of Asian languages to be incorporated in the LMF standard.

We have also contributed to the finalisation of the LMF draft revision 144 including (1) a package for derivational morphology, (2) the syntax-semantic interface with the problem of classifiers, and (3) representational issues with the richness of writing systems in Asian languages (Chung et al., 2006; Prevot et al., 2006).

As a proof-of-concept of the conceptual framework, the first version of our lexical model has been implemented in RDF-OWL and a first set of sample lexical entries has been developed in XML. The XML implementation conforms to the LMF DTD. The multilingual lexicons are intended to be used in NLP implementations and systems that support multilingual information retrieval applications for Asian languages and test usability and viability of the proposed framework (Tokunaga et al., 2006).

8.2.2. Data Categories

The activity of designing a high-level conceptual model for harmonised lexicons in this paper has been conducted in connection with the formulation of a set of low-level standards, i.e. data categories needed for adorning this structure and populating

[4] http://www.tc37sc4.org

the different layers of the lexical data model. The relation between the lexical meta-model and the data categories is an important point to mention, the first being a specification of the structure of a lexicon, the latter being linguistic constants taken from a harmonised registry.

The property of splitting the structure and the adornment is shared by all specifications that are developed within ISO-TC37/SC4. One of our specific purposes is the identification of data categories needed for the representation of peculiar features of Asian languages. An initial set of data categories at different layers of linguistic representation was isolated and contributed in particular to ISO TDG2, the Morpho-syntactic Profile. The development of lexical suites allows implementers to combine the meta-model with the relevant data categories taken from the registry. They can thus be used as examples of the application of data categories themselves and as a reference to the best practices in the representation of a given linguistic phenomenon. Some of the data categories identified and proposed are exemplified below.

Classification of derivation Derivation is a more complicated phenomenon and less studied than inflection. Thus, a specific package has been devised to deal with it. For instance, Japanese has at least four types of derivation: affixation, compounding, reduplication and borrowing. Among those, reduplication is one of distinguishing features of some Asian languages, such as Chinese and Thai. We further investigate data categories specific for reduplication.

Reduplication Reduplication is a common linguistic phenomenon in many Asian languages realising various functions such as plurality. In Chinese, 慢(man4) 'to be slow' is a state verb, while a reduplicated form 慢慢 (man4-man4) is an adverb. 看 (kan4) 'to look' is an activity verb, while the reduplicative form 看看 (kan4-kan4), refers to the tentative aspect, introducing either stage-like sub-division or the event or tentativeness of the action of the agent. This case involves verbal aspect.

Thai also has many functions realised by reduplication. A study on contemporary Thai corpora suggests at least the following five functions of reduplication.

(a) Pluralisation (to express plurarity of objects, for example □□□□ (dek0) 'child' has a reduplication form □□□□□ (dek0-dek0) 'children'.)
(b) Generalisation (to express a vague sense of a word, for example □□ (dam0) 'black' has a reduplication form □□□ (dam0-dam0) 'blackish'.)
(c) Intensification (to express a higher degree of modification, for example □□□ (mued2) 'dark' has a reduplication form □□□□ (mued2-mued2) 'very dark'.)
(d) Continuation (to express the continuation of an action for a certain period of time literally, and implicitly suggesting a specific manner of that action. For example □□□ (khid3) 'think' can be reduplicated to form □□□□ (khid3-khid3) 'think longer'. In this case, thinking for a certain period of time implies deliberate thinking.)

(e) Individualisation (to express individual from the generic group, for example □□□ (tua0:classifier) 'one' has a reduplication form □□□□ (tua0-tua0:adverb) 'one by one'.)

To deal with such complicated variations, two data categories have been proposed for reduplication: reduplicationType and reduplicationFunction. ReduplicationType specifies the surface relations between an original form and its reduplicated form. In the previous Chinese example 慢慢 is obtained by duplicating the same character twice. This type could be labeled as type 'AA', and its function 'plural' specified as a value of ReduplicationFunction.

Classifiers Many Asian languages do not distinguish singularity and plurality of nouns, but instead use numerative classifiers to denote the number of objects. In addition, semantic agreement between classifiers and nouns should be taken into account. This agreement is not as simple as number and gender agreement in European languages; it is rather similar to a selectional restriction on arguments of predicates. It is still uncertain if we can enumerate possible agreement combinations as values of a data category. We alleviated this problem by building a linguistically motivated ontology which can be used for describing noun-classifier agreement.

We have proposed a method to construct a taxonomy based on noun-classifier agreement data. Superordinate-subordinate relations are first extracted based on subsumption relations of noun sets corresponding to classifiers, and then a taxonomy is automatically constructed using these extracted relations.

Preliminary experiments were conducted by using noun-classifier agreement data of three languages: Chinese, Japanese and Thai, and we found this approach worked well for Chinese and Japanese but not for Thai (Shirai et al., 2008). In Thai, relations between a noun and a classifier are tightly coupled and fail to produce a structure of classifiers.

Honorifics Many Asian languages have some level of distinction at the lexical level representing the differences between members of a conversation based on their social level, i.e. superior/inferior. Our research has initially focused on three Asian languages: (1) Thai, (2) Japanese and (3) Chinese. Thai has a developed honorific system. The usage of Thai honorifics depends on (1) social status, (2) seniority and (3) formal and informal relationships for social and commercial links. In summary, there are four types of honorific words in Thai:

(a) Special diction for the King and the royal family,
(b) Special diction for religious figures,
(c) Respectful forms, and
(d) Polite forms.

There are some Thai words that have their own equivalents for polite senses used in formal situations or in written language.

The Japanese honorific system has four forms: respectful, humble, polite and special diction for the Imperial Family. Respectful forms show respect to those in higher positions (e.g. a boss at work, a customer and so on). Humble forms also show respect to others, but it is achieved by the speakers abasing themselves. Polite forms show politeness without differentiating social level. The detailed categories of the Japanese honorific system are as follows.

(a) Respectful forms
(b) Humble forms concerning third persons
(c) Humble forms concerning the hearer
(d) Polite forms
(e) Beautification
(f) Special diction for the Imperial Family

Although honorific systems depend heavily on both language and culture, and therefore may vary greatly between two separate languages/cultures, we have designed a prototype of universal data categories (DC) for honorifics: (a) Respectful, (b) Polite, (c) Diction for special social strata and (d) Other. These categories are intentionally broad and are intended as a basis for all languages with honorifics. It is our intention that they be further subdivided into more detailed categories for each language as applicable.

Orthography Many Asian languages involve more than one writing script, unlike many western languages. In many cases, an original script and Latin characters are used together. Among many Asian languages, Japanese probably has the most complicated writing system; four writing scripts are used in Japanese, i.e. *hiragana*, *katakana*, *kanzi* and Latin characters in romanisation. This variety can be represented by the combination of two attributes: 'scriptName' and 'orthographyName'. The complication here is that some words can be represented by a mixture of kanzi and hiragana scripts. Therefore, an attribute value of kanzi allows for using hiragana together with the kanzi script. In addition, there can be variations in the kanzi writing system. Thus when implementing this in LMF, multiple FormRepresentation instances should be allowed with the same script and orthography values but different writtenForm values.

8.3. Upper-layer Ontology

We have constructed a conceptual core for a multilingual ontology, with the main focus on Asian language diversity and the necessary attention devoted to the ontological design of the upper level. Different from traditional approaches for designing a core lexicon, we proposed a novel approach by starting from the Swadesh List (Swadesh, 1952) of different language versions, such as Chinese,

English, Bangla, Malay, Cantonese and Taiwanese. The reason why we consider the Swadesh list as the potential core lexicon is due to the lack of available resources for many languages. The list can be seen as a least common denominator for vocabulary. Various lexical-conceptual patterns have been explored with the discussion of cultural specificities.

In order to highlight the granularity issue, we also compare the coverage of the Swadesh list with the one of the Base Concept Set (BCS) as it is proposed by the Global WordNet Association[5]. Since both the Swadesh list and BCS are linked to an upper-layer ontology, SUMO (Niles and Pease, 2001), we compared the repartition of their mappings to SUMO (Huang et al., 2007).

Given this data, we experimented with designing a core upper-layer ontology with the purpose of multilingual resources standardisation and processing (Hsieh et al., 2007). We take a hybrid approach by supplementing SUMO with MILO (Mid-Level Ontology) as the foundation. By pruning the Swadesh-SUMO/MILO mapping ontological structure, we obtain a proper ontology for representing the concepts in the Swadesh list. To attest the robustness of our proposed approach, we also apply our approach to two Austronesian languages: Seediq and Kavalan. These preliminary experiments yielded promising results which motivate our ongoing work on other Asian languages.

8.4. Evaluation Platform

We evaluated the effectiveness of LMF on a multilingual information retrieval system. The system has two significant features: dimensionality reduction by using parallel corpora and linguistically motivated query expansion.

The representation of queries and documents is a key problem for information retrieval. The vector space model (VSM) has been widely used in this domain. The VSM suffers, however, from high dimensionality. Due to this high dimensionality, the vectors built from documents are complex and can contain substantial noise. We proposed a novel method that reduces the dimensionality using parallel corpora (Xia and Yu, 2007). We introduced a new metric called frequency distance to measure the translation consistency constraints. The frequency distance is used to reduce the number of index terms to be considered, improving system performance.

The linguistically motivated query expansion system aims to refine a user's query by exploiting the richer information contained within a lexicon described using the adapted framework. For example, a user inputs a keyword 'ticket' as a query. Conventional query expansion techniques expand this keyword to a set of related words by using thesauri or ontologies. Using the framework proposed in this paper, expanding the user's query becomes a matter of following links within the lexicon, from the source lexical entry or entries through predicate-argument structures to all relevant entries. We focus on expanding the user inputted list of nouns to relevant verbs, but the reverse would also be possible using the same

5 http://www.globalwordnet.org/

technique and the same lexicon. This link between entries is established through the *semantic type* of a given sense within a lexical entry. These semantic types are defined by higher-level ontologies, such as MILO or SIMPLE (Lenci et al., 2000) and are used in semantic predicates that take such semantic types as a restriction argument. Since senses for verbs contain a link to a semantic predicate, using this semantic type, the system can then find any/all entries within the lexicon that have this semantic type as the value of the restriction feature of a semantic predicate for any of their senses. By referring to the lexicon, we can then derive any actions and events that take the semantic type 'ARTIFACT' as an argument.

First, all semantic predicates are searched for arguments that have an appropriate restriction, in this case 'ARTIFACT'. Any lexical entries that refer to these predicates are then returned. An equally similar definition would exist for 'buy', 'find' and so on. Thus, by referring to the predicate-argument structure of related verbs, we know that these verbs can take 'ticket' in the role of object. The system then returns all relevant entries, here 'buy', 'sell' and 'find', in response to the user's query.

The system itself is being developed in Java for its 'compile once, run anywhere' portability and its high availability of reusable off-the-shelf components. The most popular free open-source database was selected, MySQL, to store all lexicons imported into the system. Though still preliminary and subject to change, the schema describes the relationships between entities, and more or less mirrors the classes found within the adapted LMF framework, with mostly only minor exceptions where it was efficacious for querying the data. Further details can be found in Tokunaga et al. (2008).

A lexicon is imported into the system using an import utility. After import, this data may be immediately queried upon with no other changes to system configuration. The hope being that regardless of language, the rich syntactic/semantic information contained within the lexicon will be sufficient for carrying out query expansion on its own.

Next steps for the evaluation platform are to explore the use of other information already defined within the adapted framework, specifically sense relations. Given the small size of our sample lexicon, data sparsity is naturally an issue. However, by exploring and exploiting these sense relations properly, the system may be able to further expand a user's query to include a broader range of selections using any additional semantic types belonging to these related senses. The framework also contains information about the order in which syntactic arguments should be placed. This information should be used to format the results from the user's query appropriately.

We also conducted some additional query expansion experiments using a corpus that was acquired from Chinese LDC (No. "2004-863-009") as a base (see below). This corpus marked an initial achievement in building a multilingual parallel corpus for supporting development of cross-lingual NLP applications catering to the Beijing 2008 Olympics.

The corpus contains parallel texts in Chinese, English and Japanese and covers five domains that are closely related to the Olympics: traveling, dining, sports, traffic and business. The corpus consists of example sentences, typical dialogues and articles from the Internet, as well as other language teaching materials. To deal with the different languages in a uniform manner, we converted the corpus into our proposed LMF-compliant lexical resources framework, which allowed the system to expand the query between all the languages within the converted resources without additional modifications. For details of how this IR system functions, please refer to Tokunaga et al. (2009).

Results showed that this sort of query expansion is still too naive to apply to real IR systems. It should be noted, however, that our current aim of evaluation was in confirming the advantage of LMF in dealing with multiple languages, for which we conducted a similar run with Chinese and Japanese. It also showed that in following the LMF framework in describing lexical resources, it was possible to deal with all three languages without changing the mechanics of the system at all.

8.5. Discussion

LMF is, admittedly, a "high-level" specification, that is, an abstract model that needs to be further developed, adapted and specified by the lexicon encoder. LMF does not provide any off-the-shelf representation for a lexical resource; instead, it gives the basic structural components of a lexicon, leaving full freedom for modeling the particular features of a lexical resource. One drawback is that LMF provides only a specification manual with a few examples. Specifications are by no means instructions, exactly as XML specifications are by no means instructions on how to represent a particular type of data.

Going from LMF specifications to a true instantiation of an LMF-compliant lexicon is a long way, and comprehensive, illustrative and detailed examples for doing this are needed. Our prototype system provides a good starting example for this direction. LMF is often taken as a prescriptive description, and its examples taken as pre-defined normative examples to be used as coding guidelines. Controlled and careful examples of conversion to LMF-compliant formats are also needed to avoid too subjective an interpretation of the standard.

We believe that LMF will be a major base for various Semantic Web applications because it provides interoperability across languages and directly contributes to the applications themselves, such as multilingual translation, machine aided translation and terminology access in different languages.

From the viewpoint of LMF, our prototype demonstrates the adaptability of LMF to a representation of real-scale lexicons, thus promoting its adoption to a wider community. This paper is one of the first test-beds for LMF (as one of its drawbacks being that it has not been tested on a wide variety of lexicons), particularly relevant since it is related to both Western and Asian language lexicons. The present work is a concrete attempt to specify an LMF-compliant XML format, tested for representative and parsing efficiency, and to provide guidelines for the

implementation of an LMF-compliant format, thus contributing to the reduction of subjectivity in interpretation of standards.

From our viewpoint, LMF has provided a format for exchange of information across differently conceived lexicons. Thus LMF provides a standardized format for relating them to other lexical models, in a linguistically controlled way. This seems an important and promising achievement in order to move the sector forward. Once tested at the relatively local level of our lexical grid, it can be a candidate for integration in another wide lexical grid: in the framework of the KYOTO project (Vossen et al., 2010), different European and Asian WordNets are being interlinked through a format which is dialect of LMF. The LMF format will serve as a representational bridge to evaluate the needs and problems posed by making two lexical grids interoperable.

8.6. Conclusion

This paper presented our collaborative development of an international standard for Asian language resources in cooperation with other ISO TC37/SC4 related initiatives. By adopting LMF and with the aim to provide LMF a more comprehensive coverage of languages in the world, we achieved the following goals:

- We contributed to ISO TC37/SC4 activities and ISO 24613 by testing and ensuring the portability and applicability of LMF, based on the development of a description framework for NLP lexicons for Asian languages. Our contribution includes (1) a package for derivational morphology, (2) the syntax-semantic interface with the problem of classifiers, and (3) representational issues with the richness of writing systems in Asian languages.
- We provided description of Data Categories that were not previously available in LMF, including reduplication, classifier, honorifics and orthography, through surveying and careful analysis of Asian languages.
- We designed and implemented an evaluation platform of our description framework. We focused on linguistically motivated query expansion module. The system works with lexicons compliant with LMF and ontologies. Its most significant feature is that the system can deal with any language as far as those lexicons are described according to LMF.

In this paper, we mainly worked on three Asian languages, Chinese, Japanese and Thai, on top of the existing framework. We are going to distribute our results to HLT societies of other Asian languages, requesting for their feedback through various networks, such as the Asian language resource committee network under Asian Federation of Natural Language Processing (AFNLP)[6], and the Asian

6 http://www.afnlp.org/

Language Resource Network project[7]. We believe our efforts contribute to international activities like ISO-TC37/SC4 (Francopoulo et al., 2006) as well as LMF's wider coverage of the world's typologically different languages.

Acknowledgements
This work was partly supported by General Research Fund, Hong Kong (Ref. No: B-Q24K).

References

S. Chung, K. Hasan, T. Jiang, S. Lee, I. Su, L. Prevot and C. Huang. 2006. Extending an international lexical framework for Asian languages, the case of Mandarin, Taiwanese, Cantonese, Bangla and Malay. *第五屆數位典藏技術研研討會*, Taipei: Academia Sinica. August, 31-September, 1.

N. Calzolari, F. Bertagna, A. Lenci, and M. Monachini. 2003. Standards and best practice for multilingual computational lexicons. MILE (the multilingual ISLE lexical entry). ISLE Deliverable D2.2&3.2.

G. Francopoulo, G. Monte, N. Calzolari, M. Monachini, N. Bel, M. Pet, and C. Soria. 2006. Lexical markup framework (LMF). In *Proceedings of LREC2006*.

S. Hsieh, I. Su, C. Huang, P. Hsiao, T. Kuo, and L. Prevot. 2007. Basic lexicon and shared ontology for multilingual resources: A sumo + milo hybrid approach. In *Proceedings of OntoLex Workshop in the 6th International Semantic Web Conference*, Busan.

C. Huang, L. Prevot, and I. Su. 2007. Toward a conceptual core for multicultural processing: A multicultural ontology based on the swadesh list. In *Proceedings of the 1st International Workshop on Intercultural Collaboration (IWIC)*, Kyoto.

A. Lenci, N. Bel, F. Busa, N. Calzolari, E. Gola, M. Monachini, A. Ogonowsky, I. Peters, W. Peters, N. Ruimy, M. Villegas, and A. Zampolli. 2000. SIMPLE: A general framework for the development of multilingual lexicons. *International Journal of Lexicography, Special Issue, Dictionaries, Thesauri and Lexical-Semantic Relations*, XIII(4):249–263.

I. Niles and A. Pease. 2001. Towards a standard upper ontology. In *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*.

L. Prevot, C. Huang, K. Hasan, S. Lee, I. Su, S. Chung and T. Jiang. 2006. Meta-modeling and Standardization Issues for Asian Languages Lexical Resources. In *Proceedings of International Conference on Terminology, Standardization and Technology Transfer*. pp. 151-162. Beijing: Encyclopedia of China Publishing House. August, 25-26.

K. Shirai, T. Tokunaga, C. Huang, S. Hsieh, I. Kuo, V. Sornlertlamvanich, and T. Charoenporn. 2008. Constructing taxonomy of numerative classifiers for Asian

7 http://www.language-resource.net/

languages. In *Proceedings of the third International Joint Conference on Natural Language Processing (IJCNLP 2008)*, 397–402.

M. Swadesh. 1952. Lexico-statistical dating of prehistoric ethnic contacts: With special reference to north American Indians and Eskimos. In *Proceedings of the American Philo-sophical Society*, volume 96, 452–463.

T. Tokunaga, V. Sornlertlamvanich, T. Chareonporn, N. Calzolari, M. Monachini, C. Soria, C. Huang, Y. Xia, H. Yu, L. Prevot, and K. Shirai. 2006. Infrastructure for standardization of Asian language resources. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, 827–834.

T. Tokunaga, V. Sornlertlamvanich, T. Charoenporn, N. Calzolari, M. Monachini, C. Soria, C. Huang, S. Hsieh, K. Shirai and Y. Xia. 2008. Adapting International Standard for Asian Language Technologies. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008)*. Marrakech, Morocco. May 28-30.

T. Tokunaga, D. Kaplan, N. Calzolari, M. Monachini, C. Soria, V. Sornlertlamvanich, T. Charoenporn, Y. Xia, C. Huang, S. Hsieh and K. Shirai. 2009. Query expansion using LMF-compliant lexical resources. In *Proceedings of the 7th Workshop on Asian Language Resources (ALR7), ACL-IJCNLP 2009*. Singapore, August 2-9.

P. Vossen, E. Agirre, F. Bond, W. Bosma, C. Fellbaum, A. Hicks, S. Hsieh, H. Isahara, C. Huang, K. Kanzaki, A. Marchetti, G. Rigau, F. Ronzano, R. Segers, M. Tesconi. 2010. KYOTO: a Wiki for Establishing Semantic Interoperability for Knowledge Sharing across Languages and Cultures. In Blanchard, E. and D. Allard Eds, *Handbook of Research on Culturally-Aware Information Technology: Perspectives and Models*. IGI Global USA, p. 265-294.

Y. Xia and H. Yu. 2007. Dimensionality reduction with parallel corpora. In *Proceedings of IADIS European Conference on Data Mining*, 113–118, Lisbon, July.