# Meta-modeling and standardization issues for Asian Languages lexical resources

**Laurent Prévot, Chu-Ren Huang, Kamrul Hasan, Sophia Lee, I-Li Su**
Institute of Linguistics,
Academia Sinica, Taipei
*Contact: prevot@gate.sinica.edu.tw*

**Siaw-Fong Chung**
Graduate Institute of Linguistics,
National Taiwan University, Taipei

**Tian-Jian,Jiang**
Institute of Information Science
Academia Sinica, Taipei

## Abstract

This paper describes the experiments we are currently conducing at the Academia Sinica for extending an international lexical framework (Lexical Markup Framework). Although very rich and powerful, this framework has been first developed for European languages. This paper focuses on some important extensions that need to be included for ensuring the capacity of this framework to cope with Asian languages. The main problems addressed concern (i) morphological issues with the need for derivational morphology, the interface between morphology, syntax and semantics with the problem of classifiers and more representational issues with the richness of the writing systems in Asian languages. In this paper we propose prospective solutions for these issues and illustrates them through examples from Chinese Mandarin, Taiwanese, Cantonese, Malay and Bangla.

## 1 Introduction

This paper describes the experiments we are currently conducing at the Academia Sinica for extending an international lexical framework, the Lexical Markup Framework [1] in the context of the NEDO project 'Developing International Standards of Language Resources for Semantic Web Applications' [2]. This framework concerns machine-readable lexical resources or computational lexicon and can be used for developing simple lexicon or extremely rich ones detailing syntactic behavior, morphological aspects and semantic information. This framework is aimed at becoming an ISO international standard, and is already in an advanced development stage (CD voting). LMF framework has been developed on the base of the long standing initiative of EAGLES [3], continued into ISLE [4] European. 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 lexicon for Italian, English, and also benefited from the EuroWordNet [5] 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 in order to cope with them.

On this ground, our work consists in pointing the main difficulties of the MLF and MILE frameworks for some Asian languages (Mandarin Chinese, Malay, Bangla, Cantonese, Taiwanese) and in proposing some tentative solutions to be examined with the colleagues working on other languages (Japanese, Thai). These problems concern the different aspects of the framework. More

precisely we will address the complex issue of classifiers in the section 2, morphological issues with reduplication and derivational affixes in section 3 an 4, and more representational issues with the diversity of writing systems (compared to European languages) of Asian languages (section 5). Finally we will show how, we solved these problems in extending and revising (as few as possible) the current proposals for the framework (section 6).

The global aspect of 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) [6] extended in OWL (Ontology Web Language) [7] for developing the model. The way was paved by previous work [8] that ported the MILE model in RDFS. We initially started with this version, and later on update with parts of the LMF model (developed in UML). In section 6 we will detail some differences between LMF and MILE and we dealt with them. All the experiments described in this paper were conduced under the Protégé ontology development suite Protégé [9].

## 2 Classifiers

Classifiers in Chinese can be mainly divided into three types: individuals, kinds, and events. As mentioned in [10], the usage of individual classifiers is to indicate the salient features of their nouns behind. The neutral classifier in this type is *ge5*. For instance, in (1a-b), the classifier *tiao2* classifies for long, cylindrical, flexible objects, but it can be replaced by the neutral classifier *ge5*.

| | | | |
|---|---|---|---|
| (1a) | yi1 | tiao2 | xie2dai4 |
| | one | CL.ind | shoelace (one shoelace) |
| (1b) | yi1 | tiao2 | man2yu2 |
| | one | CL.ind | eel (one eel) |

The selection of using kind classifiers is delimited by the class of a certain nouns. *zhong3(kind)* is the neutral classifier commonly used in this type and many kind classifiers can be replaced by *zhong3*. The examples of for this type are shown in (2a-b).

| | | | |
|---|---|---|---|
| (2a) | zhe4 | lei4 | dong4wu4 |
| | this | CL.kind | anima (this kind of animal)l |
| (2b) | liang3 | kuan3 | biao3 |
| | two | CL.kind | watch (two designs of a watch) |

‘two designs of a watch’

As mentioned in [10], when a bare NP is type-shifted to represent an event, the event meaning is coerced by an event classifier and then the event type is settled. As shown in (3a) and (3b), the event classifier *ci4* classifies for “the frequency/times of event” and *dun4* classifies for “the process of a meal”, so the event meanings of (3a) and (3b) are coerced because of the event classifiers.

| | | | | | |
|---|---|---|---|---|---|
| (3a) | chi1 | le5 | liang3 | ci4 | fan4 |
| | eat | ASP | two | CL.event | rice (have meals twice) |
| (3b) | chi1 | le5 | liang3 | dun4 | fan4 |
| | eat | ASP | two | CL.event | rice (have two meals) |

The above examples demonstrate the difference between the neutral classifier, ge5, and the measure word, tian1. They also 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.

Unlike Mandarin Chinese, Cantonese classifiers also come after pronouns. When the classifier comes right after the pronoun, it means singular, as in (3). When the classifier comes right after the pronoun, it means singular.

| | | | |
|---|---|---|---|
| (4) | *ngo5* | *bun2* | *syu1* |

| | | | |
|---|---|---|---|
| | my | CL.book | book (my book) |

Indo-European languages are generally not featuring classifiers, but Bangla is atypical on this aspect. Every noun in Bangla must have its corresponding classifier when used with a numeral or quantifiers (e.g 5a-c). however the number of classifiers in Bangla does not compare to Mandarin Chinese.

| | | | |
|---|---|---|---|
| (5a) | dui | ti/ta | kukur |
| | two | CL.generic | dog |
| (5b) | pach | kana boi | |
| | five | CLS.book | book |
| (5c) | dosh | jon | manus |
| | Ten | CLS.human | man |

To sum-up, classifiers although exhibiting significant differences across languages, also present some patterns like their systematic appearance between the determiner and the noun. Also recurrent is the semantic feature they carry, scholars might disagree on the exact force of the feature: does it coerces the classified object or is it simply a matter of semantic agreement? In any case this semantic aspect has to be represented in the lexicon and it is something that is currently missing in the lexical framework as we will see in section 6.1.

## 3 Reduplication

Reduplication is a derivational process in Chinese. There are two types of such reduplication. The first one is to simply repeat the word and form a new word, but normally the new word has a different part-of-speech, as shown in (6a) and (6b).

(6a) 慢 (man4) ”slow” 慢慢 (man4-man4) ”slowly”

(6b) 想(xiang3) ”to think” 想想 (xiang3-xiang3) “tentative aspect”

The cases where we find in Mandarin Chinese apply to Cantonese as well. For example, *maan6* “slow” → *maan6-maan2* “slowly” with a change of tone; ming4 “clear” → *ming4-ming4* “clearly”. It is generally associated with ‘tentative’ or ‘delimitative’ aspect as in (7) where reduplicated verb is described to convey a tentative aspect by implying the short duration of the action.

| | | | | | |
|---|---|---|---|---|---|
| (7) | *dang2* | *ngo5* | *tung4* | *keoi5* | *king1-king1* |
| | wait | I | with | (s)he | talk.redup |
| | ‘Let me have a chat with him. | | | | |

An interesting feature of Cantonese is the reduplication of classifiers. A classifier can be reduplicated to express quantification (e.g. 8). Also interesting is the ellipse of the classified noun (e.g. 8), the interpretation of the sentence reposing entirely on the predicate restrictions and on the semantic contribution of the classifier (See [11] for more precisions.)

| | | | | | | |
|---|---|---|---|---|---|---|
| (8) | *bun2-bun2* | *(syu1)* | *ngo* | *dou1* | *soeng2* | *tai2* |
| | CL.redup | (book) | I | all | want | read |
| | “I want to read all books.” | | | | | |

In Bangla, reduplication can be used for expressing a wide range of phenomena, see examples (9a-b):

(9a) *“besi”* (more) *“besi-besi”* (a lot).

(9b) tapur (sound of one drop of rain) tapur-tapur (sound of rainfall)

Reduplication has several functions in Malay, among which are pluralization, entirety of features (in adjective), and repeated action. Examples are in (10a-c).

(10a) ***Pokok*** *ini* *tinggi*

Tree this tall (This tree is tall)

(10b) ***Pokok-pokok*** *di* *sin* *tinggi.*

tree.pl Loc here tall (‘The trees in here are tall.’)

(10c) *Poko* *di* *sin* ***tingg-tinggi***

Tree Loc. Here tall.Red. (‘The trees in here are tall).

Reduplication does not necessarily occur with nouns and adjectives. It can also occur with verb, as in (11) below.

(11a) *Adik* ***ber-main*** *bola.*

Brother/Sister BER-play ball (My) brother/sister is playing ball.’

(11b) *Dia* ***ber-main-main*** *denga* *bola* *itu*

‘3.Nom.Sg. BER-play.Red. with ball that

‘He is/was toying with the ball.’

The example in (2b) shows a repeated action of playing (thus, comes the meaning of ‘toying’). All the examples above are full reduplication. In Malay, there are two other types of reduplication – partial reduplication ((12) below) and rhythmic reduplication ((4) below). (Nik Safiah Karim et al., 1994)[21]

(12a) *kawan* ‘friend’

*kawan-kawan* ‘friends’

***ke****kawan* ‘friends’

(12b) *gunung* ‘mountain’

*gunung-****ganang*** 'mountains'

In (12b), the second word (*-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.

To sum-up, reduplication has various functions both inflectional and derivational. The POS of the reduplicated element has a special importance for determining this function but by itself might not be sufficient to explain the wide range of phenomena we are facing.

## 4 Change of POS by affixes

In Chinese, using the affixes, such as *de5* or *di5*, can change the original part-of- speech of a word to adjective or adverb. Similar to Mandarin Chinese anCantonese, some adjectives can be prefixes of verbs to become compound adjectives.

Bangla has several affixes that change the part-of-speech exemplified in (13a-c):

(13a).“bipod” (Danger ) + “janok”(father) = “bipod-janok” (dangerous)

(13b)“Jati” (nation) + “iio” (a suffix)= “jatio” (national)

(13c)“Mittha”(lie) + “uk”(a suffix) = “mithhuk”(liar)

The change of parts-of-speech through affixation is also a common feature of Malay. Examples are given in (14) below. Malay has a rich affixation system which constantly changes parts-of-speech in derivational forms.

(14) *hati* ‘heart’ (Noun)

***Ber-****hati-hati* BER-hear.Red. ‘be careful’ (Verb)

***Per-hati-an*** PER-hati-AN ‘attention’ (Noun)

## 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 is shown in (9). 姐 and 姊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 (11), there are many different written compound forms for the English word, they. It will become very difficult to distinguish them unless the real Chinese characters are seen.

(11) 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”; o個/果晚 “that night”; 依/宜家 “now”. See [12] 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 the issues raised by Asian language with MILE

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)* [4,8]. 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 [2] 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 [8] 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, we can create lexical databases (containing only instances) importing only the layers there we are interested in.

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. The remaining of this paper will be therefore devoted to the description of our proposals for handling these issues in the framework.

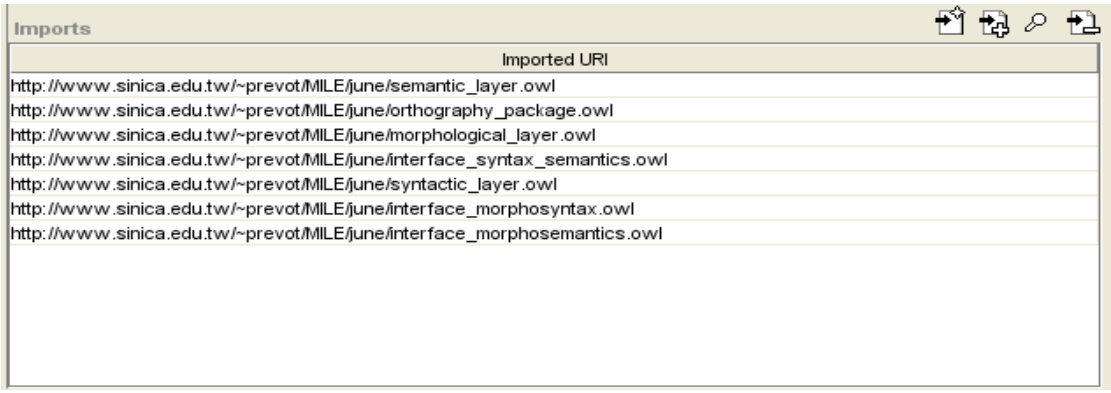


Fig.1 RDFS Import mechanism

### 6.1 Adding classifiers

Classifiers were absent in the existing framework. The idea in our proposal is to treat them as first class citizens (See Fig 2), 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 also by using the semantic collocation information provided by the original model.

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).

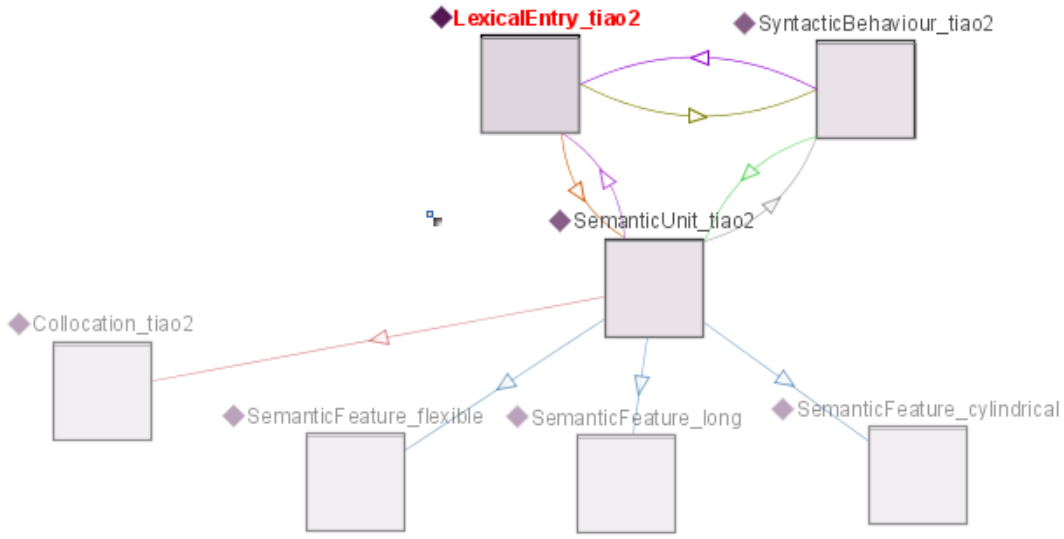


Fig. 2 Partial instantiation for a classifier

### 6.2 Adding derivational morphology

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.

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:

1. Inflection and derivation are usually described as two separate components of a morphological theory. Although it is not possible to draw a sharp line dividing the two types of morphological operation, there are at least two differences: (i) inflection does not change grammatical category, while derivation typically does, and (ii) inflection is usually peripheral to derivation.

2. Inflectional affixes are not lexical entries while derivational affixes can be.

3. Asian linguists do not consider many of their morphological phenomena as inflectional but as clearly derivational.

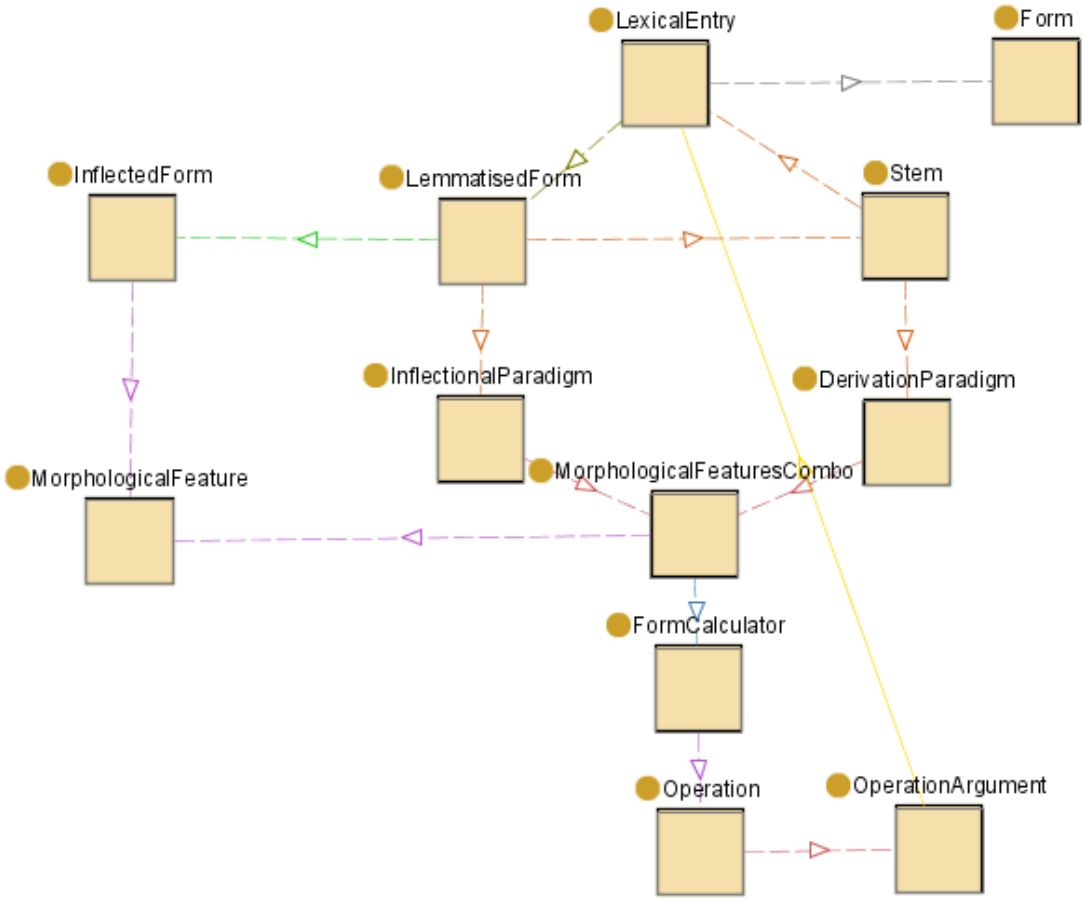


Fig.3 New morphological layer.

As a consequence we developed a solution in two steps. First, we first created a full derivational module duplicating most of the classes defined in the original inflectional paradigm. Then, while testing this solution, we realized that once the morphological operations and their arguments were defined in a generic way, most of the inflectional model could be reused in the derivational model (See Fig. 3). The final modifications were therefore minimal but allowed to deal with our reduplications (See Fig. 4) and affixes examples (See Fig. 5). It also allowed to keep separate derivation and inflection. More precisely we (i) added a class DerivationalParadigm related with

the Stem, (ii) made generic all the element 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 (as in the example of the Fig. 5).

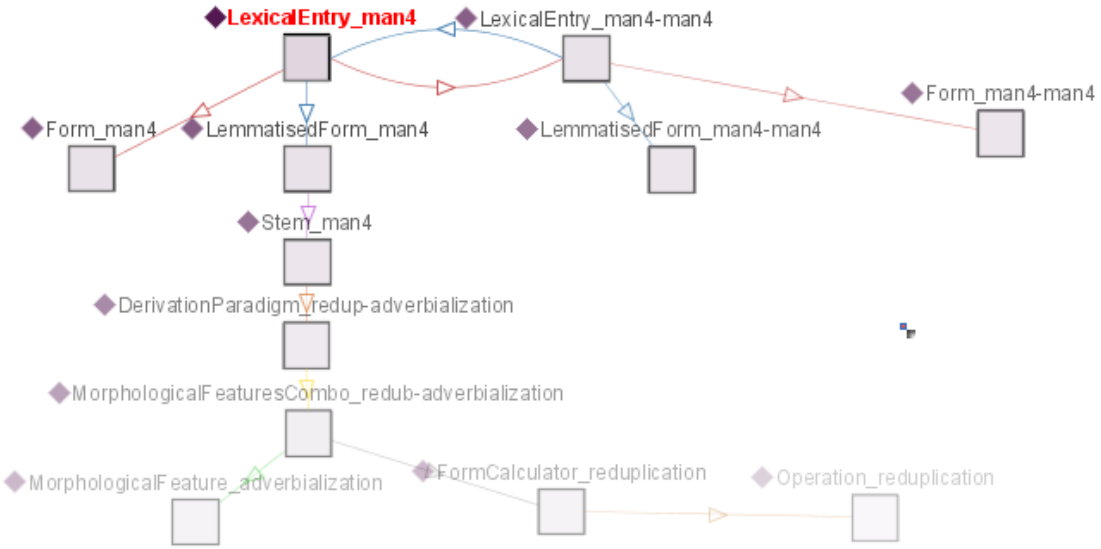


Fig.4 Reduplication example.

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).

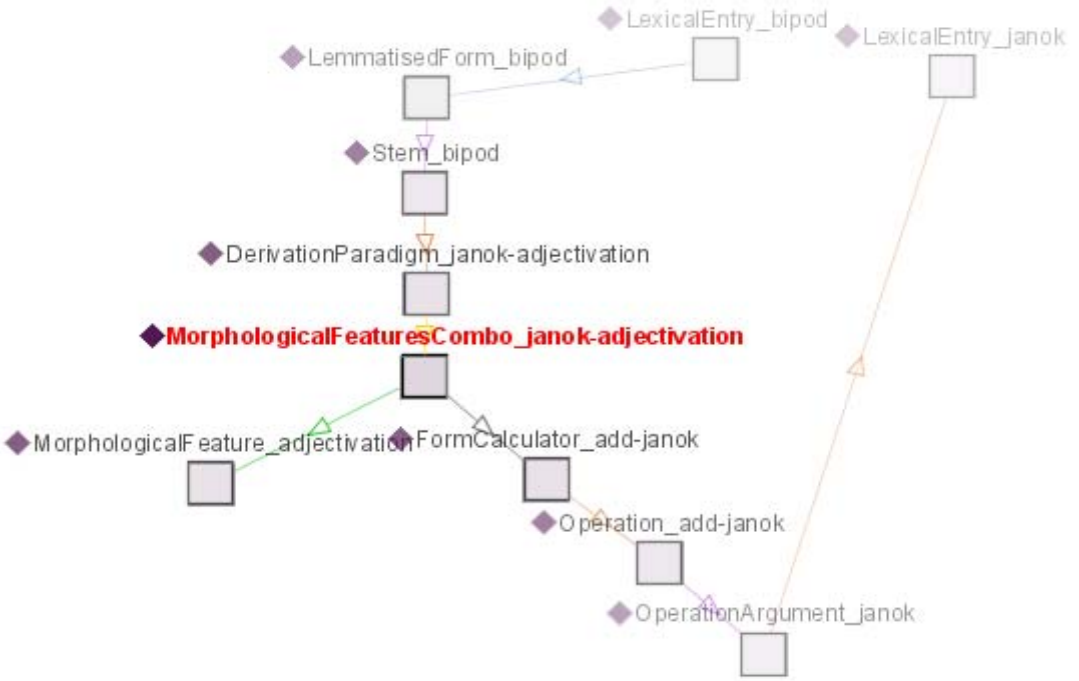


Fig.5 Derivational Affix example.

### 6.3 Handling writing systems

The uses of a rich set of writing systems in Asian languages needs also to be addressed. In the LMF framework, this issue is handled by the class RepresentationFrame that is aggregated under the class Form. Our proposal follow this idea, and follow the proposed distinction between spelling, pronunciation and writing system proposed in the LMF draft but instead of only having Representation Frame for the forms we have for anything that has a surface realization, including for example the arguments of inflectional operation (See Fig. 5)

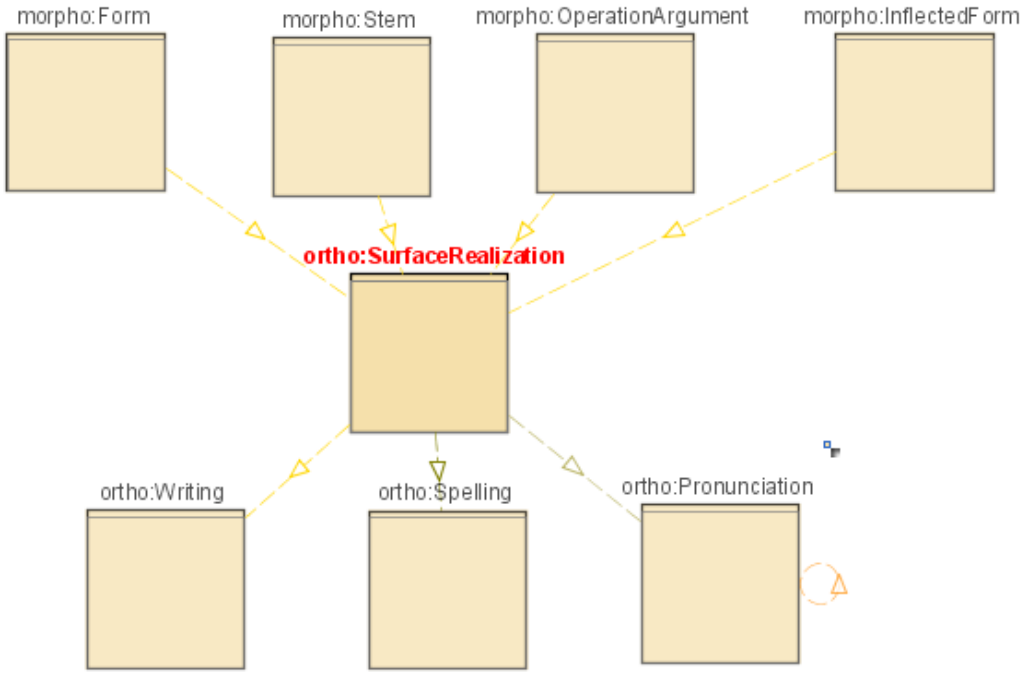


Fig.5 Proposal for an orthography module.

For accessing the detailed treatment of many of the examples presented in sections 3-5 as well as for accessing and using the full model please visit the following address: *http://www.sinica.edu.tw/~prevot/MILE/june/* .where all the owl files are stored.

## 7 Conclusion and Future work

The proposals presented in this paper are currently discussed among the members of the NEDO, compared and evaluated against the other proposals emanating from the other members. Ultimately it will contribute in formulating the Asian chapter suggestions to ISO committee about the Lexical Markup Framework (ISO TC 37). This work is therefore continuously evolving on the base of our experiments and of the comments from other teams involved.

About the future work we plan to encode more examples in order to check empirically the model. In particular we are currently inputing the Swadesh list in the model. Another important issue is the mapping of the model with already existing lexical formats in order to facilitate the encoding of existing resources (e.g WordNet, FrameNet,.. as explained in [1]) in the standardized framework.

## Acknowledgments

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

[2] Takenobu, T.; Sornlertlamvanich, V.; Charoenporn, T.; Calzolari, N.; Monachini, M.; Soria, C.; Huang, C.; YingJu, X.; Hao, Y.; Prevot, L. & Kiyoaki, S. (2006) Infrastructure for standardization of Asian language resources. Proceedings of ACL-COLING.

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

[4] 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.

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

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

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

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

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

[10] Chu-Ren Huang, Kathleen Ahrens.(2003) Individuals, kinds and events: classifier coercion of nouns, Languages Sciences, 25:353-373.

[11] Killingley, S.Y. (1983). Cantonese Classifiers: Syntax and Semantics. Newcastle upon Tyne. Grevatt and Grevatt.

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