Adapting International Standard for Asian Language Technologies

Tokunaga Takenobu[†], Dain Kaplan[†], Chu-Ren Huang[‡], Shu-Kai Hsieh[‡], Nicoletta Calzolari[§], Monica Monachini[§], Claudia Soria[§], Shirai Kiyoaki[¶], Virach Sornlertlamvanich[*], Thatsanee Charoenporn[*], Xia YingJu[◇]

[†]Tokyo Inst. of Tech., Tokyo Japan, {take,dain}@cl.cs.titech.ac.jp
[‡]Academia Sinica, Taipei Taiwan, {churenhuan,shukai}@gmail.com
[§]ILC/CNR, Pisa Italy, {glottlo,monica,soria}@ilc.cnr.it
[¶]JAIST, Ishikawa Japan, kshirai@jaist.ac.jp
[*]TCL/NIST, Bangkok Thailand, {virach,thatsanee}@tcllab.org
[◇]Fujitsu R&D Center, Beijing China, yjxia@cn.fujitsu.com

Abstract

Corpus-based approaches and statistical approaches have been the main stream of natural language processing research for the past two decades. Language resources play a key role in such approaches, but there is an insufficient amount of language resources in many Asian languages. In this situation, standardisation of language resources would be of great help in developing resources in new languages. This paper presents the latest development efforts of our project which aims at creating a common standard for Asian language resources that is compatible with an international standard. In particular, the paper focuses on i) lexical specification and data categories relevant for building multilingual lexical resources for Asian languages; ii) a core upper-layer ontology needed for ensuring multilingual interoperability and iii) the evaluation platform used to test the entire architectural framework.

1. Introduction

Natural language processing research on Asian languages has been thriving in recent years. The ALR workshop series (2001–)[1], a special track and a panel dedicated to Asian languages in COLING/ACL 2006, as well as special double issues of *Journal of Language Resources and Evaluation* (Huang and Tokunaga, 2006) are good evidence of these flourishing research activities.

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 used are less language specific than classical rule-based approaches where a human analyses the behaviour of target languages and constructs rules manually. 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 a situation, standardisation of language resources would be of great help in developing resources in new languages.

Against such background, we have launched a three year project, funded by NEDO, to create a common standard for Asian language resources that adapts an international standard (Tokunaga and others, 2006). Four research items are addressed by the project, (1) building a description framework of lexical entries, (2) building sample lexicons, (3) building an upper-layer ontology and (4) evaluating the proposed framework through an application. This paper presents the latest developments of the project, focusing in particular on i) lexical specification and data categories relevant for building multilingual lexical resources for Asian languages; ii) a core upper-layer ontology needed for ensuring multilingual interoperability and iii) the evaluation platform used to test the entire architectural framework.

[1]http://www.cl.cs.titech.ac.jp/alr

2. Lexical specification

Our lexical specification 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[2] and ISO TC37/SC4[3]. 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 contributed to the finalisation of the LMF draft revision 14[4] including (1) a package for derivation-

[2]http://lirics.loria.fr/

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

[4]LMF is now in the Final Draft of International Standard

al morphology, (2) the syntax-semantic interface with the problem of classifiers, and (3) representational issues with the richness of writing systems in Asian languages.
As a proof-of-concept of the conceptual framework, a first version of the NEDO 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 NEDO 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.

3. Data categories

The activity of designing a high-level conceptual model for harmonised lexicons in our project 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 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. A specific purpose of the NEDO project 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 currently 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.

(FDIS) status. It is expected to be International Standard in 2008.

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 เด็ก (dek0) 'child' has a reduplication form เด็กๆ (dek0-dek0) 'children'.)

(b) Generalisation (to express a vague sense of a word, for example ดำ (dam0) 'black' has a reduplication form ดำๆ (dam0-dam0) 'blackish'.)

(c) Intensification (to express a higher degree of modification, for example มืด (mued2) 'dark' has a reduplication form มืดๆ (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 คิด (khid3) 'think' can be reduplicated to form คิดๆ (khid3- khid3) 'think longer'. In this case, thinking for a ceartain period of time implies deliberate thinking.)

(e) Individualisation (to express individual from the generic group, for example ตัว (tua0:classifier) 'one' has a reduplication form ตัวๆ (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 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 alleviate 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. The correspondences between our universal data categories and the categories for the honorific systems of Thai and Japanese are shown in Table 1.

Universal DC	Thai	Japanese
Respectful	(c) Respectful	(a) Respectful, (b) Humble/3rd persons, (c) Humble/hearer
Polite	(d) Polite	(d) Polite
Diction for special social strata	(a) Diction for the King, (b) Diction for religious figures	(f) Diction for the Imperial Family
Other		(e) Beautification

Table 1: Proposal of Universal DC for Honorific

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 correspondence between the writing systems and the combinations of the attributes is summarised as Table 2. The

Writing system	scriptName	orthographyName
Hiragana	hiragana	-
Katakana	katakana	-
Kanzi	kanzi	-
Romanisation	latin	Japanese style kunrei style Hepburn style

Table 2: Japanese writing systems

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. Figure 1 shows an example of the entry ''tizirege'' (curly hair).

4. Multilingual resources

We are constructing 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

[5]http://www.globalwordnet.org/

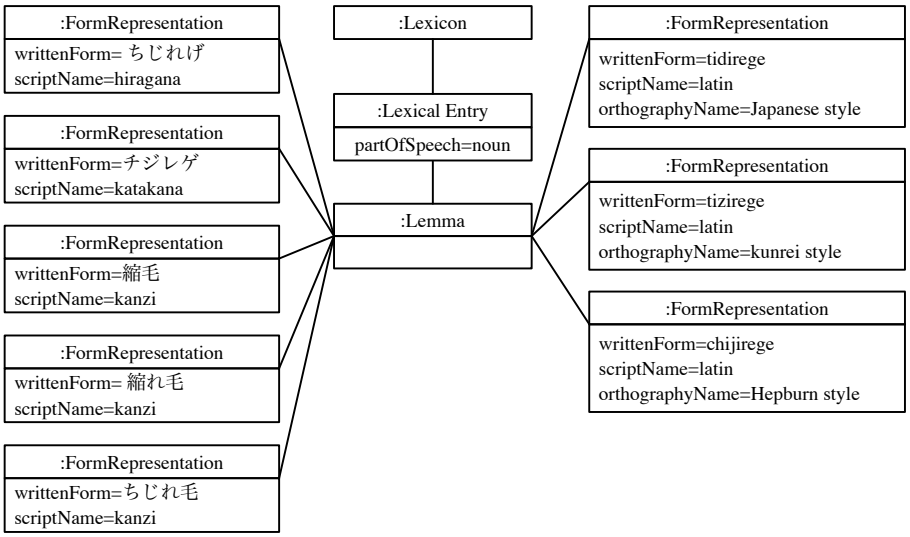

Figure 1: Example of a Japanese entry with multiple scripts

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.

5. Evaluation platform

We are evaluating our research results on a multilingual information retrieval system which is under development. 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 by this project, 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 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 (refer to section 4.) 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. As a concrete example, let us continue using the example from above. The lexical entry for ''ticket'' might contain a semantic type definition something like in Figure 3.

```
<LexicalEntry ...>
  <feat att="POS" val="N"/>
  <Lemma>
    <feat att="writtenForm" val="ticket"/>
  </Lemma>
  <Sense ...>
   <feat att="semanticType" val="ARTIFACT"/>
    ...
  </Sense>
  ...
</LexicalEntry>
```

Figure 3: Lexical entry for ''ticket''

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 ''AR-

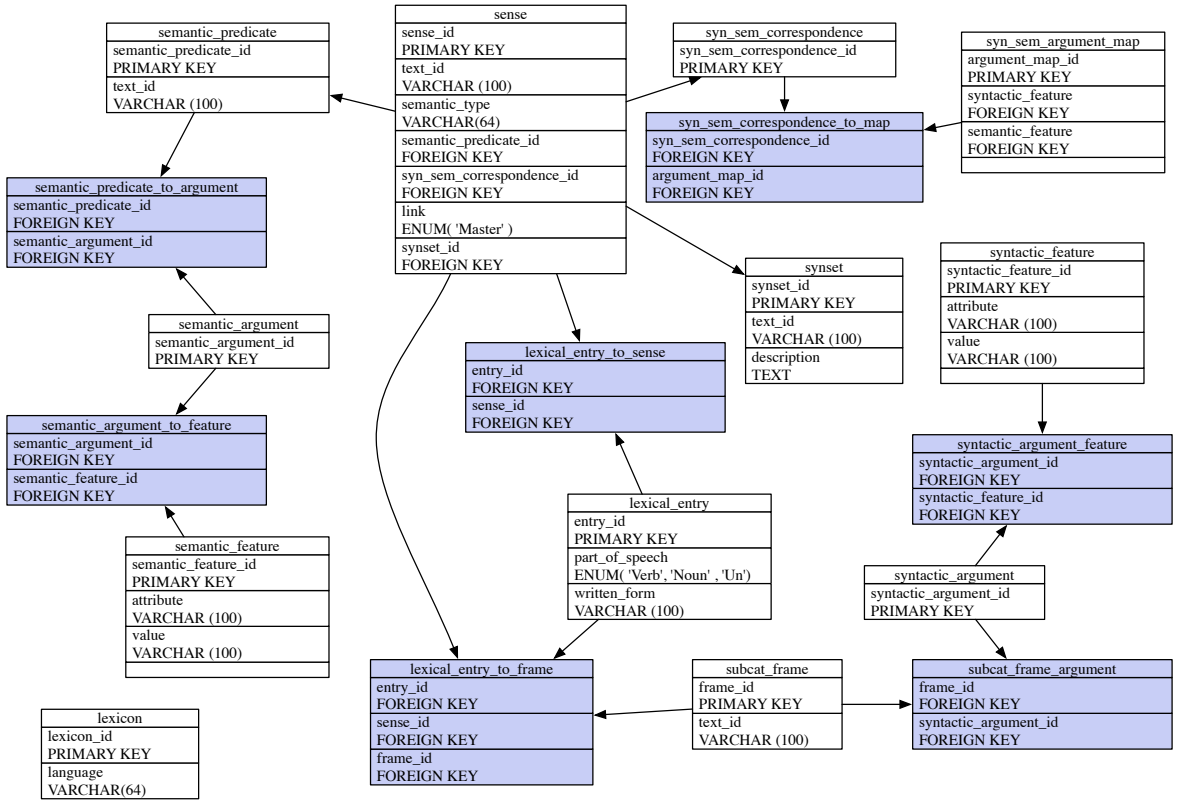

Figure 2: Database schema

```
<LexicalEntry ...>
  <feat att="POS" val="V"/>
  <Lemma>
    <feat att="writtenForm" val="sell"/>
  </Lemma>
  <Sense id="sell-1" ...>
    <feat att="semanticType"
          val="Transaction"/>
    <PredicativeRepresentation
      predicate="pred-sell-1"
      correspondences="map-sell1">
  </Sense>
</LexicalEntry>

<SemanticPredicate id="pred-sell-1">
  <SemanticArgument ...>
    ...
    <feat att="restriction" val="ARTIFACT"/>
  </SemanticArgument>
</SemanticPredicate>
```

Figure 4: Lexical entry for ‘‘sell’’ with its semantic predicate

TIFACT’’ as shown in Figure 4, and then any lexical entries that refer to these predicates are 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 in Figure 2 is as the system stands today. It 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.

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 to the small size of our sample lexicon, data sparsity is naturally an issue, but hopefully 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.

This new type of expansion requires rich lexical information, but we are expecting that once this information is described within our framework for each language, the

expansion mechanism will work regardless of language. The information retrieval system would be a good touchstone to show the portability of systems with the lexical information described with our framework.

6. Concluding remarks

This paper outlined a project for creating a common standard for Asian language resources in cooperation with other initiatives. We start with three Asian languages, Chinese, Japanese and Thai, on top of an existing framework which was designed originally for European languages. We plan to distribute our draft 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 Language Resource Network project[7]. We believe our efforts contribute to international activities like ISO-TC37/SC4.

Acknowledgment

This research was carried out through financial support provided under the NEDO International Joint Research Grant Program (NEDO Grant).

7. References

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

Shu-Kai Hsieh, Su I-Li, Chu-Ren Huang, Hsiao Pei-Yi, Kuo Tzu-Yi, and Laurent 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.

Chu-Ren Huang and Takenobu Tokunaga. 2006. Asian language processing: State of the art resources and processing. *Journal of Language Resources and Evaluation*, 40(3-4). Special issue.

Chu-Ren Huang, Laurent Prevot, and Su I-Li. 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)*.

Kiyoaki Shirai, Takenobu Tokunaga, Chu-Ren Huang, Shu-Kai Hsieh, Ivy Kuo, Virach Sornlertlamvanich, and Thatsanee Charoenporn. 2008. Constructing taxonomy of numerative classifiers for Asian languages. In *Proceedings of the third International Joint Conference on Natural Language Processing (IJCNLP 2008)*, pages 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, pages 452–463.

Takenobu Tokunaga et al. 2006. Infrastructure for standardization of Asian language resources. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 827–834.

YingJu Xia and Hao Yu. 2007. Dimensionality reduction with parallel corpora. In *IADIS European Conference on Data Mining*, pages 113–118, Lisbon, July.

[6] http://www.afnlp.org/

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