Extending an international lexical framework for Asian languages, the case of Mandarin, Taiwanese, Cantonese, Bangla and Malay

Siaw-Fong Chung[1,3] Tian-Jian Jiang[2] Kamrul Hasan[3] Sophia Lee[3] I-Li Su[3] Laurent Prévot[3] Chu-Ren Huang[3]

[1.] Graduate Institute of Linguistics, National Taiwan University
[2.] Institute of Information Science, Academia Sinica, Taipei
[3.] Institute of Linguistics, Academia Sinica, Taipei
Contact: prevot@gate.sinica.edu.tw

Abstract

This paper describes the experiments we are currently conducing at Academia Sinica for extending a proposed international standard for lexical framework (Lexical Markup Framework) and shows the relevance of this work for the Digital Archives Program. Although this framework is very rich and powerful, it originally has been developed for European languages. Hence some important extensions are needed to ensure the robustness of this framework to cope with Asian languages. The issues addressed include: (i) the need for derivational morphology, (ii) the interface between morphology, syntax and semantics with the problem of classifiers, and (iii) representational issues with the richness of the writing systems in Asian languages. In this paper we propose prospective solutions for these issues and explain how lexicon meta-models are useful for digital archives.

Keywords：Lexical resources, Standardization, Asian Languages, Multi-linguality

1. Introduction

This paper describes the experiments we are currently conducing at the Academia Sinica for extending a proposed international standard for lexical framework, the Lexical Markup Framework (LMF) [8] in the context of the NEDO project 'Developing International Standards of Language Resources for Semantic Web Applications' [19] and their potential applications for the Digital Archives Program. This framework is aimed at becoming an ISO international standard, and is already in an advanced development stage (Committee Document voting). LMF framework has been developed on the base of the long standing European initiative of EAGLES [6], and continued with international participation in ISLE [3] that proposed the MILE (Multi-lingual ISLE Lexical Entry). As a natural consequence, this framework is extremely detailed and fitted for European languages (earlier versions of the model have been used for building real-scale lexica for Italian, English, and also benefited from the EuroWordNet [20] experience). However, the fast growing interest for NLP applications in Asian languages, and the crucial issue of massive multi-linguality made clear the need of checking how far the current model is fitted for Asian languages and how to extend or revise it to increase its robustness.

To achieve this goal, our work focuses on the main difficulties of the MLF and MILE frameworks when applied to Asian languages (Mandarin Chinese, Taiwanese, Malay, Bangla, Cantonese). We also propose some tentative solutions to be examined by the colleagues working on other Asian languages (Japanese, Thai) in order to achieve broader consensus and robustness.

The development of such rich, multi-lingual, robust and interoperable lexica is key element in the development of multilingual question-answering and information retrieval systems that can be used for searching the digital archives. More precisely, once the meta-model established it is easy to translate strategic lexical resources such as WordNet [7] or FrameNet [1] in the framework and possible to relate them across languages as demonstrated in [8].

This work with its multilingual aspect and its future as an international standard combined with the Semantic Web applications of the resources built within this framework, made obvious the choice of the W3C RDFS language (Resource Description Framework Schema) [18] extended in OWL (Ontology Web Language) [15] for developing the model. The way was paved by previous work [12] that ported the original MILE model in RDFS. All the experiments described in this paper were conduced under the Protégé ontology development suite [16].

The problems we faced when using the model for Asian languages are described in sections 2 to 5 and our solutions are presented in the section 6. The usefulness of this type of lexicon meta-models for Digital Archives is explained in section 7.

2. Classifiers (CL)

Classifier system exists in many different languages. In general, the location of a classifier is

usually between the determiner and the noun. The cross-lingual examples for such construction are shown in (1). Basically, we can explore and classify the complex semantic concepts of nouns through their classifiers.

(1) determiner + CL+ noun

a. Mandarin
liang2(two) ge5(CL) ren2(person)
"two people"

b. Taiwanese
nng1(two) e5(CL) lang5(person)
"two people"

c. Cantonese
jat1(one) gaa3(CL) ce1(car)
"one car"

d. Bangla
pach(five) kana(CL) boi(book)
"five books"

c. Malay
Dua(two) biji(CL) durian(durian)
"two durians"

Take Mandarin as the example. The classifier system in Mandarin is very rich and complex. The dispute about distinguishing the classifiers and the measure words (MW) has always existed in Mandarin. The first traditional view about classifiers and measure words is to treat classifiers as measure words and vice versa. Another different view thinks that classifiers and measure words are still distinguishable. According to Huang and Ahrens (2001), the method to examine the difference between CL and MW in Mandarin is to insert the genitive *de5* particle between the classifier/measure word and its following noun. The following example (2) demonstrates the difference between the neutral classifier, ge5, and the measure word, tian1. Those examples show the genitive *de5* particle indeed can be inserted in the position between the measure word and its following noun but not for the position between the classifier and its following noun.

(2)

(a) liang2(two) ge5(CL) ren2(person)
*liang2(two) ge5(CL) de5(GEN) ren2(person)
"two people"

(b)
san1(three) tian1(MW:'day') jia4qi2(holiday)
san1(three)tian1(MW:'day')de5(GEN)jia4qi2(holiday)
"three day holidays"

There are 500 classifiers in Mandarin, but less than 200 are widely used. Many references have their own classification for the classifiers. However, the classifiers in Mandarin can be roughly divided into three main types: individuals, kinds, and events. Basically, the classifier systems in Taiwanese and Cantonese are quite similar to the one in Mandarin. However, Cantonese has an exceptional classifier construction that does not exist in Mandarin or Taiwanese. Cantonese classifiers can appear after the pronouns. As shown in (3), when a pronoun is followed by CL+noun, the pronoun becomes possessive. Besides, CL + noun can also come after a noun to express singularity.

(3)

(a)*ngo5* *bun2* *syu1*
I CL book
"my book"

(b) nei5 gin6 saam1
you CL clothes
"your clothes"

(c) lou2 si1 bun2 syu1
teacher CL book
"the teacher's book"

Unlike Taiwanese, Mandarin, Cantonese and Malay, Bangla has default classifiers, such as ta, ti, to, appear in the language as shown in (4).

(4) dui ti/ta kukur
Two CLS dog
"Two dogs"

3. Reduplication

Reduplication serves several functions in the Asian languages. For example, it may be related to the aspects of verbs (Mandarin and Cantonese) as well as quantification such as pluralization (Bangla and Malay) or entirety of features (Malay). Some reduplications may involve part-of-speech changes.

3.1 Mandarin, Taiwanese and Cantonese

Reduplication involving aspects of verbs can be seen in example (5) for Mandarin.

(5)
(a) 想 *xiang3* ('to think')
(b) 想想 *xiang3-xiang3* (tentative aspect)

For 5(b), *xiang3xiang3* refers to the tentative aspect of thinking.

Some reduplications involve part-of-speech changes. One example for Mandarin is in (6) below.

(6)
(a) 慢 *man4*(adj)
(b) 慢慢 *man4man4*(adv) 'slow'

In (6a), *man4* is an adjective where in (6b), *man4man4* has become an adverb. This also occurs in Taiwanese with the example in (7).

(7)
(a) *ban7* (adj)
(b) *ban7ban7* (adv) 'slow'

In Cantonese, *maan6* "slow" has the reduplicated from of *maan6maan2* 'slowly' with a change of tone. In

addition, the adjective *ming4* ‘clear’ has the reduplicated adverb of *ming4ming4* ‘clearly.’ As in Mandarin, this form is associated with ‘tentative’ or ‘delimitative’ aspect.

(8) *dang2 ngo5 tung4 keoi5 king1 king1*
wait I with (s)he talk talk
‘Let me have a chat with him.

In (8), the reduplicated verb *king1 king1* is described to convey a tentative aspect by implying the short duration of the action. However, it somehow links with a longer duration in some cases as in (9).

(9) *zyu2 zyu2 keoi5 laa1*
cook cook it PT
‘Cook it further!

In some cases in Taiwanese, however, verb reduplications may not be always seen as aspect. For instance, example (10a) is not allowed, which should be represented as *siuN7 chit8 e7* ‘think for a while.’ However, the example in (10b) is correct. (10b) can also appear in the form *che7 chit8 e7* (see Cheng [4]).

(10)
(a) siuN7 *siuN7siuN7 ‘think’
(b) che7 che7che7 ‘sit for a while’

As Huang [10] has suggested, Taiwanese reduplication is a lexical process. In example (11), where *khiau2* ‘able, smart’ can be reduplicated, but its synonym *gau3* ‘able’ cannot. These kind gaps are random and can only be encoded in lexicons.

(12)
(a) *khiau2 khiau2khiau2*
‘able, smart'
(b) *gau3 *gau3gau3* ‘able’

As for Cantonese, the nouns classified by certain classifiers may be omitted [13]:

(13) *bun2-bun2 (syu1) ngo*
CL.red (book) I
dou1 soeng2 tai2
all want read
“I want to read all books.”

(14) *go3-go3 (jan4) dou1 jau5*
CL.red (person) all have
bou6 din6nou5gaa3 laa1
CL computer PRT
“Everyone has a computer.”

3.2 Bangla and Malay

In Bangla and Malay, there are basically three types of reduplications – full and rhythmic (for Malay, see [14]).

A Bangla example of full reduplication is seen in (15) below.

(15) *“**besi besi** kore khao”*
*(Eat **a lot**)*

In (15), the meaning of *besi* is ‘more’ but its full reduplicated form *besi-besi* means ‘a lot.’ The same is found in Malay, in (16) below.

(16)
(a) *kawan* ‘friend’
kawan-kawan ‘friends’
***ke**kawan* ‘friends’

(b) *sabut* ‘husk’
sabut-sabut ‘fiber’
*s**er**abut* ‘fiber’

In (16) above, both *kekawan* is the shorter form of *kawan-kawan* and *serabut* is the shorter form of *sabut-sabut*. These shorter forms are sometimes called ‘partial reduplication [14]. The reduplicated morphemes can appear as prefix (as in 16a) or infix (as in 16b).

The other type of reduplication is the rhythmic type. Examples of Bangla are in (17a-b).

(17)
(a) *tapur-tupur* ‘sound of rainfall’
(b) *fit-fat* ‘smart’

In (17a), *tapur* indicates ‘sound of one drop of rain’ but *tapur-tapur* refers to ‘sound of rainfall.’ In (17b), *fit* means ‘fine’ but *fit-fat* means ‘smart.’ Similarly, in Malay, the following examples are given.

(18)
(a) *sayur* ‘vegetable’
*sayur-**mayur*** 'vegetables'
(b) *gunung* ‘mountain’
*gunung-**ganang*** 'mountains'

In (18), the second word (*-mayur* and *-ganang*) are usually not used on its own. It is also worth noting that the choice of reduplication (partial or full) is not random, i.e., only certain words can be fully or partially or even rhythmically reduplicated.

3.3 Functions of Reduplication in the Asian Languages

Reduplication serves various functions in the Asian languages, among which are pluralization, entirety of features (in adjectives), and repeated actions. The followings in (19) show some Malay examples.

(19)
(a) ***Pokok** ini tinggi*
tree this tall
‘This tree is tall.’

(b) Pluralization

Pokok-pokok *di sini tinggi.*
tree.pl. Loc. here tall
'The trees in here are tall.'

(c) Entirety of Feature in Adjective
Pokok di sini ***tingg-tinggi****.*
tree Loc. here tall.Red. entirely
'The trees in here are tall.'[1]

In addition, reduplication does not necessarily occur with nouns and adjectives. It can also occur with verb, as in (20) below.

(20)
(a) *Adik* ***ber-main*** *bola.*
Brother/Sister BER-play ball
'(My) brother/sister is playing ball.'

(b) *Dia* ***ber-main-main***
'3.Nom.Sg. BER-play.Red.
dengan bola itu
with ball that
'He is/was toying with the ball.'

The example in (20b) shows a repeated action of playing (thus, comes the meaning of 'toying'). All the examples above are full reduplication.

The followings in (21) show the overall functions of reduplication in different languages.

(21)

(a) Pluralization

(b) Augmentation (with classifier, see section 3.0).

(c) Entirety

(d) Repeated Actions

The reasons for reduplications various but some are related to emphasis pragmatically (such as the number in pluralization; the action in repeated actions, etc.)

4. Change of POS by affixes

In Chinese, Taiwanese and Cantonese, certain affixes are used to change the part-of-speech of a word, as shown below:

(22)ADJ → ADV
a. Mandarin
you3 xiao4 + de5 → you3 xiao4 de5

b. Cantonese
yau5 haau6 + gam2 → yau5 haau6 gam2
'effective' + AFFIX → 'effectively'

(23)N → V
a. Mandarin
dian4 nao3 + hua4 → dian4 nao3 hua4

b. Taiwanese
tian7 nau2 + boa3 → tian7 nau2 hoa3

c. Cantonese
din6 nou5 + faa3 → din6 nou5 faa3

'computer' + AFFIX ('computerize'

Mandarin (24) and Cantonese (25), but not Taiwanese, also allow time words, such as 'year' and 'day', to be reduplicated to form adverbs with habitual meaning:

(24) wo3 tian1-tian xiang3 ni3.
I day-day miss you
'I miss you everyday.'

(25) ngo5 nin4-nin4 heoi3 taai3 gwok3
I year-year go Thailand
'I go to Thailand every year.'

There are also some affixes in Bangla that change the part-of-speech.

(26) N (ADJ
bipod + janok (bipod-janok)
'danger' + AFFIX ('dangerous')

(27) N (ADJ
jati + iio → jatio
'nation' + AFFIX → national

The change of POS through affixation is a common feature of Malay. Examples are given in (28) below.

(28) hati 'heart' (N)
Ber-hati-hati BER-hear.Red.
'be careful' (V)

Per-hati-an PER-hati-AN
'attention' (N)

In fact, Malay has a rich affixation system which constantly changes POS in derivational forms.

5. Orthography

Many Chinese words have orthography variants. For instance, when the words sheng1(升) and sheng1(昇) are used as verbs and both refer to the concept of "raising," but in certain compound forms, such as liter "公升", is only allowed the sheng1(升) rather than sheng1(昇). The similar situation with 姐 and 姊 that have both have the same pronunciation 'jie3', and they are usually used to call " the elder sister". However, for the compound form, Miss "小姐," only jie3(姐) is allowed.

Using pinyin to replace the real Chinese characters may cause the confusion about distinguishing the words that have the same pronunciation. For example, as shown in (29), there are many different written compound forms for the English word, 'they'. It will

[1] Plural meaning of 'trees' comes from reduplication *tinggi-tinggi* 'all tall'

become very difficult to distinguish them unless the real Chinese characters are seen.

(29) ta1men2“他們(male/neutral)/它們(thing)/她們(female)/牠們(animal)/祂們(god)”
‘they’

Written Cantonese is not used in formal forms of writing. However, written colloquial Cantonese does exist; it is used mostly for transcription of speech, subtitles and informal forms of communication. Therefore, apart from the orthographic variants found in Mandarin, there are more variants for written Cantonese. For instances, 琴/擒日 “yesterday”; 個/果晚 “that night”; 依/宜家 “now”. See [5] for more examples.
Some Cantonese words lack a written form, for examples, *leu1* “to split”, *he3* “to kill time“. This leads to inclusion of English words or “non-standard” Cantonese romanization. In the case of *he3*, it is usually written as “hea”.

6. Handling Asian languages within the lexicon meta-model

Before presenting our extensions to the existing framework for Asian languages, we have to give some details about the starting point. There are actually several versions of this model that are currently compared, and evaluated by instantiating them with various languages. The initial version we worked with is a RDFS implementation of the MILE (*Multilingual ISLE Lexical Entry*) designed by the computational Lexicons Working Group (CLWG) of ISLE (*International Standards for Language)* [3,12]. Two essential features of the framework are its modularity and its inclusion in the Semantic Web by the usage of RDFS and OWL. Based on the same grounds, but distinct, the LMF [8] is being developed with the objective of proposing it at an ISO standard (TC 37 SC 4). The LMF has been developed in XML but not ported in RDFS yet. However these frameworks are very similar. Most of the experiences and extensions of this paper were primarily done on the MILE model. However, lately we coded a significant part of the LMF framework in OWL for benefiting from the best parts of both models.

MILE framework is divided into the semantic, syntactic and morphological layers. While this design was established in [12] its implementation in OWL as three independent modules was remained to be done. It’s what we did first by using the import mechanisms of OWL (See Fig. 1). Equipped with this model, the designer can create lexical databases importing only the layers relevant for the current purposes. Once this done, we started encoding lexical entries from various languages in the model and quickly we faced the issues that we presented in sections 3 to 6.

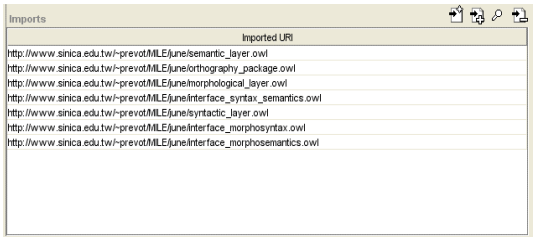


Fig.1 RDFS Import mechanism

6.1 Adding Classifiers

Classifiers were absent in the existing framework. The idea in our proposal is to treat them as first class citizens, having a lexical entry for them but also a semantic unit where we can describe their semantic features. Although our treatment is still preliminary it will be very handy to have information represented in this way for explaining the semantic agreement between the classifier and the noun it classifies (who has himself a set of semantic features) or by using the semantic collocation information provided by the original model.

6.2 A derivational module for the morphological layer

As made clear in sections 3 and 4, Asian languages have important derivational phenomena that need to be handled. An important aspect of the meta-model development is that the model should remain flexible enough to allow the lexicographer to choose between the different possible implementations. More precisely, for handling inflection, one lexicographer might want to enumerate all inflected forms of a given lemmas and associate them with the corresponding morphological features, while another will simply provide the rule for calculating the inflected forms, a third one could decide to enumerate the irregular forms and to provide the inflectional paradigm of the regular forms. This has been done nicely in both MILE and LMF. However, these models are restricted to inflection phenomena.

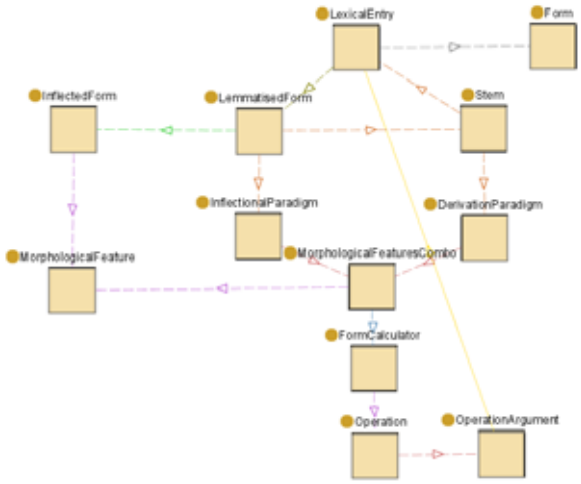


Fig.2 Derivational Module

Technically speaking the derivational morphology phenomena could be described in the current model by using the classes designed for inflection. However, there is a need for distinguishing between inflectional

and derivational morphology as we did (See Fig 2.).

In our diagrams (generated with Jambalaya plug-in under Protégé 3.2), the squares with named labelled with rounds are classes, those labelled with diamonds are instances, the arrows are object properties in RDFS terms (they correspond to the relations in UML).

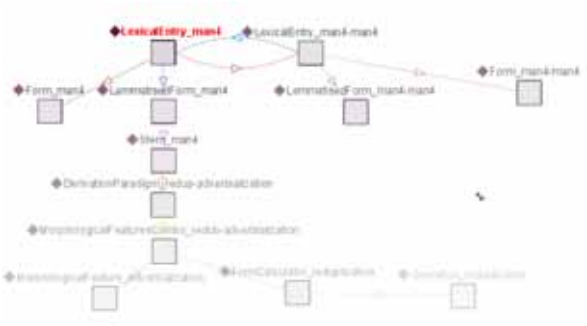


Fig.3 Reduplication treatment

The final modifications to the original model are as discrete as possible but allow to deal with our reduplications (See Fig. 3) and affixes examples (See Fig. 4).[2] It also allowed to keep separate derivation and inflection phenomena. More precisely we (i) added a class DerivationalParadigm related with the Stem, (ii) made generic all the elements that can be shared by inflection and derivation, and (iii) allowed the operation argument to be a lexical entry for capturing the fact that derivational affix can be treated as such.

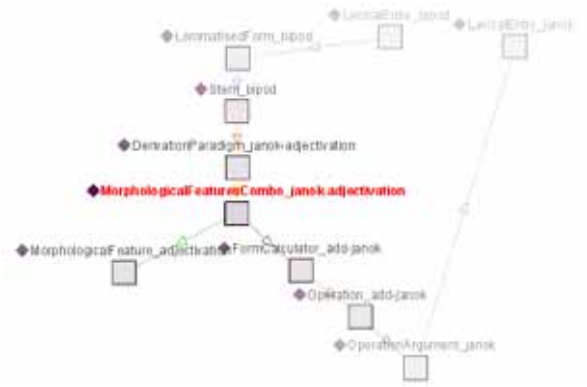
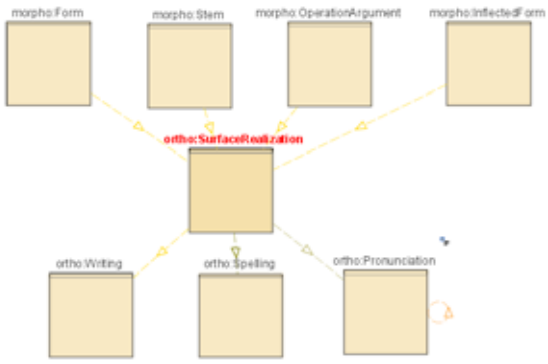


Fig.4 Change of POS through affix treatment

This division between derivation and inflection is not a theoretical choice on our part, and it let the liberty for the lexicographer to handle a phenomenon where he wants. For example, reduplication can manifest features that are considered traditionally as inflectional (e.g. plurality) but another view point could be to treat even this one as derivational on the base of their similarity with other reduplications that are typically derivational (e.g. change of POS).

6.3 An orthography module

As illustrated in section 6, Asian languages have much wider range of orthography, and specially writing, systems than European ones. To deal with this issue we developed an orthography module, as envisioned (but not practically developed to our knowledge) in preliminary ISO meetings distinguishing spelling, writing and pronunciation systems. Each form has a surface realization that can be encoded in these various systems. However, not only form has surface realization but also inflection arguments from the existing MILE and derivation affixes from our derivation extension (See Fig. 5)



. Fig.5 Orthography Module

7. Application to Digital Archives and the web Information Retrieval

Digital Archives important repositories for a various type of information sources. They are key components of the fast growing Semantic Web as they constitute high-quality resources that are very well organized. They constitute therefore the main resources for users to access cultural heritage of their country but also to explore cultures from places geographically remote. To achieve this goal, one of the main strategic lock is the language barrier. Although the archives are interesting to everyone the way to search them and to display the results is improvable, both in terms of quantity and relevance of the retrieved documents and in terms of language accessibility.

Cross Language Information Retrieval (CLIR) and Multi Lingual Information Retrieval (MLIR) can be facilitated by using standard model for the lexicon. Actually, one of the goals of MILE is to facilitate all the HLT application areas, including Information Retrieval. As discussed in earlier sections, MILE frame-work is using the W3C RDF language [18], which is designed for semantic web applications and interoperability.

Most CLIR approaches translate queries into the document language, and then perform monolingual retrieval. There are three main approaches in CLIR and MLIR: using machine translation, a parallel corpus, or a bilingual dictionary [17]. One of the popular and widely used methods of CLIR is dictionary-based approach [2]. Dictionary methods for cross-lingual IR and other IR techniques use Query Terms and Query expansion based on linguistic information contained in Machine-readable dictionaries that can be developed or ported within the meta-model described here.

[2] For seeing the detailed treatment or the examples we presented above and the full OWL instantiation of the model, please see the OWL example file available at the address http://www.sinica.edu.tw/~prevot/MILE/june/

All the information contained in the framework for the lexical entries can be used for query expansions. The most important lexical information contained in the proposed MILE framework that may guide the Query Expansion include morphological and, of course, semantic layers.

The morphological layer helps to expend queries with irregular derived or inflected forms that could not be caught by regular string searches.

However, the main layer concerned with query expansion is the semantic layer. Primarily synonym links, hypernyms and hyponyms and more cautiously all other semantic relations can be used for basic query expansion (See Fig. 6 for an illustration of the MILE semantic layer). In a more prospective way, the semantic features of the queried terms, their semantic frame information or the semantic collocations can be explored.

Fig.6 MILE Semantic Layer

The highly modular MILE architecture is divided mainly into two parts: Mono MILE and Multi MILE. The Mono MILE contains language specific information, whereas the Multi MILE contains cross-language information. Cross language lexical information gathered in MULTI mile is very important to extend the query for multiple languages. Moreover according to the level of development of the resources of a given language, the meta-model allows for a simple mapping that will use the semantic hierarchy (or ontology) developed for the language of the original application or for a richer integration of both languages since the meta-model allows for both direct bilingual links or for truly multi-lingual resources using a Inter-Lingual-Index (ILI) and able to cope with sophisticated multilingual mappings.

As a consequence of this robustness, flexibility and intrinsic interoperable design, the framework enables us to integrate various existing resources of different types and from different languages like WordNet [7], Euro-WordNet [20], and Sinica Bow [11] to facilitate cross-lingual Information Retrieval. Our contribution to the development of such meta-models is an important step toward this next generation of lexical resources, deeply modular, highly interoperable, and massively multi-lingual.

8. Conclusion and Future Work

The proposals presented in this paper are currently discussed among the members of our NEDO project. They will be compared and evaluated against proposals from other members. Ultimately it will contribute in formulating the national member and liaison suggestions to ISO committee about the Lexical Markup Framework (ISO TC 37 SC4).

Finally, our work can be support the on-going work of other members of our group to support develop meaning-based cross-language query system of the NDAP Union Catalogue. Our work, when combined with the above-mentioned work based on WordNet [7] and Sinica BOW [11], will allow queries to be expanded in languages other than English and Chinese.

Acknowledgements

We would like to thank the people involved in this project in Taipei, Katarzyna Horszowska, and Yong-Xiang Chen, the NEDO meeting participants and the NEDO project members that answered many questions regarding the MILE and LMF frameworks. We remain entirely responsible for the problems remaining in the current version and for the potential residual misunderstanding of the original model. This research was carried out through financial support provided under the NEDO International Joint Research Grant Program (NEDO Grant)

References

[1] Baker, C.F, C. Fillmore, C. & J.B. Lowe. (1998) The Berkeley FrameNet project. *Proceedings of the COLING-ACL.*

[2] Ballesteros, L. & B Croft. (1996) Dictionary Methods for Cross-Lingual Information Retrieval", *in Proc. of the 7the DEXA Conference an Database and Expert Systems Applications*, pp. 791-801.

[3] Calzolari, N., 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.

[4] Cheng, R. L (1988) Semantic and Grammatical Features of Reduplicated Verbs in the Taiwan Dialect, *Zhongguo Yuwen*, No. 6 439-444, 1988 (In Simplified Chinese)

[5] Cheung, L.Y.. (1983). (In Chinese) "A Total Count of Cantonese Syllables with no character representations", Yuwen Zazhi 10: 28-35.

[6] Eagles project: http://www.ilc.cnr.it/EAGLES/home.html

[7] Fellbaum, C. (ed.) (1998) Wordet, a lexical database. The MIT Press.

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

[9] Huang, C. and K. Ahrens.(2003) Individuals, kinds and events: classifier coercion of nouns, Languages Sciences, 25:353-373.

[10] Huang, C. R. (1992) Adjectival Reduplication in Southern Min: A Sutdy of Morpholexical Rules with Syntactic Effects, Chinese Language and Linguistics, Vol. 1, 407-422.

[11] Huang; C., R. Chang. & S. Lee. (2004) Sinica BOW (Bilingual Ontological Wordnet): Integration of Bilingual WordNet and SUMO *4th International Conference on Language Resources and Evaluation (LREC2004).*

[12] Ide, N.; A. Lenci, & N. Calzolari. RDF instanciations of ISLE/MILE lexical entries (2003) Proceedings of the ACL'03 workshop on Linguistic annotation: getting the model right.

[13] Matthews, S. & V. Yip. (1994). Cantonese: a Comprehensive Grammar. London: Routledge.

[14] Nik Safiah Karim, Farid M. Onn, Hashim Hj. Musa and Abdul Hamid Mahmood. (1997). Tatabahasa Dewan. Edisi Baharu. Kuala Lumpur: Dewan Bahasa and Pustaka. [Grammar reference]

[15] OWL overview: http://www.w3.org/TR/owl-features/

[16] Protégé: http://protege.stanford.edu/

[17] Qin, Jialun, Yilu Zhou, Michael Chau, Hsinchun Chen (2003) Supporting Multilingual Information Retrieval in *Web Applications: An English-Chinese Web Portal Experiment 6th International Conference on Asian Digital Libraries, ICADL*

[18] RDFS: http://www.w3.org/TR/rdf-schema/

[19] Tokunaga, T.; V. Sornlertlamvanich; T. Charoenporn; N. Calzolari,; M. Monachini; C. Soria; C. Huang; Y. Xia; H. Yu; L/ Prevo & K. Shirai. (2006) Infrastructure for standardization of Asian language resources. Proceedings of ACL-COLING.

[20] Vossen, P. (ed.) (1998) EuroWordNet: a multilingual database with lexical semantic networks. Kluwer Academic Publisher.