

Annotating Linguistic Data with ImageSpace for the Preservation of Endangered Languages

Shiyong Lu, Rong Huang, and Farshad Fotouhi
Department of Computer Science
Wayne State University
Detroit, MI 48202, USA
{shiyong, f10272, fotouhi}@cs.wayne.edu

Abstract

Many languages are in serious danger of being lost and as a result, there has been a significant increase in language documentation projects, and also in attempts to preserve language documentation via the use of digital technologies. We are developing a distributed digital library for endangered languages which will contain various data of endangered languages in the forms of text, image, video, audio and include advanced tools for intelligent cataloging, indexing, searching and browsing information on languages and language analysis. As part of this project, we developed ImageSpace, an ontology creation and annotation tool that currently supports both texts and images. In this paper, we share our experience of annotating linguistic data with ImageSpace.

Keywords: Endangered languages, ontology, annotation, the Semantic Web, digital library.

1. Introduction

Many languages are in serious danger of being lost. Language data is central to the research of a large social science community, including linguists, anthropologists, archeologists, historians, sociologists, and political scientists interested in the culture of indigenous people. When a language disappears there are two major effects. First, there is the loss of valuable scientific data about the cultural system that produced the language. The death of a language usually entails the loss of a community's traditional poetry, songs, images, stories, proverbs, laments, and religious rites. Second, any language loss represents a serious scientific loss: studies of linguistic diversity and cross-linguistic comparisons drive much of linguistic theory. In addition, linguistic material provides

valuable information about population movements, contacts, and genetic relationships. When a language becomes highly endangered, efforts to preserve the existing documentation become critically important not only to that community and to academic linguistics, but also to sister sciences such as anthropology, archaeology, history, and ethnobiology.

Digital technologies offer the best promise for the preservation of endangered languages' data, for they give permanent storage, wide dissemination, and flexible access. But to realize that promise, it is important to digitize the material in a way that conforms to best practice within the language technology community, for otherwise it will not be generally accessible, or machine-interpretable. We expect that after various endangered language archives are created, the next step is to integrate these data sources and support intelligently, linguistically searching these archives across the Internet. One major problem with searching a distributed set of archives is the incompatibility of markup of data, and the incompatibility of the queries each system requires. One of the most obvious ways that this problem manifests itself is in multilingual searches: a query might require the use of English, or French, or Spanish terms, for example. One simple approach is to limit query terms to English, and then to search translated terms in collections of other languages based on dictionaries that translate from English to other languages. This is not, however, a sophisticated enough strategy, since what we require is not just mapping between words, but between senses of a word. To take an example, the word "morphology" has many meanings in English, most of which are irrelevant to linguistic data. Its translation into French, "morphologie" has a somewhat different set of senses. How do we know whether a use of "morphologie" is relevant to our search? The solution is to use an ontology

that precisely defines meaning in a domain of knowledge, and then to map the terminology of other languages to that ontology.

As part of the effort of preserving endangered languages at Wayne State University, we have developed *ImageSpace*, an ontology creation and annotation tool that supports both texts and images. Details of *ImageSpace* can be found in [10, 4]. In this paper, we share our experiences of annotating linguistic data with *ImageSpace* for the preservation of endangered languages.

Organization. The rest of the paper is organized as follows. Section 2 describes related work in linguistic ontologies and annotation tools for linguistic data. Section 3 presents linguistic ontology creation and linguistic data annotation with *ImageSpace* using various examples. Finally, Section 6 concludes the paper and presents some future work.

2. Related work

The first and most ambitious effort to develop standards for linguistic markup was the Text

Encoding Initiative (TEI), which was started in 1987. The first widely distributed TEI guidelines [6] provide recommendations for various linguistic markups using SGML.

In the past few years, as the World Wide Web becomes the primary source for storing and accessing data. XML (extensible Markup Language), emerged as the standard language for data representation and exchange on the World Wide Web. Ontologies are used to standardize the semantics markup tags used in different application domains. Ontologies are typically specified in one of the Web ontology languages that have been recommended by the World Wide Web Consortium, including RDF-S [9], DAML+OIL [2], and recently OWL [3]. The goal is that, once data are annotated using the terms defined in a particular ontology, they are not only interpretable by humans, but also by computer programs, a vision of the Semantic Web [1, 5].

Recently, as part of the effort of preserving endangered languages, the University of Arizona developed the GOLD linguistic ontology [7],

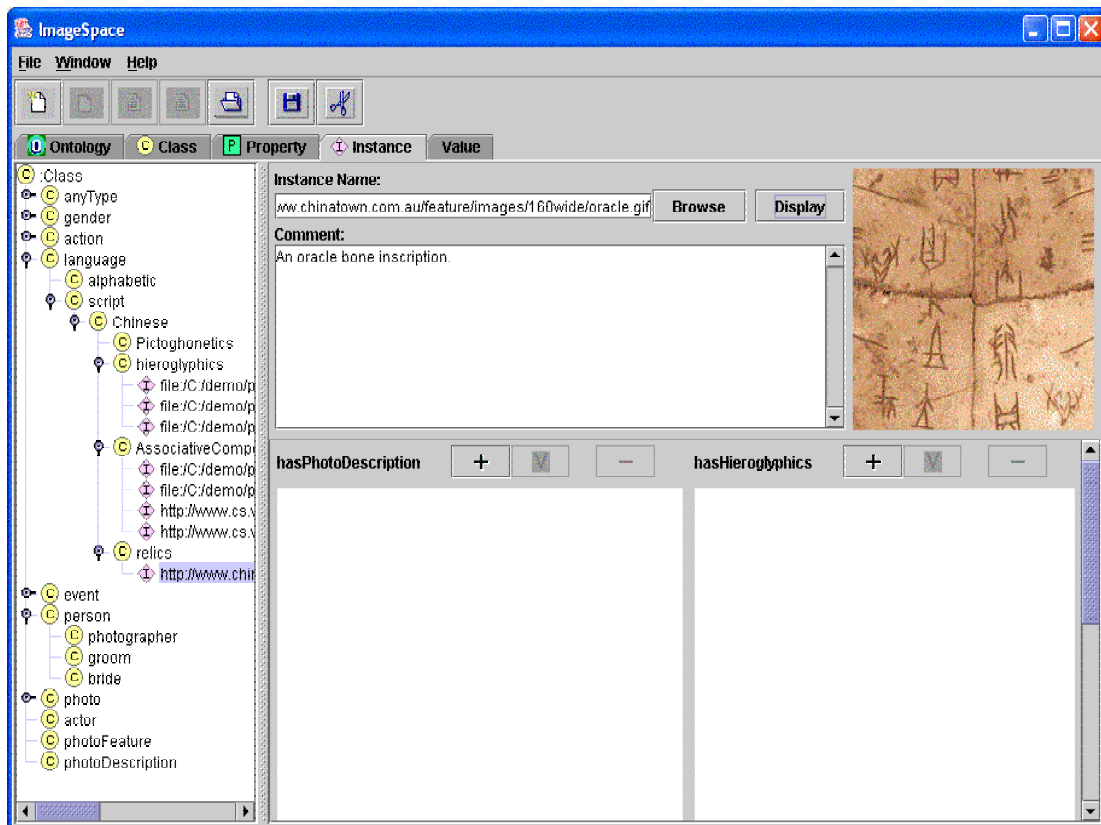


Figure 1. A snapshot of annotating an oracle bone inscription

which is based on the upper ontology SUMO (<http://ontology.teknowledge.com>). It is believed that ontologies will continue to play an important role in the development of the Semantic Web.

Numerous ontology creation tools have been developed. Among them, Protégé (<http://protege.stanford.edu/>), developed at Stanford University, and OntoEdit [8] are two well-known representatives. In addition, it is worth mentioning that Elan [11] is a linguistic multimedia annotation tool that is developed at the Max Institute for Psycholinguistics in Nijmegen. While some of these tools provide partial support to DAML+OIL, *ImageSpace* provides full support for this language, and integrates image ontology creation, image annotation and display in one framework. The tool is built particularly with image support in mind and features a user-friendly interface support for image display and ontology-driven annotation capabilities.

3. Annotating linguistic data

ImageSpace provides a user-friendly interface to the user to annotate resources on the Semantic Web including images and texts. Figure 1 displays a snapshot of annotating an oracle bone inscription with *ImageSpace*. In the following, we illustrate how to annotate linguistic data by annotating some old Chinese characters. We use DAML+OIL [2], one of the standard Web ontology language based on XML syntax as the markup language for our annotation.

It is well known that written Chinese is not an alphabetic language, but a script of ideograms. Their formation follows three principles (<http://www.chinavista.com/experience/hanzi/hanzi.html>):

(1) *Hieroglyphics or the drawing of pictographs*. This category of old Chinese characters was created like *pictures* with their forms and shapes mimicking the physical objects that they refer to. For example, the following four old Chinese characters fall into this category.



(2) *Associative compounds*. Although the drawing of pictographs is intuitive, it is difficult to express abstract ideas with such a principle.

So the ancient Chinese invented associative compounds, characters formed by combining two or more elements, each of which expresses one idea. For example, as shown below, the sun and the moon combined together became the character *ming* (*bright*), the sun placed over a horizontal line forms the ideogram *dan* that means “sunrise” or “morning”.



(3) *Pictophonetics*. Though pictographs and associative compounds can express their meanings by their forms and combinations of forms, they do not give any hint about pronunciation. The pictophonetic method was developed to create new characters by combining one element indicating meaning and the other sound. For example, 爸 (ba) the Chinese character for "papa" is formed by the element 巴 (ba), which represents the sound, and the element 父 (fu), which represents the meaning (father). Likewise the character 芭 (ba) is formed by 巴 (the sound) and 艹, indicating a plant. In this way, more and more characters were made. Today, pictophonetics constitute about 90 percent of all Chinese characters.

As part of the linguistic ontology, we specify that *Hieroglyphics*, *AssociativeCompound* and *Pictophonetics* are *daml:subClassOf* *OldChineseCharacter*:

```
<daml:Class rdf:ID='Hieroglyphics'>
  <daml:subClassOf>
    <daml:Class
      rdf:resource='#OldChineseCharacter'>
    <daml:subClassOf>
  </daml:Class> ...
```

Based on this ontology, we can annotate various kinds of old Chinese characters. For example, the pictograph of *sun* can be annotated by:

```
<Hieroglyphics rdf:about='http://a.b.c/ri.gif'>
  <daml:comment>The sun.</daml:comment>
  <hasSense>Sun</hasSense>
  ...
</Hieroglyphics>
```

For associative compounds, the annotation can indicate what elements constitute a particular

character. For example, the *ming* (bright) can be annotated as follows:

```
<AssociativeCompound
  rdf:about='http://a.b.c./ming.gif'>
  <daml:comment>Bright. </daml:comment>
  <hasSense>Bright</hasSense>
  <hasHieroglyphics rdf:resource='http://a.b.c./ri.gif/'>
  <hasHieroglyphics
    rdf:resource='http://a.b.c./yue.gif/'>
  ...
</AssociativeCompound>
```

For pictophonetics, the annotation can indicate both the meaning element (a pictograph) and the sound element.

```
<Pictophonetics rdf:about='http://a.b.c./father.gif'>
  <daml:comment>Father. </daml:comment>
  <hasSense>Father</hasSense>
  <hasHieroglyphics rdf:resource='http://a.b.c./fu.gif/'>
  <hasHieroglyphics
    rdf:resource='http://a.b.c./ba.gif/'>
  ...
</Pictophonetics>
```

The next version of *ImageSpace* will support the annotation of other multimedia data so that one can associate the annotation of a pictophonetic character with a sound file.

With the above annotation, endangered language data can be marked in a standard form and exported to the Semantic Web, and a semantic search engine will be able to answer queries such as the following:

- return all the old Chinese pictophonetic characters that contains the pictograph *sun*.
- return the most frequently used sound elements in all pictophonetic characters.
- return all the associative compounds that consist of more than two pictographs.
- return all the oracle bone inscriptions in which there is a pictophonetic character ‘爸’.

4. Conclusions and future work

As part of the effort of preserving endangered languages at Wayne State University, we have developed *ImageSpace*, an ontology creation and annotation tool that supports both texts and

images. In this paper, we shared our experience of annotating linguistic data with *ImageSpace*. Future work includes the development of *MultimediaSpace* that will not only support the annotation of images, but also other multimedia resources such as videos, audios, etc. Future version will also support OWL [3], the successor of DAML+OIL. In the meanwhile, the Gold ontology will be extended with terms to describe linguistic multimedia data.

References

- [1] S. Lu, M. Dong and F. Fotouhi, “The Semantic Web: Opportunities and Challenges for Next-Generation Web Applications”, *International Journal of Information Research*, 7(4), 2002.
- [2] F. Harmelen, P. Patel-Schneider and I. Horrocks, “Reference Description of the DAML+OIL Ontology Markup Language”, <http://www.daml.org/2001/03/reference>, March, 2001.
- [3] S. Bechhofer, F. Harmelen, J. Hendler, I. Horrocks, D. McGuinness, P. Patel-Schneider and L. Stein, “OWL Web Ontology Language Reference”, *W3C Candidate Recommendation*. <http://www.w3.org/TR/owl-ref/>. August, 2003.
- [4] R. Huang, “ImageSpace: A DAML+OIL Based Image Ontology Creation and Annotation Tool”, master thesis, advisor: Dr. Shiyong Lu, *Department of Computer Science, Wayne State University*. Dec., 2003.
- [5] T. Berners-Lee, J. Hendler, and O. Lassila. “The Semantic Web”, *Scientific American*. May 2001.
- [6] C. Sperberg-MCQueen and L. Burnard, “Guidelines for Electronic Text Encoding and Interchange (TEI P3)”. *Chicago and Oxford*. Text Encoding Initiative. 1994.
- [7] S. Farrar, W. Lewis, and D. Langendoen, “An Ontology for Linguistic Annotation”, *AAAI Workshop on Semantic Web Meet Language Resources*, Edmonton, Alberta, Canada, July, 2002.
- [8] Y. Sure, M. Erdmann, J. Angele, S. Staab, R. Studer, and D. Wenke, “OntoEdit: Collaborative Ontology Development for the Semantic Web”, *Proc. of the first International Semantic Web Conference 2002 (ISWC 2002)*, June 9-12 2002, Sardinia, Italia.
- [9] D. Brickley and R. Guha, “Resource description framework (RDF) schema specification”. *W3C Working Draft*, April 2002. <http://www.w3.org/TR/rdf-schema>
- [10] Rong Huang, Shiyong Lu and Farshad Fotouhi, “ImageSpace: An Image Ontology Creation and Annotation Tool”, in *Proc. of the 19th International Conference on Computers and Their Applications (CATA'2004)*, Seattle, WA, USA, March, 2004.
- [11] “Elan: A linguistic multimedia annotator”, the Max Institute for Psycholinguistics, Nijmegen. <http://www.mpi.nl/tools/elan.html>