

## Online Dictionary and Ontology Building for Austronesian Languages in Taiwan

D. Victoria Rau  
*Wheaton College*

Meng-Chien Yang      Hui-Huan Ann Chang  
*Providence University, Taiwan*

Maa-Neu Dong,  
*National Museum of Natural Sciences, Taiwan*

This paper provides a model of language documentation and conservation in Taiwan to illustrate how online dictionaries have been produced by a collaborative team, and how technology has been used in the process to create a formalized model of existing indigenous knowledge. Our interactions with the Yami community over the past decade have led us to believe that a cooperation framework involving three groups of experts provides necessary “scaffolding” before an “egalitarian” wiki style of online dictionary or ontology building can be attempted. In addition, ontology building requires triangulation of various sources of human interpretations. It is not possible to build an ontology only based on sophisticated machine reasoning. We hope this model of collaboration can serve as a feasible model for other projects in language revitalization and capacity building in the future.

**1. INTRODUCTION.** <sup>1</sup> Traditional dictionary-making by linguists in consultation with community members is a familiar model that has been used by academics for quite some time (e.g., Egerod 1980, Wolff 1972). However, recent attempts at language revitalization and capacity building have placed great emphasis on initiatives from and empowerment of the communities (Florey, Penfield, and Tucker 2009, Rau and Florey 2007). With the development of cross-disciplinary collaboration, several projects have created online dictionaries and have built online ontologies—formal models of structures of information, including terms and the relationships among them, in specific domains of knowledge (Noy and McGuinness 2001)—of indigenous languages. These include work on languages such as Marquesan and Tuamotuan of French Polynesia (Cablitz et al. 2009, Zinn et al. 2008) and

---

<sup>1</sup> The Yami ontology project was funded to the first two authors by the National Science Council (NSC 96-2411-H-126-009-MY2) from 8/1/2007–7/31/2009. The Yami dictionary project was funded to the first and last authors by the Council of Indigenous Peoples in Taiwan from 2007–2009. We would like to acknowledge the generous support of the two funding agencies. This paper was presented at the First International Conference on Language Documentation and Conservation (ICLDC), March 12–14, 2009, Honolulu, Hawai‘i, partially supported by a travel grant from G. W. Aldeen Memorial Fund at Wheaton College to the first author. We would like to thank the editors of LD&C, especially Professor Ken Rehg, and the two anonymous reviewers, whose valuable suggestions have helped us revise this paper. All remaining errors are our own.

Warlpiri of Australia (Corris et al. 2004, Manning et al. 2001); however, there has not been much research in this line on Austronesian languages in Taiwan.

In response to the need to revitalize the endangered Austronesian languages of Taiwan, the Council of Indigenous Peoples (CIP) in Taiwan has implemented a four-year dictionary project, beginning in 2007, which is funding various teams of linguists and community members to complete individual dictionaries. After the completion of these traditional paper dictionary projects, the CIP's next plan is to research indigenous knowledge for the sake of building an indigenous encyclopedia.<sup>2</sup> As Austronesian linguists in Taiwan have accumulated substantial numbers of texts and have developed at least three archives (Hsieh and Huang 2007), we have begun integrating linguistic research with computer technology to build online dictionaries and construct models of domains of indigenous knowledge. Using the Yami corpora from the websites that we have established (Yami language archive, <http://yamiproject.cs.pu.edu.tw/yami>; Yami e-Learning, <http://yamiproject.cs.pu.edu.tw/elearn>) as a point of departure, we have developed a cooperation framework that can be adapted to meet the needs of individual dictionary projects for language documentation and conservation.

The cooperation framework consists of language activists, linguists, and computer scientists. We have developed and are developing several products for indigenous language revitalization: two Yami glossaries, two Yami online dictionaries, a wiki dictionary for Formosan language activists, and a set of online descriptions of ontologies of indigenous knowledge. In this paper, we will first describe the framework, and then we will describe the steps involved in building online dictionaries and creating online ontology models using the resources from the dictionaries.

**2. A COOPERATION FRAMEWORK.** The cooperation framework involves three core members in the collaborative team. The combined efforts of each dyad and the synergy of all three partners have resulted in different productions, which would not have been possible if they had been attempted individually. Although the cooperation framework has been adopted by other projects (e.g., Cablitz et al. 2009), it is by no means easy. The following account comes from our experiences over a decade of collaboration.

One of the crucial factors that has undergirded our cooperation over a decade is the initiative of the Yami language expert (Dong, the fourth author of this paper), who had the vision of compiling a dictionary of her own language in 1994, and invited a linguist (Rau, the first author of this paper), to work on this task. This endeavor has turned out to be a lifelong commitment. When the collaboration pattern of the dyad of the language expert and the linguist had been firmly established, Rau in 2002 recruited a linguistics graduate student with training in computer science (Chang, the third author of this paper) to build a database, and finally invited a computer scientist (Yang, the second author of this paper) on board in 2004. The addition of the computer scientist allowed us to transform the Yami documentation into a digital archive that met the standards of current information and communication technology. The Orchid Island community members were the last to join this

---

<sup>2</sup> Chiu-Yi Wang, Director of Department of Education and Culture, CIP, Taiwan, personal communication.

project, in 2004, as we launched the digital archiving project. Two local language consultants in Orchid Island were recruited to join the research team.

The community members had been suspicious and critical of the collaborative efforts between Rau, a Chinese female linguist, and Dong, the Yami female<sup>3</sup> language expert who had left the community at age 16 to attend high school in Taiwan.<sup>4</sup> Although Rau was able to analyze the linguistic structure of the language, she is a non-speaker of Yami. Dong, on the other hand, along with all other community members who served as consultants had been accused of assisting Chinese anthropologists in gathering Yami cultural knowledge for their own scholarly advancement. The negative attitudes toward “outside” researchers had been a stumbling block for forming trusting relationships between the linguist and the community, and attempts to involve community members, especially males, in collaborating on a dictionary project had never been successful. The linguist and the Yami language expert decided to take up the task on their own, with the goal of “throwing out a minnow to catch a whale.” They went ahead with their work on a Yami textbook (Dong and Rau 2000) and a preliminary project of dictionary-making (Rau and Dong 1999), hoping to bring about criticism and competition from community members if the community perceived the language products as inadequate. However, after the publication of a volume of Yami texts with a reference grammar and vocabulary based on a corpus of twenty selected texts (Rau and Dong 2006) from over a decade of collaboration, the work has finally gained the approval of the community.<sup>5</sup> This has paved the way for successfully recruiting some community members to participate in the ensuing dictionary project. At the same time, the community members who had considered compiling a Yami dictionary on their own, and had actively collected information toward that goal, gradually shifted their plans to publishing books on Yami culture.

Interaction between community members and the computer scientist has resulted in the development of a wiki online dictionary<sup>6</sup> using the concepts and techniques of Web 2.0 (Millard and Ross 2006, Treese 2006) to provide a platform for Formosan language activists to begin documenting and archiving their languages (Chou 2008, Yang et al. 2008). Indigenous communities expressed concerns about the qualifications of the gatekeepers/editors to deal with the input data from unmonitored dictionary contributors (a similar situation is described in Cablitz et al. 2009). In addition, indigenous language activists, who generally lack Chinese word-processing skills, also worried about potential abuse of the intellectual property rights to the materials put online. The greatest concern is “ownership.” Even our language expert (i.e., Dong) was hesitant to use the wiki system, because

---

<sup>3</sup> For a detailed study of gender and sound change in Yami, see Rau, Chang, and Dong 2009.

<sup>4</sup> Orchid Island is a small offshore island southeast of Taiwan and at the northern tip of the Batanes Province of the Philippines. Politically, the island is under the administration of Taiwan.

<sup>5</sup> The publication of Rau and Dong 2006 was significant enough to be reported by Liberty Times in Chinese, (<http://www.libertytimes.com.tw/2006/new/oct/22/today-life1.htm>).

<sup>6</sup> [http://dicts.cs.pu.edu.tw/ada\\_e/](http://dicts.cs.pu.edu.tw/ada_e/)

she felt that the lexical entries that she painstakingly entered seemed to disappear and be taken away by others as soon as she pushed “send.” She felt she could not monitor her progress constantly as she would have been able to do if she were working independently. The printout of the wiki version of the dictionary at this stage is also not as attractive as the one produced by using Toolbox. For these reasons, the wiki dictionary has not gained enough support from any of the current dictionary-making teams of indigenous peoples to be used as their major database, although two teams (those documenting Thao and Seediq) have inquired about the use of it.

The collaboration between the linguist and the computer scientist (i.e., Rau and Yang) has focused on research on ontology, using Protégé to formalize a model of existing indigenous knowledge. Not surprisingly, this is the easiest part in the cooperation framework, as cross-disciplinary collaboration is still within the realm of shared academic culture. Finally, using the same cooperation framework, as illustrated in figure 1, the trio finished a Yami language documentation project (Rau, Yang, and Dong 2007), a Yami e-Learning project (Rau and Yang, forthcoming 2009), two versions of the Yami online dictionaries, and a model of an ontology (to be described in this paper), and went back to the community in May 2009 to teach elementary school students to use animated stories from the Yami e-Learning website to study their own language.

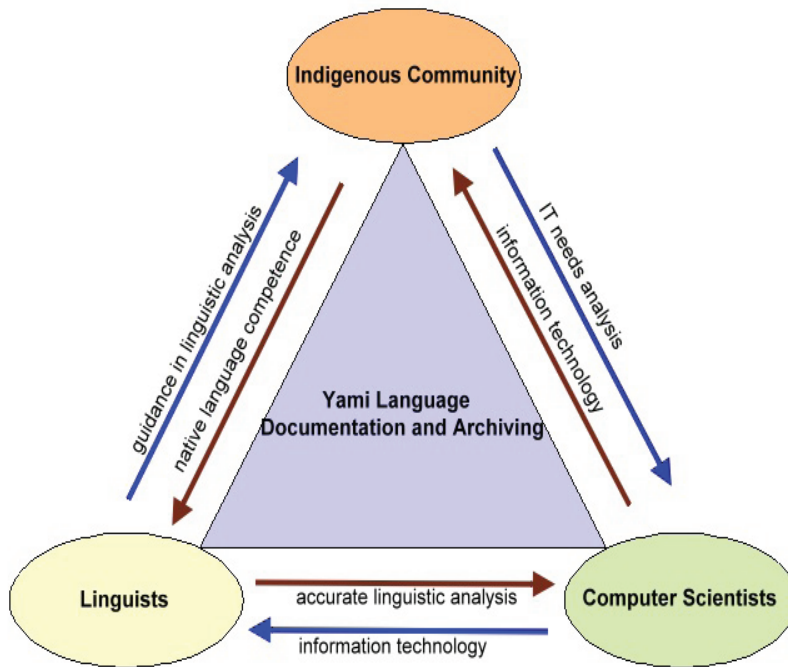


FIGURE 1: A cooperation framework

The cooperation framework allowed us to form interdependent work relationships. The linguist relied on the computer scientist to provide information technology and the language expert for native language competence. The computer scientist relied on the linguist to provide accurate linguistic analysis and the language expert for IT needs analysis. The Yami language expert relied on the linguist for guidance in linguistic analysis and the computer scientist for information technology. The research team's focus on communication and strong connections among participants are what made it possible for us to produce the materials for language conservation based on language documentation.

Our cooperation framework has gradually been articulated as a model of community-based language documentation and conservation, as illustrated in figure 2, with a workflow definition of an indigenous language documentation project that involves all parties and product delivery. The workflow is as follows:

(1) Establish a website for text collection. Two community language consultants videotape four texts (~20 minutes each) each month, the topics of which are preplanned based on the community calendar to cover a wide range of speech events. The linguist and language expert transcribe, annotate, and translate two of them each month. The corpus serves as the basis for writing a reference grammar and developing language teaching materials.

(2) Establish an e-Learning website on language learning. The linguist and the language expert develop language lessons from the transcribed texts and field test them with college students taking a course on Austronesian languages. The computer scientist designs the e-Learning system as a set of teaching materials to supplement classroom instruction and elicit feedback from the speech community.

(3) Establish a wiki online dictionary interface for community members to share their indigenous knowledge. More research is still needed to overcome the obstacles we have encountered in this area.

(4) Establish a website for the online dictionary and ontology building projects. The linguistics sub-team<sup>7</sup> and the language expert input new words from the texts into the dictionary database on a monthly basis, and take pictures of artifacts to illustrate some lexical items. Three representative community members are recruited to check the accuracy of the dictionary and record the pronunciation of lexical entries and example sentences. The computer scientist helps design the interface of the online dictionary to incorporate ontology building, leading toward a model of a collective knowledge system (Gruber 2008).

---

<sup>7</sup> The linguistics sub-team consists of the third author and other graduate students in linguistics recruited as assistants for the project.

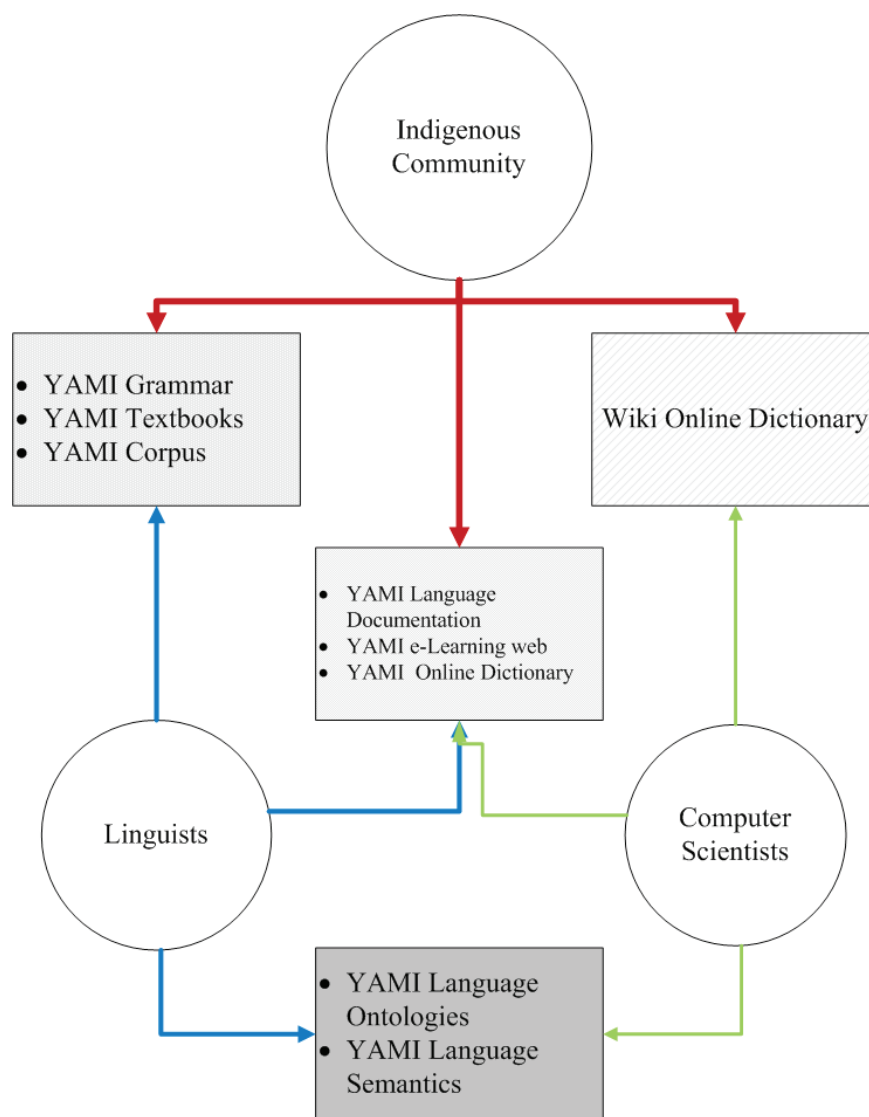


FIGURE 2: An integrated model of language documentation and conservation

The digital model of the collective knowledge system for the Yami language is an ideal online platform for using and sharing the Yami language. It has the potential to build an environment in which user-generated content can play a role similar to that of a community elder, in that it can teach about, explain, and provide new information about the Yami language. As more contributions are made, the available information increases, and so the system is becoming more useful. The scalability of the system can automatically adjust to the structure to compile and arrange new information.

To develop this system, much collaborative research, such as that begun in Yang et al. 2009, is still needed. However, our cooperation framework has paved the way to integrating knowledge of Yami language and culture, linguistic analysis, and computer technology into a newly accessible form of knowledge.

**3. ONLINE GLOSSARIES AND DICTIONARIES.** Two online glossaries and two online dictionaries have been produced, supported by various grants, to meet different purposes. The first glossary is a keyword search<sup>8</sup> from the texts gathered for Digital Archiving Yami Language Documentation,<sup>9</sup> funded by ELDP, SOAS, University of London (August 1, 2005–July 31, 2007). This was designed by the computer scientist (Yang) to make it possible to find all relevant examples in the corpus by typing in the keyword of interest. One special feature of this search is the inclusion of the Yami translation of the New Testament, the most widely read literature on Orchid Island. The Bible verse lookup in Yami is accompanied by both Chinese and English translations.

The second online glossary was initiated at the same time that our team was implementing the ELDP project. An online platform for learning the Yami language was established. Three levels of materials for beginning, intermediate, and advanced learners were developed and put in the platform. In addition, for quick reference, a concise online Yami-Chinese-English glossary,<sup>10</sup> based on the contents of the Yami language textbook materials, was created. This was also designed by the computer scientist to display the Chinese or English translation when a Yami keyword is input. It has provided resources for Yami teenagers preparing to take a Yami language proficiency test to enter better schools, and for Taiwan college students who are taking Austronesian linguistics courses to find supplemental information.

A Yami online dictionary project (<http://yamibow.cs.pu.edu.tw>), funded by the CIP (July 1, 2007–June 30, 2009) to Providence University, Taiwan, was launched after Yami texts and teaching materials had been established. The online dictionary<sup>11</sup> is an essential website for Chinese speakers learning the Yami language. During the two-year-project period, two dictionaries for second language learners were produced. At the end of the first year, a small-scale learning dictionary was produced, containing 1,786 lexical entries, with 780 roots and 1,006 derivatives, and 2,204 example sentences. At the end of the second year, the learning dictionary was expanded into a culture dictionary, containing 6,022 lexical entries, with 3,902 roots, 2,120 derivatives, and a total of 6,450 example sentences. Completing the learning dictionary provided an opportunity to field test our dictionary-making procedures. We noticed that it was necessary to leave six to eight months for proof-

---

<sup>8</sup> <http://yamiproject.cs.pu.edu.tw/yami/database.htm>

<sup>9</sup> <http://yamiproject.cs.pu.edu.tw/yami>

<sup>10</sup> <http://yamiproject.cs.pu.edu.tw/elearn/search.php>

<sup>11</sup> [http://yamibow.cs.pu.edu.tw/index\\_en.htm](http://yamibow.cs.pu.edu.tw/index_en.htm)

reading and revision, so we scheduled extra time for that task when a large-scale culture dictionary was attempted in the second year.

**3.1 WORKFLOW OF DICTIONARY MAKING FOLLOWING THE COOPERATION FRAMEWORK.** The workflow of making both the learning dictionary and the culture dictionary follows the cooperation framework and integrated model described previously. First of all, two community language consultants in Orchid Island videotaped four speech events each month. The texts were screened and selected by the Yami language expert (Dong) for transcription, annotation, and translation. The lexical entries for the online dictionary project were selected from these texts and entered into Toolbox, along with sample sentences. Second, the linguist (Rau) examined the use of orthography, parts of speech, and classification of synonymy and polysemy, while the computer scientist (Yang) designed the online presentation. Third, a draft of the dictionary was presented to the two community language consultants who had recorded the events for reviewing and sound recording.<sup>12</sup> After the drafts were proofread and revised, a linguist with good computer skills (Chang) transformed the output of the database from Toolbox to web pages using Lexique Pro.

In the learning dictionary, a user can choose from a catalogue of twenty semantic classes: animals, boats, body parts, clothing, fish, fruit, houses, insects, kinship, nature, numbers, plants, dimensions, time, weather, colors, fishing, emotion, sickness, and sensation. Figure 3 illustrates the search result for the category “fish.” Each entry in the dictionary has both Chinese and English explanations. In addition to an English index, an index organized by Chinese pinyin spelling was included to make it easier for a Chinese user to look for words. Sound files<sup>13</sup> are also included of pronunciation of words in Iratay,<sup>14</sup> Iraralay, and Ivalino. The recordings were made in these three villages in Orchid Island to cover the two major dialect areas, defined by the rising of the two diphthongs /ay/ and /aw/ (Rau, Chang, and Dong 2009).

**4. ONTOLOGY.** The last part of this paper demonstrates how endangered language data can be mined to expound on theoretical ontologies. After the online dictionaries had been fully established, we began working with the knowledge structure of one semantic domain, with the goal of expanding to cover the whole Yami ontology. As this ontology project is integrated with the online Yami dictionary project (<http://yamibow.cs.pu.edu.tw>), we hope

---

<sup>12</sup> This is a strategy of killing two birds with one stone, as the community members are usually busy with their work (i.e., farming and fishing) during the weekdays and church activities during weekends, and are thus slow to respond to any inquiry on Yami lexical semantics. When they were asked to record their reading of the lexical entries and the example sentences, they were paid to do it, and were willing to set aside two or three days to concentrate on the dictionary draft. As they were reading it aloud, they checked accuracy of the data and made corrections at the same time.

<sup>13</sup> [http://yamibow.cs.pu.edu.tw/ResearchResults\\_en.htm](http://yamibow.cs.pu.edu.tw/ResearchResults_en.htm)

<sup>14</sup> Dong, who is female, recorded the Iratay accent, while the two language consultants, a male and a female, recorded the Iraralay and Ivalino accents, respectively.

to provide an integrated framework for online dictionary building and the construction of online models of ontologies for other Austronesian languages.

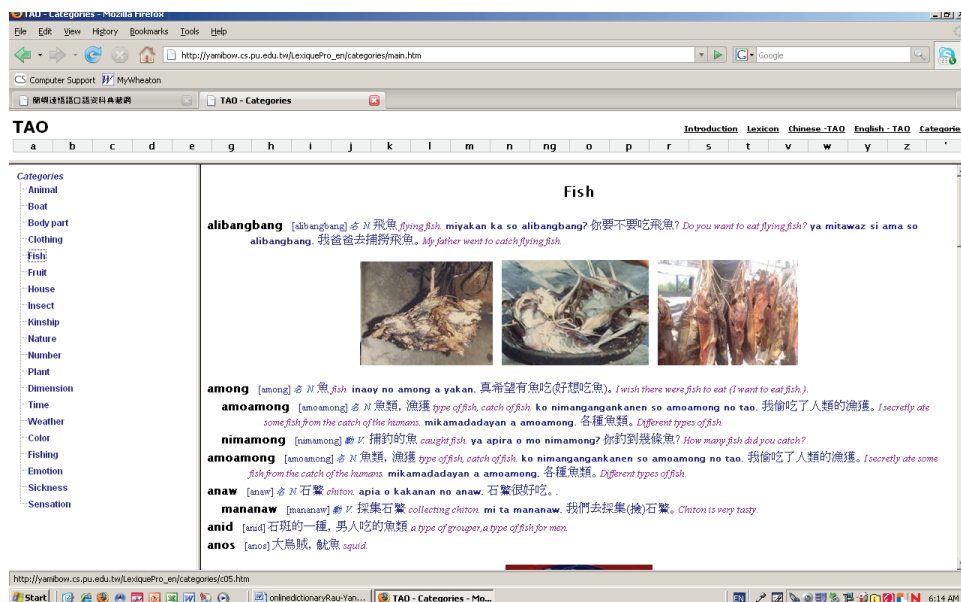


FIGURE 3: An index organized by semantic categories

Following Zinn et al. 2008, we aimed at linking the online dictionary to ontological resources, and allowing community members to play an active role in the process. However, we are fully aware of the fact that human epistemology and ontology are not ahistorical and fixed, but are multifaceted and dynamic (Denzin and Lincoln 2000; Kramsch 2008; Rossman and Rallis 2003). It is not reliable research to accept one or two community members' interpretations of the category of interest without triangulation. Therefore, we decided to start with fish names, as fishing culture is one of the most prominent semantic domains of Yami culture, and is highly integrated into the lives of Orchid Islanders. In addition, previous anthropological studies (e.g., Chen 1994) and ethnobiological studies (e.g., Shao et al. 2007) have accumulated enough lexical materials for us to formalize the existing ontology of fish names. It will become clear in the next section that it is not possible to use complicated machine reasoning to formalize an ontology without anthropological or linguistic interpretations of indigenous knowledge. Although ontology exists in the mind of native language users, it requires using multiple methods to examine multiple sources to validate its representation and categorization.

**4.1 YAMI FISH NAMES.** There have been two ways of classifying fish in Yami, proposed by Chinese and Japanese scholars and local Yami activists. The first is a three-way<sup>15</sup> dis-

<sup>15</sup> If the category of inedible fish is included, this can be seen as a four-way distinction.

inction: fish suitable for young men, women, and old men to consume (Chen 1994, Dong and Rau 2000, Wei and Liu 1962, Yu 1994), and the second is a two-way distinction: the good/real fish or *oyod* and the bad ones or *rahet* (Hsu 1987, Tsuchida 1984, Xie 2003).

Gao (2004) reexamined and reinterpreted previous classifications and proposed that the local classification of fish is a reflection of the organization of the Yami society. In other words, the distinction between *oyod* and *rahet* represents life/death (edible/inedible), holiness/secularism (migratory/nonmigratory), and marked/unmarked (adult males/others). Whereas the two-way distinction focuses on the attributes of the two types of fish, the three-way distinction focuses on the roles of the consumers of fish. The following method uses computer software to represent the current classifications of Yami fish names.

**4.2 NUTS AND BOLTS OF REPRESENTING THE YAMI FISH ONTOLOGY.** The workflow of online ontology building also followed the cooperation framework and integrated model described for online dictionary building. We began with a cooperative effort between Rau, the Chinese linguist and Dong, the Yami expert to provide information<sup>16</sup> on the names of fish, the identity of the group of people who are permitted to eat them, and the consequences of violating the taboo by a pregnant woman, her husband, a woman who has recently given birth, or a woman who is breast feeding.

Before we began work on the fish ontology, a list of 109 Yami fish names based on Chen 1994 and Shao et al. 2007 had been entered into the Toolbox database as part of the online dictionary building process. Two local language consultants on Orchid Island reviewed the Yami fish names to confirm their accuracy. Yang, the computer scientist and Chang, the linguist with good computer skills designed the interface for the ontology. We set up a “Yami Fish Project” with Protégé,<sup>17</sup> following the seven step process in Noy and McGuinness 2001 to generate a Yami fish ontology.

#### **Step 1: Determine the domain and scope of the ontology:**

The following questions were generated from Rau and Dong’s Yami teaching materials, described in footnote 13.

- Which fish are edible and inedible for Yami people?
  - Which gender can eat what kind of fish?
  - What kind of fish can be eaten by Yami old men?
  - What kind of fish can Yami pregnant women eat?

---

<sup>16</sup>The information can be retrieved from the following url addresses: <http://yamiproject.cs.pu.edu.tw/elearn/reading.php?sn=825>  
<http://yamiproject.cs.pu.edu.tw/elearn/reading.php?sn=863>

<sup>17</sup> Protégé is an open-source ontology editor for constructing domain models and knowledge-based applications developed by Stanford Center for Biomedical Informatics Research at the Stanford University School of Medicine (<http://protege.stanford.edu/>).

**Step 2: Consider reusing existing ontologies:**

We adopted the scientific names for fish from the Fish Database of Taiwan (<http://fishdb.sinica.edu.tw/>).

**Step 3: Enumerate important terms in the ontology:**

The important fish-related terms used in this ontology include:

(a) classification of Yami fish:

*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',

*kakanen no rarakeh* 'lit. food for old people, referring to fish for old men';

(b) named Yami fish, such as *ilek* 'rudderfish', and *cilat* 'jackfish';

(c) Yami people: young men, women, and old men, and

(d) women of three stages: not pregnant, pregnant, and breast feeding.

**Step 4: Define classes and the class hierarchy:**

Based on previous studies of Yami fish (Chen 1994, Dong and Rau 2000, Gao 2004, Shao et al. 2007, Yu 1994) and on consultation with the Yami expert (Dong) as the final authority, we generated an initial classification of Yami fish terms (see table 1) that describes the relationship between Yami people and fish. A plus or a minus sign indicates whether the fish is edible or inedible for a specific group of people. It forms an implicational scale with gender placed higher than age in terms of restriction on fish consumption.

<i>Yami people</i>	<i>Yami fish</i>			
	Good fish	Bad fish	Food for old people	Inedible fish
<b>Women</b>	+	-	-	-
<b>Young Men</b>	+	+	-	-
<b>Old men</b>	+	+	+	-

TABLE 1: The classification of Yami fish

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

The coding system is a simple taxonomy that does not describe any property. To find out who can eat what kind of fish and which fish is not edible, an association with the top class “Fish properties” described in the database had to be made. Thus, every kind of fish inherits certain of these properties. We consulted the Fish Database of Taiwan (<http://fish-db.sinica.edu.tw/>) to create the datatype properties to describe the relationships between an individual and its data values.

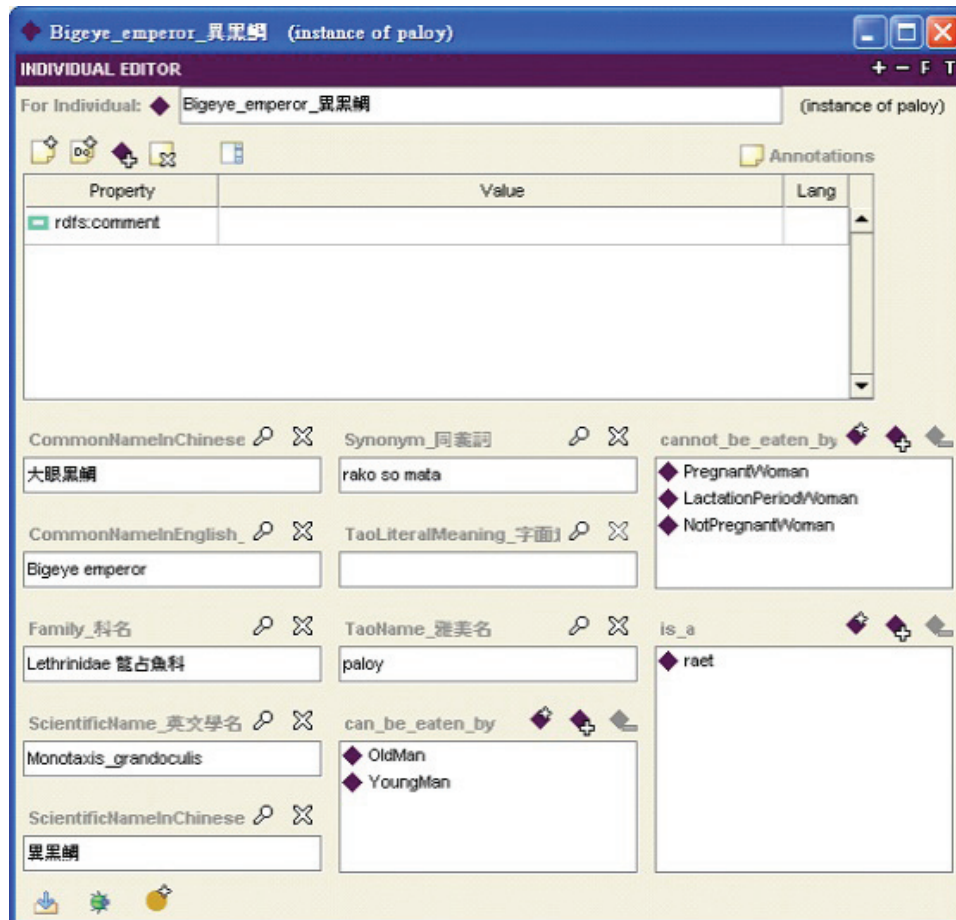


FIGURE 4: An example of a class editor

Based on the concept of table 1, we defined the following restricting conditions to describe 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 some old men class,” and the necessary condition is “cannot\_be\_eaten\_by some women and cannot\_be\_eaten\_by some young men class.”

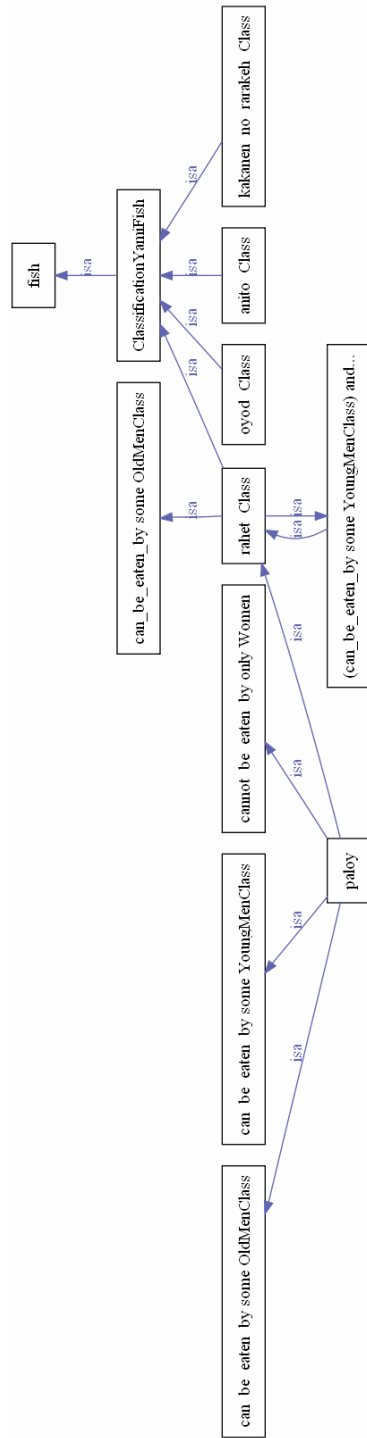


FIGURE 5: OntoViz display for *paloy* fish

### Step 7: Create instances (individuals):

Instances (individuals) were created to represent objects in the fish domain. The description of the ontology needs to represent properties of fish in the hierarchy. Therefore, we modeled basic fish types as instances (individuals). For these fish, we depicted the properties connecting to the Yami people information. As figure 4 illustrates, the Yami fish named *paloy* is a *rahet*.<sup>18</sup> It cannot be eaten by women, but it can be eaten by old men and young men. Others are datatype properties, for example *paloy*'s English common name is 'big eye emperor', and its scientific name is *Monotaxis grandoculis*. Figure 5 is a complete display of the online information we have on the *paloy* fish in relation to the whole Yami fish taxonomy.

Having created the representation of the ontology, we applied reasoning software, RacerPro 1.9.2,<sup>19</sup> to keep the online ontology models in a maintainable and modular state, and to minimize the human errors that are inherent in maintaining a multiple inheritance hierarchy. After checking the logical consistency to ensure that there was no inconsistent class, the reasoning software automatically generated the inferred hierarchy, based on the

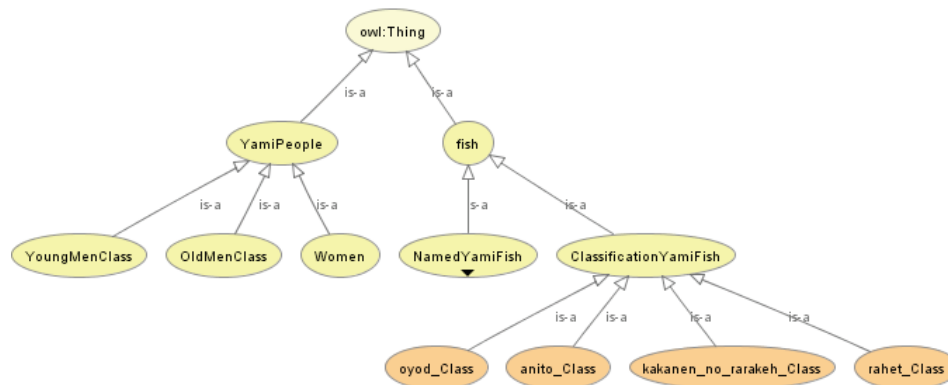


FIGURE 6: OWLViz display of the inferred hierarchy

explicitly defined pre-asserted hierarchy, as shown in figures 6 and 7, respectively.

The four subclasses (oyod, anito, kakanen no rarakeh, and rahet) under the “classification of Yami fish” in the inferred hierarchy has been reclassified under both the “named Yami fish” and the “classification of Yami fish” in the asserted hierarchy.

<sup>18</sup> The information here is based on Chen’s (1994) anthropological study, and confirmed by the fourth author.

<sup>19</sup> RACER stands for Renamed ABox and Concept Expression Reasoner, created by Racer Systems GmbH & Co. KG (<http://www.racer-systems.com/index.phtml>).

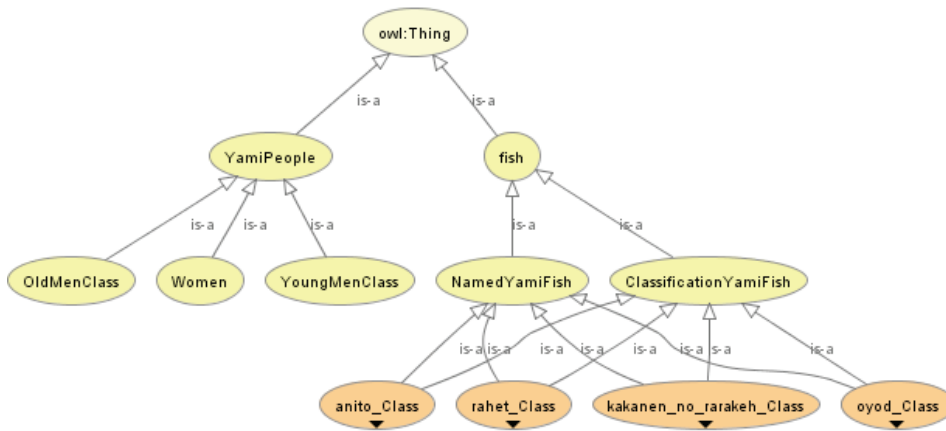


FIGURE 7: OWLViz display of the asserted hierarchy

### Contents

- [Ontology](#)
- [All Resources](#)
- [All Classes](#)
- [All Object Properties](#)
- [All Datatype Properties](#)
- [All Annotation Properties](#)
- [Individuals](#)
- [\[+\] Resources](#)

## Individual: Bigeye\_emperor

### Types

paloy

### Relationships

can be eaten by	<a href="#">OldMan</a>
can be eaten by	<a href="#">YoungMan</a>
cannot be eaten by	<a href="#">PregnantWoman</a>
cannot be eaten by	<a href="#">LactationPeriodWoman</a>
cannot be eaten by	<a href="#">NotPregnantWoman</a>
CommonNameInChinese 中文俗名	大眼黑鯛
CommonNameInEnglish 英文俗名	Bigeye emperor
Family 科名	Lethrinidae 鰹占魚科
is a	<a href="#">rahets</a>
ScientificName 英文學名	Monotaxis grandoculis
ScientificNameInChinese 中文學名	黑鰹
Synonym 同義詞	rako so mata
TeotName 雅賓名	paloy

FIGURE 8: Ontology browser window generated by Protégé

Finally, HTML documentation for the online ontology model was created using OWL Doc plug-in. In other words, ontology documents in the form of HTML pages (OWL Doc) were generated by Protégé to provide links between the elements of the ontological domain (i.e., concepts (classes), properties, and individuals). The OWL document generated for the Yami fish online ontology model is available online at [http://yamibow.cs.pu.edu.tw/fish\\_en/index.html](http://yamibow.cs.pu.edu.tw/fish_en/index.html). Figure 8 shows the Yami Fish HTML ontology browser window for classes and properties.

**4.3. LIMITATIONS AND FUTURE STUDIES.** One of the most ideal approaches to constructing models of ontologies is Zinn et al.’s (2008:2) community-based project, which has the goal of providing “simple but effective tools that empower a broad base of users to describe those concepts.” However, our interactions with the Yami community over the past decade have led us to believe that a cooperation framework involving three groups of

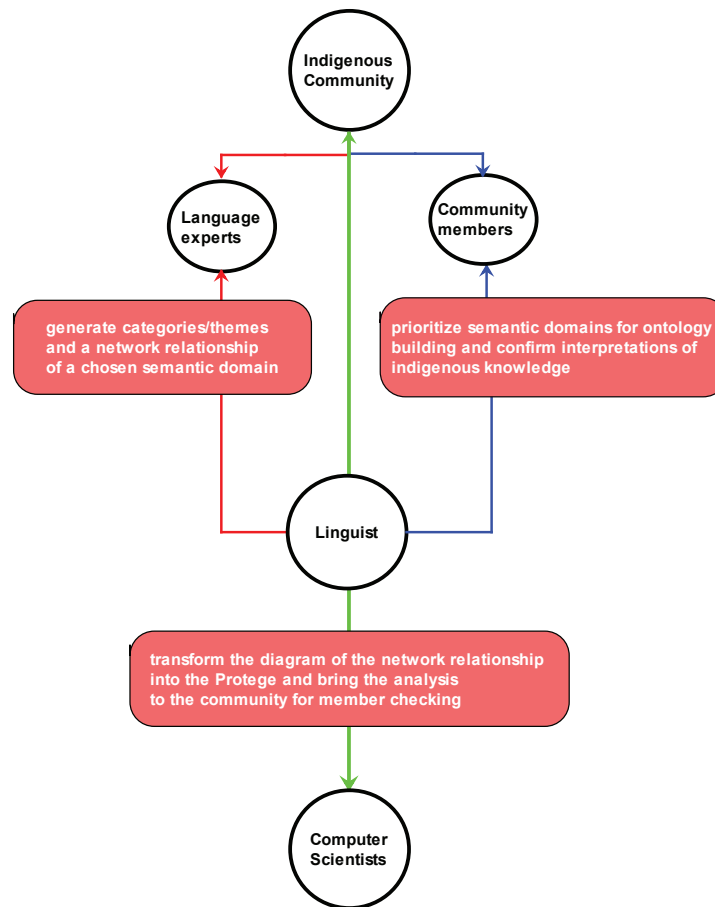


FIGURE 9: A cooperation framework for ontology building

experts, as described in this paper, provides necessary “scaffolding” before an “egalitarian” wiki style of online dictionary or ontology building can be attempted. The community language activists who are interested in developing their own dictionaries are not impeded by complicated database software. In fact, most of the indigenous language dictionary teams in Taiwan are trying to use Toolbox for their own projects. The computer scientist of our team has been contacted by those teams when they have needed troubleshooting assistance in using Toolbox and Lexique Pro since May 2009, when our research team was invited to lead the Workshop on Indigenous Language Dictionary Making at National Dong Huang University in Taiwan. Interestingly, it is the unresolved problems of intellectual property rights and uncoordinated gate-keeping that have prevented the indigenous communities from being willing to investigate the wiki interface at this stage.

In the process of building the online representation of a traditional domain of Yami ontology, we were also confronted with the problem of epistemology. We have come to realize that it is not possible to describe an ontology based on sophisticated machine reasoning. Any ontology description, as with any qualitative research, requires triangulation of various sources of human interpretations. The use of Protégé was mainly a representation of the prefigured classification of the four types of fish, based primarily on Dong’s “indigenous knowledge,” derived partially from her insider’s view, but most crucially from her scrutiny of previous literature. Since an online database of Yami fish names and various anthropological studies on Yami fish culture are available in the literature, it has made our first attempt at creating a formal model of an existing ontological system less daunting.

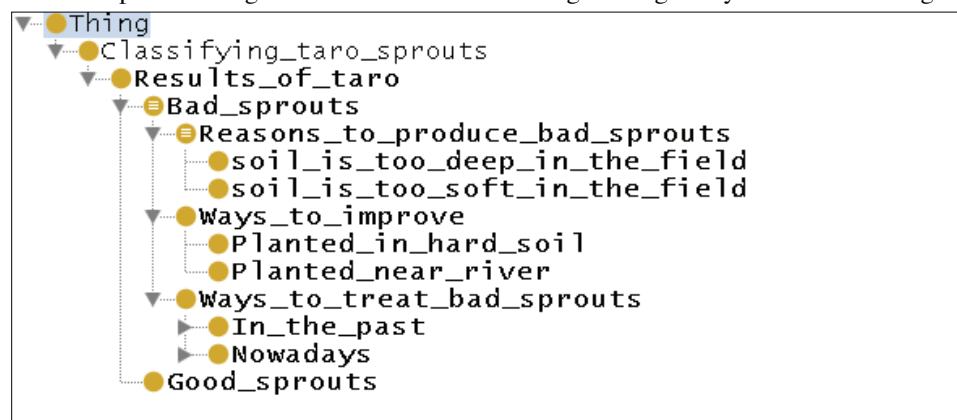


FIGURE 10: Hierarchy of classifying taro sprouts ontology

Although our practical goal of finding cultural items in Yami language is similar to that of Cablitz et al.’s (2007) study on Marquesan and Tuamotuan languages, we hope to continue with theoretical contributions to the work on ontologies in computer science. We hope to extend our findings in Yami to develop a representation of a specific domain of indigenous knowledge in Taiwan, and to find a possible approach to integrating lexical semantics of Yami into English WordNet and Chinese-English bilingual BOW.

For immediate future research, we will continue to use our cooperation framework for building online ontology representations, as illustrated in figure 9. We should begin with the list of semantic domains identified in the online dictionary, prioritizing them by

consulting with community members, conducting a discourse analysis of texts on a chosen domain to generate categories/themes and a network relationship of the semantic domain, transforming the diagram of the network relationship into Protégé, and bringing the analysis to the community for member checking.

This process can be briefly demonstrated based on analysis by Protégé of a text about taro planting, as illustrated in figure 10. A fuller explanation of this methodology will be provided in a future study. The study will continue to find the semantic relations of Yami and other languages. At the end, the set of explicit representations of Yami ontologies that we produce will further understanding of the knowledge domain of the Yami people and the Yami language.

**5. CONCLUSION.** In this paper, we have presented a model of language documentation and conservation in Taiwan to illustrate how online dictionaries have been produced by a collaborative team, and how technology has been used in the process to create a formalized model of existing indigenous knowledge.

As an ongoing project, our next step is to find solutions to concerns about the qualifications of the gatekeepers/editors of the input data from unmonitored dictionary contributors and about potential abuse of the intellectual property rights of the materials put online. Meanwhile, we will continue to document the semantic infrastructure of the Yami language and to expand the scope of our work to other ontological domains and to other endangered Austronesian languages in Taiwan.

This paper provides a case of language documentation and conservation to illustrate how online dictionaries have been produced by a collaborative team and how technology has been used in the process to create formal online representations of existing indigenous knowledge. We hope this model of collaboration can serve as a feasible model for other projects in language revitalization and capacity building in the future.

## REFERENCES

- CABLITZ, GABRIELE, FASAN CHONG, and EDGAR TETAHIOTUPA. 2009. Problems and benefits of web-based tools for language documentation. Paper presented at the First International Conference on Language Documentation and Conservation. (ICLDC), March 12–14, 2009, University of Hawai‘i at Mānoa. <http://scholarspace.manoa.hawaii.edu/handle/10125/5003>. Last accessed October 30, 2009.
- CABLITZ, GABRIELE, JACQUELIN RINGERSMA, and MARC KEMPS-SNIJDERS. 2007. Visualizing endangered indigenous languages of French Polynesia with LEXUS. In *11<sup>th</sup> International Conference Information Visualization*, (IV '07), 409–414. IEEE Computer Society.
- CHEN, SHU-FEN (陳淑芬). 1994. More than food: Yami’s fish. MA thesis, Tsing Hua University, Taiwan.
- CHOU, HSIN-TA. (周信達). 2008. A Formosan multimedia dictionary designed via a participatory process. MA thesis, Providence University, Taiwan.
- CORRIS, MIRIAM, CHRISTOPHER MANNING, SUSAN POETSCH, and JANE SIMPSON. 2004. How useful and usable are dictionaries for speakers of Australian indigenous languages? *International Journal of Lexicography* 17(1):33–68.
- DENZIN, NORMAN K., and YVONNA S. LINCOLN, eds. 2000. *The handbook of qualitative research*, 2nd ed. Thousand Oaks, CA: Sage.
- DONG, MAA-NEU, and D. VICTORIA RAU (董瑪女, 何德華). 2000. *Ciriciring no Tao* [Yami textbook in Chinese]. Taipei: Crane Publishing.
- EGEROD, SØREN. 1980. *Atayal-English dictionary*. 2 volumes. Scandinavian Institute of Asian Studies Monograph Series No. 35. London and Malmo: Curzon Press.
- GAO, XIN-JIE (高信傑). 2004. Rethinking the classification of Yami fish [in Chinese]. *Field Materials, Institute of Ethnology, Academia Sinica*, 18:93–147. Taipei: Institute of Ethnology, Academia Sinica
- FLOREY, MARGARET, SUSAN PENFIELD, and BENJAMIN TUCKER. 2009. Towards a theory of language activism. Paper presented at the First International Conference on Language Documentation and Conservation (ICLDC), March 12–14, 2009, Honolulu, Hawai‘i.
- GRUBER, THOMAS R. 2008. Collective knowledge systems: Where the social web meets the semantic web. *Journal of Web Semantics* 6:4–13.
- HSIEH, FUHUI, and SHUANFAN HUANG. 2007. Documenting and revitalizing Kavalan. In *Documenting and revitalizing Austronesian languages*, ed. by D. Victoria Rau and Margaret Florey, 93–110. *Language Documentation and Conservation* Special Publication No. 1. Honolulu: University of Hawai‘i. <http://hdl.handle.net/10125/1354>
- HSU, SHAO-YING (徐韶讖). 1987. *Heaven, Yami, and fish* [Unpublished, in Chinese; translated by Rong-Yao Huang]. Institute of Ethnology, Academia Sinica.
- KRAMSCH, CLAIRE. 2008. Third places in applied linguistics. Keynote speech at AILA 2008, 15<sup>th</sup> World Congress of Applied Linguistics, August 24–29, Essen, Germany.
- MANNING, CHRISTOPHER D., KEVIN JANSZ, and NITIN INDURKHYA. 2001. Kirrkirr: Software for browsing and visual exploration of a structured Warlpiri dictionary. *Literary and Linguistic Computing* 16(1):123–139.
- MILLARD, DAVID E. and MARTIN ROSS. 2006. Web 2.0: Hypertext by any other name? Paper presented at ACM Conference on Hypertext and Hypermedia 2006, August 23–25, Dense, Denmark.

- NOY, NATALYA F., and DEBORAH L. MCGUINNESS. 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>. Last accessed October 30, 2009.
- RAU, D. VICTORIA, HUI-HUAN CHANG, and MAA-NEU DONG. 2009. A tale of two diphthongs in an indigenous minority language: Yami of Taiwan. In *Variation in indigenous minority languages*, ed. by James N. Stanford and Dennis R. Preston, 259–279. Amsterdam: John Benjamins.
- RAU, D. VICTORIA, and MENG-CHIEN YANG. 2009. Digital transmission of language and culture. In *Language endangerment and maintenance in the Austronesian region*, ed. by Margaret Florey, 207–224. Ontario, Canada: Oxford University Press.
- RAU, D. VICTORIA, and MARGARET FLOREY, eds. 2007. *Documenting and revitalizing Austronesian languages. Language Documentation and Conservation*, Special Publication No. 1. Honolulu: University of Hawai'i. <http://nflrc.hawaii.edu/ldc/sp01>. Last accessed October 30, 2009.
- RAU, D. VICTORIA, MENG-CHIEN YANG, and MAA-NEU DONG. 2007. Endangered language documentation and transmission. *The Journal of National Council of Less Commonly Taught Languages* 4 (Spring 2007):53–76. University of Wisconsin at Madison.
- RAU, D. VICTORIA, and MAA-NEU DONG. 2006. *Yami texts with reference grammar and dictionary*, Language and Linguistics. Special Monograph A-10. Taipei: Institute of Linguistics, Academia Sinica.
- RAU, D. VICTORIA, and MAA-NEU DONG. 1999. Analysis of Yami lexical counts and dictionary making [in Chinese]. *The Newsletter of the Form for Museum Ethnography of Taiwan* 2(3):30–58.
- ROSSMAN, GRETCHEN B., and SHARON F. RALLIS. 2003. *Learning in the field: An introduction to qualitative research*, 2nd ed. Thousand Oaks, CA: Sage.
- SHAO, GUANG-ZHAO, JING-YI CHEN, ZHENG-YI CAI, YU-WEN QIU, JIAN-CHENG YE, and LAI-YU XIE (邵廣昭, 陳靜怡, 蔡正一, 邱郁文, 葉建成, 謝來玉). 2007. Maomaoran no karakowan no wawa do pongso: The ethnobiology and ecology of marine creatures in Orchid Island [in Chinese]. Taitung County Government, Taiwan.
- TREESE, WIN. 2006. Web 2.0: Is it really different? *netWorker, COLUMN: Putting It Together* 10(2):15–17.
- TSUCHIDA, SHIGERU. 1984. Fish names in Yami (Imorod dialect): Second interim report. *Tokyo University Linguistics Papers* 1984:11–90. Tokyo: Department of Linguistics, University of Tokyo.
- WEI, HUI-LIN, and PIN-HSIUNG LIU. 1962. *Social structure of the Yami Botel Tabago* [in Chinese]. Taipei: Institute of Ethnology Academia Sinica.
- WOLFF, JOHN. 1972. *A dictionary of Cebuano Visayan*. Ithaca, NY: Southeast Asia Program, Cornell University. Volume 1, <http://seapdatapapers.library.cornell.edu/cgi/t/text/text-idx?c=seap;idno=seap085a>, Volume 2, <http://seapdatapapers.library.cornell.edu/cgi/t/text/text-idx?c=seap;idno=seap085b>. Last accessed October 30, 2009.
- XIE, YONG-QUAN (謝永泉). 2003. There is no distinction among fish for old man, woman, or man. Yami people only distinguish oyod fish from raet fish [in Chinese]. *Fei Wen Newsletter*, 4.

- YANG, MENG-CHIEN, HSIN-TA CHOU, HUI-SHIUAN GUO, and GIA-PING CHEN. 2008. On designing the Formosan multimedia word dictionaries by a participatory process. In *Documenting and revitalizing Austronesian Languages*, ed. by D. Victoria Rau and Margaret Florey, 111–133. *Language Documentation and Conservation* Special Publication No. 1. Honolulu: University of Hawai'i. <http://hdl.handle.net/10125/1359>
- YANG, MENG-CHIEN, CHIA-YING TSENG, and D. VICTORIA RAU. 2009. A semantic framework for translating web resources for endangered languages: A Yami language prototype. Paper presented at the Fourth International Conference on Computer Sciences and Convergence Information Technology (<http://www.aicit.org/iccit>). November 24–26, Seoul, Korea.
- YU, GUANG-HONG (余光弘). 1994. Classification of Yami food and its social meaning [in Chinese]. *Bulletin of the Institute of Ethnology Academia Sinica*, 76:21–42. Taipei: Institute of Ethnology, Academia Sinica.
- ZINN, CLAUS, GABRIELE CABLITZ, JACQUELIJN RINGERSMA, MARC KEMPS-SNIJDERS, and PETER WITTENBURG. 2008. Constructing knowledge spaces from linguistic resources. Paper presented at CIL 18 Workshop on Linguistic Studies of Ontology: From Lexical Semantics to Formal Ontologies and Back, July 21–26, 2008, Seoul, Republic of Korea.