A Proposed Model for Constructing a Yami Wordnet

Meng-Chien Yang[1], D. Victoria Rau[2], Ann Hui-Huan Chang[2]
[1] Providence University, Taichung, Taiwan ROC
[2] Natioanl Chung Cheng University, Chiayi, Taiwan ROC
mcyang2@pu.edu.tw, lngrau@ccu.edu.tw, ann-ronald@msa.hinet.net

Abstract

This paper describes an attempt to build a lexical database for the Yami language, an Austronesian endangered language. As the Yami language documentation and conservation projects have produced substantial corpora, we are now ready to construct the Yami online knowledge database based on the knowledge we have accumulated in the language.
In this paper, we propose a model to build the WordNet-like Yami lexical semantics and database. The model is first described in detail, followed by an illustration of how to construct an ontology for Yami fishery at the implementation phase.

Keywords

Wordnet; ontology; Yami; Formosan Language.

1 Introduction

An online lexical database for a specific language is an invaluable resource for many research projects and web and natural language processing applications. For example, WordNet (Fellbanum 1998), a lexicon database for English language produced by Princeton University, has brought about many new applications and research projects(Smith et al.,2004). Many researchers are following the WordNet experience to produce online lexicon databases for other languages (e.g., EuroNet). Most of these research projects started with large online corpora and distilled the lexical entries from the contents of the corpora. As a result, many major languages have many archives and documents from the past to support their semantic analyses (Kovecses 2005).

Creating an online lexical database for an endangered language, on the other hand, is a totally different story. An endangered indigenous language is used by a small group of people and is likely to be extinct in the next decade or so. It is a very challenging task to collect a comprehensive corpus for the endangered language, as the scarce resources and lack of archives present a big problem in establishing a sizable corpus. However, many organizations, such as SIL, Dobes and UNESCO, have sponsored projects to help researchers carry out endangered language documentation (Byrne 2009). Since 2005, our research team has been involved in documenting Yami (Tao), a Philippine Batanic language spoken by 3000 people on Orchid Island off the southeast coast of Taiwan. We have published our results on the following three web sites:

http://yamiproject.cs.pu.edu.tw/yami/, http://yamiproject.cs.pu.edu.tw/elearn,

http://yamobow.cs.pu.edu.tw/.

Having collected a substantial number of Yami texts for our corpus, we have produced an online dictionary and created an ontology of fish (Rau et al., 2009; Rau and Yang 2009). The reasons to bootstrap our study of Yami semantics are the following:

1. Studying indigenous knowledge in the Yami language is important for understating their knowledge of Taiwan ecological environments.
2. Semantic classifications of Yami words will provide valuable information in the study of other Austronesia languages in Taiwan.
3. Classifications of the word senses in Yami is important for building a hierarchical semantic structure among the languages of the world, as it is different from the English and Chinese WordNet.
4. The study itself can increase understanding of the Yami language and possibly facilitate language revitalization.

As the knowledge and concepts in Yami cannot be interpreted correctly by merely translating it into English or Chinese, we need to carry out a systematic study of lexical semantics to create a Yami WordNet. Therefore, the Yami lexical database can represent a model of semantic connections between the endangered language and other world languages.

In this paper, a new approach to establishing an online lexical database for an endangered indigenous language is proposed. We aim to start with a feasible corpus size for the endangered language. Our goal is to create a lexical database with 80,000 to 200,000 word entries with links to form a semantic network of the Yami language knowledge domain. In addition, this new language database will be connected with several major language lexical databases to form a worldwide semantic network (Fellbanum 1998; Huang et al., 2003). Specifically, the database will be connected with the English WordNet, and BOW, a Chinese-English bilingual database.

The rest of this paper is organized as follows. Section 2 describes the methodology of finding target and source domains based on Chung's (2009) research . Section 3 describes the proposed framework for constructing the Yami lexical database. Section 4 describes the process of transforming the constructed ontologies into the lexical database. Section 5 shows the evaluation of our model. The conclusion and future studies are in Section 6.

2 Research Finding of Formosan Language Metaphore

Chung (2009) proposed a way to identify concreteness of a source domain, based on SUMO and WordNet. She suggested that lexical and computational methods were able to reduce human subjectivity in determining source domains through both top-down and bottom-up approaches and hypothesized that the top-down approach would return general source domains, while the bottom-up approach would return specific source domains.

Her lexical knowledge bases consisted of SUMO, WordNet, SinicaBow and other in-house Chinese corpora. First, she identified the most frequent words in her corpora, as the more frequent a word/construction is, the more likely this word/construction is to be used metaphorically. Then she selected senses (e.g., economy) from SinicaBow as target domains for metaphor analysis and connection with the WordNet explanation and SUMO node. Thirdly, she used both top-down and bottom-up approaches (manual identification) to do automatic grouping of metaphorical expressions in terms of source domains.

Her study shows that ontology can be seen as a coarse-grained categorization of concepts, which can be examined in fine-grained corpora analyses; however, ontology is unable to predict regional differences and construction differences. As some data from this

two-pronged approach will not match the ontological representation of human knowledge, certain adjustments will need to be made to SUMO.

3 Model for Determining Yami Semantics

To accommodate the different approaches to semantic analysis, we propose a model to integrate them. One is a top-down approach by adopting the semantics from other languages' lexical databases and the other is a bottom-up approach of constructing the semantics from the Yami indigenous knowledge. The bottom-up approach attempts to find the semantics that is different from the general semantics of word senses collected for other lexical databases and the semantics standard, IEEE SUMO(Niles and Pease 2001).

In the proposed model, a Yami ontology is constructed, using two approaches. The first one is the ontological integration approach, whereby a systematic process following the seven steps in Noy and McGuinness (2001) is used to integrate the semantic word senses from the Yami corpora and Yami dictionaries (Rau et al., 2007; Rau and Dong 2006) with the semantic hierarchies from the databases of WordNet, Academia Sinica BOW, and the indigenous knowledge database of the NADP project. The results are arranged following the SUMO (IEEE Upper-level Semantics ontology) semantic structure.

The advantages of this approach are twofold. On the one hand, it can develop the semantics ontology systematically and computationally, which is suitable for the process of natural language processing (NPL). On the other hand, the ontology created by this approach can be computationally integrated into the other domain knowledge following a similar hierarchy.

The second approach is an attempt to cover metaphorical expressions and the source domains which are not included in the hierarchy of WordNet-- SUMO alignment in the previous approaches (Kato et al., 2009; Pei et al., 2004). We plan to create a knowledge database called the Yami upper-level semantic metaphorical database. In this database, the word senses related to the unique Yami culture and usages are categorized into a relational structure similar to the semantic structure of SUMO. To create this database, the conceptual metaphors in the Yami words and their contexts are retrieved by using several mapping and alignment technologies to crawl the Yami language resources in the Internet to reinterpret the Yami archives and lyrics. The items in the Yami upper-level semantic metaphorical database can be used to create a different ontology to interpret the word senses and contextual information. As the contextual information is likely to be different from the semantic ontology created in the first approach, it is necessary to consolidate the differences of these two approaches. In this model, we propose a Yami ontology analyzer for integrating the ontologies from the two approaches.

The ontology analyzer is a system containing the following functions: (1) To perform ontological integration for the nodes with different meanings from different ontologies; (2) To resolve the semantic metaphorical differences among the ontologies; and (3) To realign some ontologies with the semantic links from these two approaches. The ontology analyzer will generate a set of Yami ontologies with metaphorical semantics extracted from Yami corpora. Some ontologies will have links with the word senses and semantics from WordNet and BOW. Figure 1 illustrates the framework for building Yami ontologies.

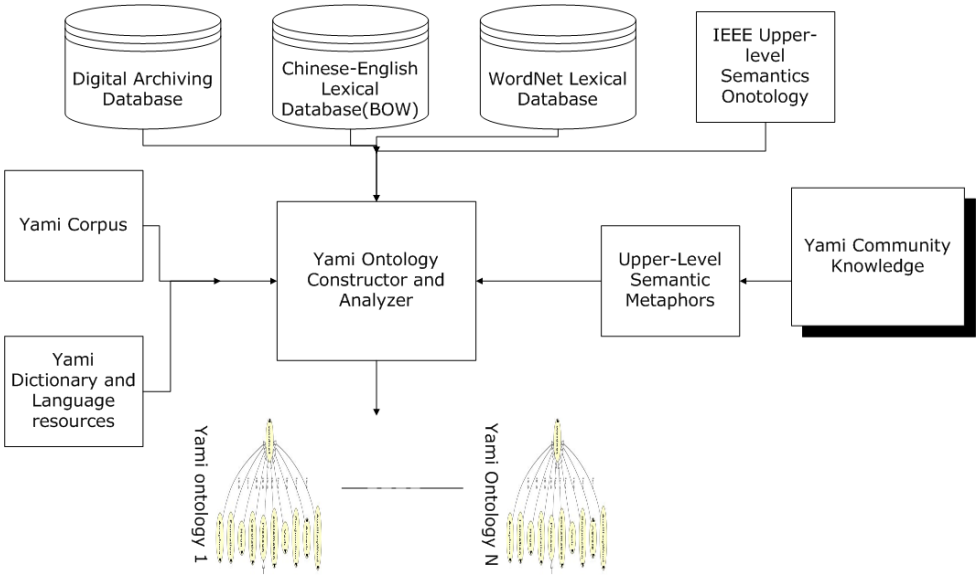

Figure 1. The Framework for building Yami ontologies

After the Yami ontologies have been created, they will be used to create the Yami lexical database, which is illustrated in Figure 2. First, the Yami ontologies and the Yami upper level semantics are used as the semantic guideline to create the Yami lexical database. Then, the entries in Yami dictionaries, WordNet, and Academia Sinica BOW are put into the Yami Semantic Mapping System to construct the Yami lexical database.

In the Yami Semantic mapping system, the Yami ontologies are collected and woven into a Yami semantic knowledge database. The Yami semantic knowledge database includes all metaphorical inductions and the semantic links from the information in the Yami ontologies. The Yami semantic knowledge database is, in turn, interpreted and transformed into the Yami lexical database. This proposed model can produce a computerized Yami lexical database automatically with less human intervention, with the hope of reducing ambiguities of the Yami metaphorical meanings.

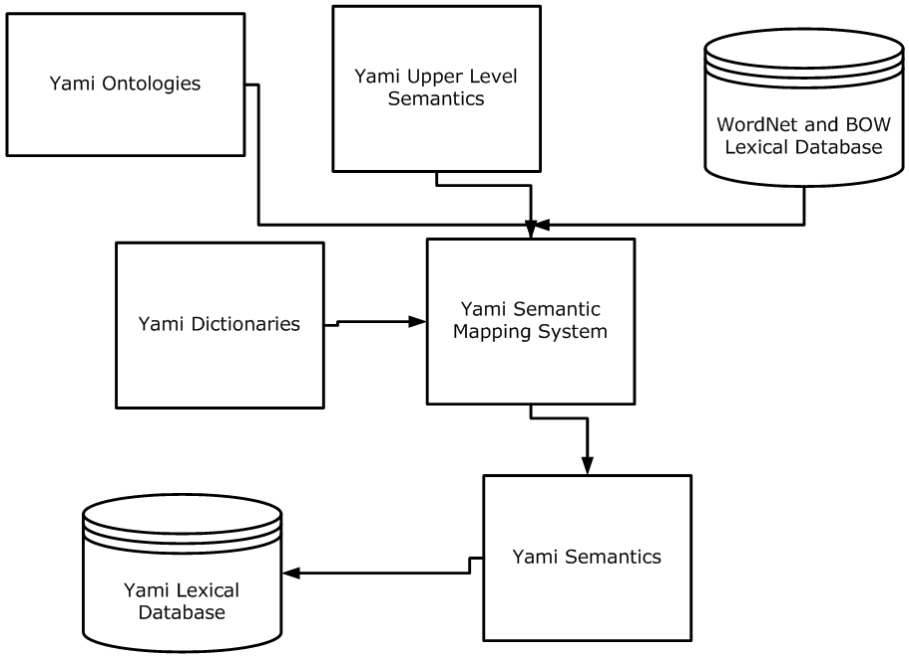

Figure 2. The Process of Building the Yami Lexical Database

Currently, the details of the proposed model are being developed. In the pilot test stage, we found the proposed model was able to produce a Yami lexical database. The process of creating Yami ontologies has also increased our understandings of Yami semantics, as discussed in the next section.

4 Approaches to Constructing Yami Semantic Ontologies

In this section, the methods for creating the Yami ontologies are presented. Two bottom-up approaches have been or will be used in our projects. The first approach followed the steps proposed by Noy and McGuinness (2001) . The second approach is proposed to transform the indigenous upper-level semantic metaphors to ontologies.

4.1 The seven steps for building ontologies

The seven steps from the "Ontology 101 development process" by Noy and McGuinness (2001) used to generate the Yami ontologies are presented as follows:

Step 1: Determine the domain and scope of the ontology

The ontology development starts with basic questions which should be addressed to limit the scope of the model:

- What domain will the ontology cover?
- What is the purpose of the ontology?
- For what types of questions should the information in the ontology be able to provide answers?
- Who will use and maintain the ontology?

Here is an illustration of a fish ontology in the implementation phase.

The Yami Fish Ontology covers the integration of Yami semantic knowledge database that is associated with knowledge in Yami fish. The main purpose for building the ontology was to help preserve Yami cultural heritage for the speech community. Based on Rau and Dong (2006) study , an initial list of competency questions for which the Yami Fish ontology can provide answers was generated:

- Which fish are considered edible and inedible for Yami people?
- Which sex can eat what kinds of fish?
- What kinds of fish can be eaten by Yami elderly males?
- What kinds of fish can pregnant women eat?

Step 2: Consider reusing existing ontologies

We then found fish names in Yami from the database of the Yami online dictionary project (http://yamibow.cs.pu.edu.tw) and their scientific names from the Fish Database of Taiwan (http://fishdb.sinica.edu.tw/).

Step 3: Enumerate important terms in the ontology

Three important fish-related terms were identified in this ontology. They are: (a) classification of Yami fish from the perspectives of Yami consumers: *anito* 'lit. ghost, referring to inedible fish', *raet* 'lit. bad, referring to fish for men', *oyod* 'lit. true or good, referring to fish for women', and *kakanen no rarakeh* 'lit. food for old people, referring to fish for old men', (b) Yami fish names, such as *paloy* 'big eye emperor', and *cilat* 'jackfish' and (c) classification of Yami people as fish consumers: women, young men, and old men.

Step 4: Define classes and the class hierarchy

The classification of Yami Fish was generated (see Tables 1-2).

Named Yami fish / **Classification of Yami fish & Yami People**			***paloy*** **"big eye emperor"**
Yami People	Women	not pregnant	× cannot_eat
		pregnant	×
		breast feeding	×
	Young Men		+ can_eat
	Old Men		+
Classification of Yami fish	*oyod* (Good fish)		×
	raet (Bad fish)		+ has
	kakanen no rarakeh (Food for old people)		×
	anito (Inedible fish)		×

Table 1. The classification of Yami fish

As shown in Table 1, *oyod* 'good fish' can be eaten by all three groups of people, while *anito* 'inedible fish' cannot be eaten by anybody; *raet* 'bad fish' can be eaten by men but not women, while *kakanen no rarakeh* 'food for old people' can only be eaten by old men. Table 2 presents the example classification of *paloy*. As a *raet* fish, *paloy* cannot be eaten by women.

Yami fish / Yami people	*oyod* (Good fish)	*raet* (Bad fish)	*kakanen no rarakeh* (Food for old people)	*anito* (Inedible fish)
Women	+ (can eat)	– (cannot eat)	–	–
Young Men	+	+	–	–
Old Men	+	+	+	–

Table 2. The example classification of *paloy*

Steps 5-6: Define the properties of classes and slots and define the facets of the slots

There are two main types of properties representing relationships between two individuals, i.e., object properties and datatype properties. These properties are readily available from the list produced as a result of Step 3 and Step 4. Thus object properties, as shown in Table 3, were generated based on Tables 1 and 2. The relationship between an old man and a *paloy* is "can_eat" while a *paloy* and an old man is in the relationship of "can_be_eaten_*by*."

Relationship: **object properties**	**Relationship Description**
can_be_eaten_*by*	X <can_be_eaten_by> Y. E.g. "*paloy*" < can_be_eaten_by> "Old_Men;" This is used as the inverse relationship of < can eat >.
cannot_be_eaten_*by*	X <cannot_be_eaten_by> Y. E.g. "*anito*" < cannot_be_eaten_by> "Old_Men;" This is used as the inverse relationship of < cannot_eat>.
has	Y <has> X. E.g. "*rahet*" <has> *"paloy*." This is used as the inverse relationship of < is_a >.

Table 3. Object properties

The creation of the datatype properties was made by consulting the Fish Database of Taiwan to describe the relationship between an individual and its data values. The datatype properties consist of common, family, and scientific names in Chinese and English, synonyms, Yami literal meanings, and Yami names.

What's more, based on the concepts of Tables 2 and 3, we defined the restricting conditions to describe the relationships between the individuals of fish and individuals of Yami people. For instance, for the inedible fish "anito_class," the necessary and sufficient condition is "cannot_be_eaten_by old men class," "cannot_be_eaten_by women," and "cannot_be_eaten_by young men class."

Step 7: Create instances (individuals).

The last step was creating individual instances of classes in the hierarchy. We chose a class, created an individual instance of that class, and filled the slot values.

As Figure 3 illustrates, the Yami fish named *paloy* is a *rahet*, which cannot be eaten by women. The others are datatype properties, for example, *paloy*'s English common name is 'big eye emperor' and its scientific name is *Monotaxis grandoculis*.

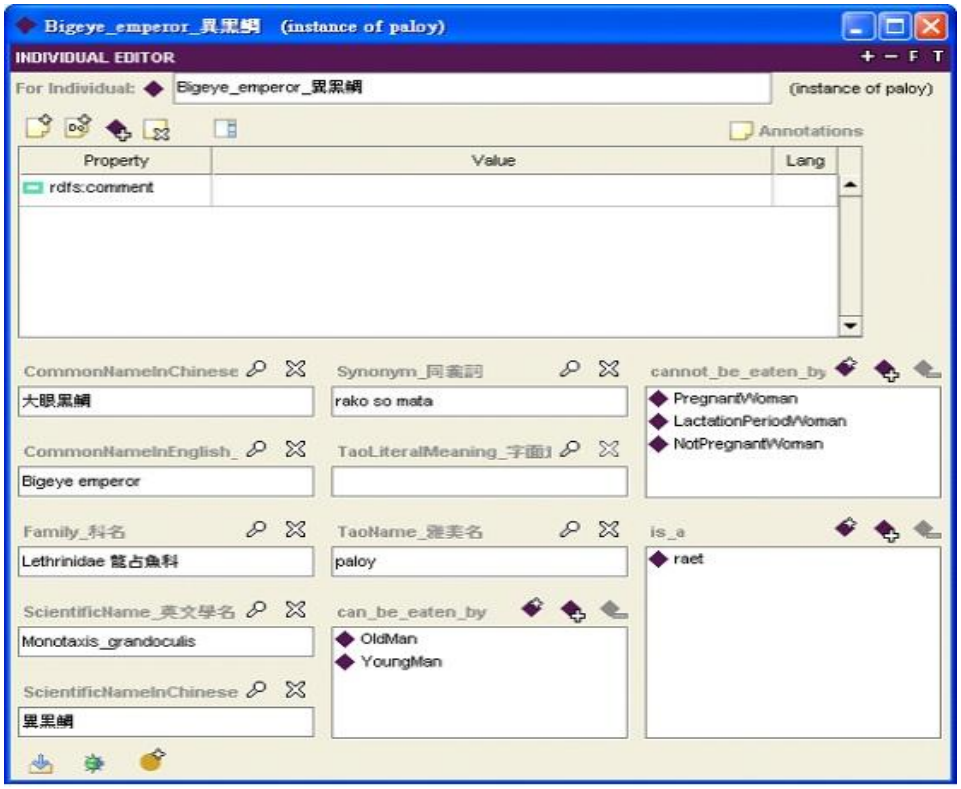

Figure 3. Example of Individual Editor

An example of *paloy* from the Yami fish ontology is provided in Figure 4. It is specified that a big eye emperor is a *paloy*, which is grouped into the concept of *rahet*, a Yami fish which cannot be eaten by Yami women.

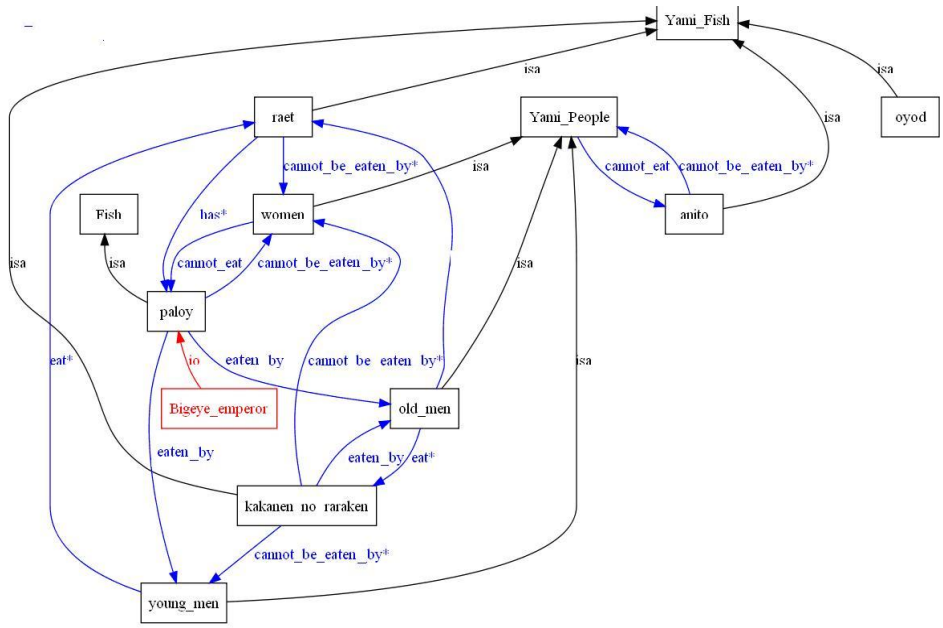

Figure 4. OntoViz display for *paloy* fish

4.2 Reasoning the consistency

Finally, having created the representation of the ontology, we applied reasoning software, RacerPro 1.9.2,1 to keep the online ontology models in a maintainable and modular state and to minimize human errors that are inherent in maintaining a multiple inheritance hierarchy. Figure 5 presents the Ontology browser window generated by Protégé.

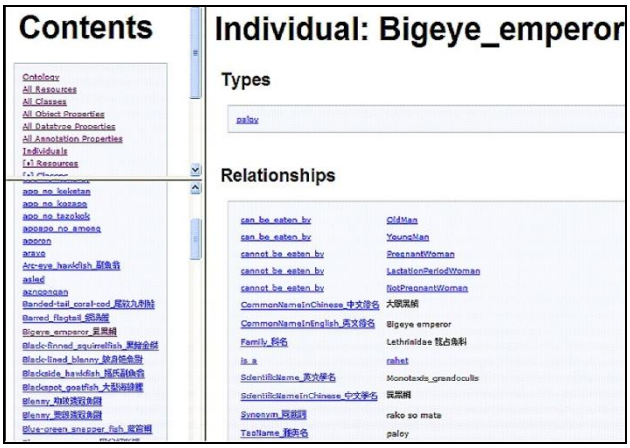

Figure 5. Ontology browser window generated by Protégé

[1] RACER stands for **R**enamed **A** Box and **C**oncept **E**xpression **R**easoner, created by Racer Systems GmbH & Co. KG (http://www.racer-systems.com/index.phtml).

To summarize, the general stages of the Yami ontology construction can be presented as follows in Figure 6.

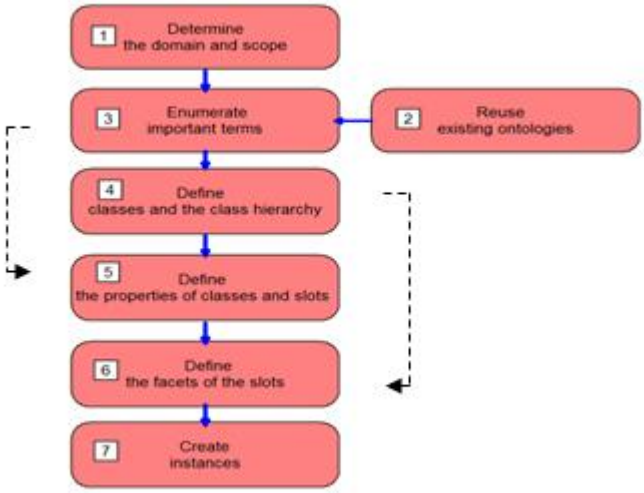

Figure 6. Yami Ontology development process

The seven-step process has been applied to represent formally a Yami fish ontology (Rau et al., 2009) and will be used to discover a Yami plant ontology in a future study. As we begin to build a Yami WordNet, we will encounter metaphorical differences, as described in Section 2, and thus will require a different approach, as described in the following section.

4.3 The bottom-up concept for creating ontologies

The bottom-up approach for creating the ontologies starts from organizing and classifying the semantic entries in the database of the upper-level metaphors. The database consists of major semantic entries collected from our various field studies (Su 2002). These selected entries are likely to be different from those entries in WordNet and SUMO and are unlikely to find matching Chinese or English translations. The entries are collected in a loosely bottom-up approach in that some are from the collected texts, while others can be found by the data mining system.

A reinforcement strategy will be developed to emphasize the semantic structure of these entries and to transform these entries into several groups of induction rules. The strategy follows the study in (Rau et al., 2007) to write rules for creating the Yami ontologies.

5 Framework Evaluation

We have conducted two empirical evaluations of the proposed model. First, an experimental system was constructed to find the Yami lexical database with the information on fish (Rau et al., 2009). As shown in Figure 7, the ontology of the fish entries was created. On the right subtree, the semantics of Yami Fish names are shown as branch nodes. On the left subtree, the traditional classification of Yami people based on fish consumption is shown. In this

ontology, the concept of old people did not show up, but we can find it in another ontology of Yami grammar (Tseng 2009; Yang et al. 2009), as shown in Figure 8, created by the use of Yami upper-level semantics to classify Yami old people.

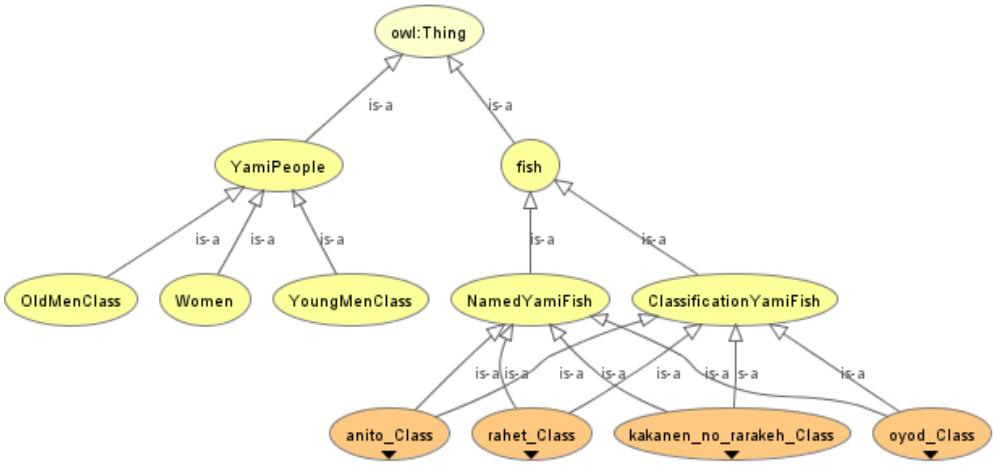

Figure 7. Ontology diagram of Yami Fish

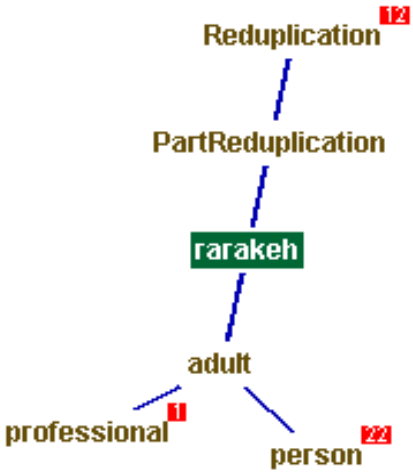

Figure 8. ontology diagram of "old people"

In the proposed approach, the two ontologies can be connected using the Yami upper-level semantic metaphors on Figure 1 and a semantic link between the node "rarakeh" and the "OldMenClass" is created. Then the Yami ontology constructor and analyzer will create a new ontology with the expanded list of fish that can be eaten by old people. The Yami semantic mapping system would use the new ontology to construct the semantics between the synsets of "OldMenClass" and the synsets of the entries in the new ontology.

6 Conclusion

In this paper, a model for constructing a Yami Lexical database for creating Yami ontologies was proposed. This has provided the foundation for us to begin implementing the whole model to construct a sizable Yami lexical database. In this paper, the proposed model is illustrated and the steps of constructing the ontology of the Yami fishery concepts are described.

In future, all modules in the proposed framework will be implemented and the first Yami lexical database will be created. We will use this study to apply to other Formosan languages.

7 Acknowledgement

This research was supported in part by Taiwan E-learning and Digital Archives Programs (TELDAP) sponsored by the National Science Council of Taiwan under NSC Grants: NSC 98-2631-H-126-001.

8 References

Byrne, A. The importance of culture in digital ecosystems: managing indigenous research data, Proceedings of MEDES 2009, pp. 1-7.

Chung, S-F. A corpus-driven approach to source domain determination, Special Monograph A-22, *Language and Linguistics*. Taipei: Academia Sinica. 2009.

Fellbanum, C. Wordnet: An electronic lexical database, The MIT Press, 1998.

Huang, F., Vogel, S. and Waibel, A. Extracting named entity translingual equivalence with limited resources, ACM Trans. On ASIAN Language Information Processing, Vol. 2, No. 2, pp. 124-129, 2003.

Kato, M. P., Ohshima, H., Oyama, S., Tanaka, K. Query by analogical example: Relational search using Web search engine indices, Proceedings of CIKM '09, pp. 27-36, 2009.

Kovecses, Z. Metaphor in culture: Universality and variation, Cambridge University Press, 2005.

Melo, G., and Weikum, G. Towards a universal Wordnet by learning from combined evidence, Proceedings of the 18th ACM conference on Information and knowledge management, pp. 513-522, 2009.

Niles, I, and Pease, A. Toward a standard upper ontology**,** Proceedings of the international conference on Formal Ontology in Information Systems, 2001.

Noy, N. F. and McGuinness, D. L. 2001. Ontology development 101: A guide to creating your first ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880 (http://www-ksl.stanford.edu/people/dlm/papers/ontology101/ontology101-noy-mcguinness.html).

Pei, J., Han, J., Mortazavi-Asi, B., Wang, J., Pinto, H., Chen, Q., Dayal, U., Hsu, M-C. Mining sequential patterns by pattern-growth: The PrefixSpan approach, IEEE Trans. On Knowledge and Data Engineering, Vol. 16, No. 11, pp. 1424 -1440, 2004.

Rau, D. V., Yang, M.-C., Chang, H.-H. and Dong, M.-N. Online dictionary and ontology building for Austronesian languages in Taiwan. *Journal of Language Documentation*

and Conservation, University of Hawaii 3.2: 192-212, December 2009. (http://nflrc.hawaii.edu/ldc/)

Rau, D. V. and Yang, M.-C. Digital transmission of language and culture. In Language endangerment and maintenance in the Austronesian region. Ed. by Margaret Florey. Oxford University Press. pp. 207-224.2009.

Rau, D. V., Yang, M.-C., and Dong, M.-N. Endangered language documentation and transmission. Journal of National Council of Less Commonly Taught Languages (NCOLCTL). University of Wisconsin at Madison. pp. 53-76, 2007.

Rau, D. V. and Dong, M.-N. Yami texts with reference grammar and dictionary, Language and Linguistics. Special Monograph A-10. Taipei: Institute of Linguistics, Academia Sinica 2006.

Smith, B. and Fellbaum, C. Medical WordNet: A new methodology for the construction and validation of information resources for consumer health, Proceedings of the 20th international conference on Computational Linguistics, 2004.

Su, L. I. What can metaphors tell us about culture? Language and Linguistics 3:3, pp. 589-613, 2002.

Tseng, C.-Y. A Semantic framework of translating Web resources for endangered languages: A Yami language prototype, Masters Thesis, Providence University, ROC, 2009.

Wei, C-P, Shi H., Yang, C. C., Feature reinforcement approach to poly-lingual text categorization, LNCS 4822, pp. 99-108, 2007.

Yang, M.-C, Tseng, C.-Y. and Rau, D. V. A Semantic framework for translating web resources for endangered languages: A Yami language prototype. Proceedings of the 4th ICCIT: 2009: International Conference on Computer Sciences and Convergence Information Technology (http://www.computer.org/portal/web/csdl/proceedings/i#4). November 24-26, Seoul, Korea. 256-261, 2009.