Modeling Wordlists via Semantic Web Technologies

Shakthi Poornima
Department of Linguistics
State University of New York at Buffalo
Buffalo, NY USA
poornima@buffalo.edu

Jeff Good
Department of Linguistics
State University of New York at Buffalo
Buffalo, NY USA
jcgood@buffalo.edu

ABSTRACT

We describe an abstract model for the traditional linguistic wordlist and provide an instantiation of the model in RDF/XML intended to be usable both for linguistic research and machine applications.

Categories and Subject Descriptors

E.2 [**Data**]: Data Storage Representations

Keywords

wordlists, interoperation, RDF

1. INTRODUCTION

Lexical resources are of potential value to both traditional descriptive linguistics as well as computational linguistics.[1] However, the kinds of lexicons produced in the course of linguistic description are not typically easily exploitable in natural language processing applications, despite the fact that they cover a much larger portion of the world's languages than lexicons specifically designed for NLP applications. In fact, one particular descriptive linguistic product, a wordlist, can be found for around a third to a half of the world's seven thousand or so languages, though wordlists have not played a prominent role in NLP to the best of our knowledge.

Wordlists are widely employed by descriptive linguists as a first step towards the creation of a dictionary or as a means to quickly gather information about a language for the purposes of language comparison (especially in parts of the world where languages are poorly documented). Because of this, they exist for many more languages than do full lexicons. While the lexical information they contain is quite sparse, they are relatively consistent in their structure across resources. As we will see, this makes them good candidates for exploitation in the creation of a multilingual database consisting of rough translational equivalents which lacks precision, but has coverage well-beyond what would otherwise be available.

This paper describes an effort to convert around 2700 wordlists covering more than 1500 languages (some wordlists represent dialects) and close to 500,000 forms into an RDF format to make them more readily accessible in a Semantic Web context.[2] This may well represent the largest single collection of wordlists anywhere and certainly represents the largest collection in a standardized format. While the work described here was originally conceived to support descriptive and comparative linguistics, we will argue that the use of Semantic Web technologies has the additional beneficial effect of making these resources more readily usable in other domains, in particular certain NLP applications.

We approach this work as traditional, not computational linguists, and our current goal is to encode the available materials not with new information but rather to transfer the information they contain in a more exploitable format. Semantic Web technologies allow us to represent traditional linguistic data in a way we believe remains faithful to the original creator's conception and, at the same time, to produce a resource that can serve purposes for which it was not originally intended (e.g., simplistic kinds of translation). Our work, therefore, indicates that Semantic Web offers a promising approach for representing the work of descriptive linguists in ways of use to computational linguists.

2. MODELING A WORDLIST

We illustrate the basic structure of a wordlist in (1), which gives a typical presentation format. Here, the language being described is French, with English labels used to index general meanings.

(1)

MAN	*homme*
WOMAN	*femme*

The information encoded in a wordlists is quite sparse. In general, they give no indication of morphosyntactic features (e.g., part of speech), nor of fine-grained semantics. Meanings are most usually indexed simply by the use of labels drawn from languages of wider communication (e.g., English or Spanish), though the intent is not to translate between languages but, rather, to find the closest semantic

[1]Funding provided for the work described here has been provided by NSF grant BCS-0753321 in the context of a larger-scale project, Lexicon-Enhancement via the Gold Ontology, headed by researchers at the Institute for Language Information and Technology at Eastern Michigan University. More information can be found at `http://linguistlist.org/projects/lego.cfm`.

[2]These wordlists were collected by Timothy Usher and Paul Whitehouse in the context of traditional comparative linguistic research, and represent an enormous effort without which the work described here would not have been possible.

WWW2010, April 26-30, 2010, Raleigh, North Carolina.

match in the target language for what is presumed to be a general concept. The notional relationship between a meaning and a form in a wordlist is not one of defining (as is the case in a monolingual dictionary) or translating (as is the case of a bilingual dictionary), but rather something we term *counterpart* following [1]. This is not a particularly precise relation, but it is not intended to be. Specifying too much precision in the meaning-form relationship would make it difficult to collect wordlists rapidly, which is otherwise one of their most desirable features.

The concepts that one sees in traditional linguistic wordlists have often been informally standardized across languages and projects through the use of what we call here *concepticons*. Concepticons are curated sets of concepts, minimally indexed via words from one language of wider communication but, perhaps, also described more elaborately using multiple languages (e.g., English and Spanish) as well as illustrative example sentences. They may include concepts of such general provenance that counterparts would be expected to occur in almost all languages, such as TO EAT, or concepts only relevant to a certain geographical region or language family. For instance, Amazonian languages do not have words for MOSQUE, and Siberian languages do not have a term for TOUCAN [1, p.5-6].

To the extent that the same concepticon can be employed across wordlists, it can be understood as a kind of interlingua, though it is not usually conceptualized as such by descriptive linguists. The concepticon we are employing is based on three available concept lists. The most precise and recently published list is that of the Loanword Typology (LWT) project [1], which consists of around 1400 entries.

3. WORDLISTS AND SEMANTIC WEB

Each wordlist in our RDF datanet consists of two components: metadata and a set of entries. The metadata gives relevant identifying information for the wordlist e.g., a unique identifier, the ISO 639-3 code, the related Ethnologue language name, alternate language names, reference(s), the compilers of the wordlist, etc. The entries set consists of all entries in the wordlist. The structure of our entries is quite simple, consisting of a reference to an external concepticon entry in the concepticon employed by our project paired with a form in the target language using the counterpart relationship discussed above. Obviously, this structure could be elaborated. However, it is sufficient for this first stage of a work and, we believe, serves as an appropriate baseline for further specification.

In cases where there is more than one form attached to a concept, we create two concept-form mappings. For instance, the entry in (2) from a wordlist of North Asmat, a language spoken in Indonesia, associates the concept GRANDFATHER with two counterparts, whose relationship to each other has not been specified in our source.

(2) GRANDFATHER: *-ak*, *afak*

An RDF/XML fragment describing one of the two forms in (2) is given in Figure 1 for illustrative purposes. In addition to drawing on standard RDF constructs, we also draw on descriptive linguistic concepts from GOLD[3] (General Ontology for Linguistic Description), which is intended to be a sharable ontology for language documentation and description. The key data encoded by our RDF representation of wordlists is the counterpart mapping between a particular wordlist concepts (`lego:concept`) drawn from our concepticon and a form (`gold:formUnit`) found in a given wordlist.

```
<rdf:RDF xmlns:rdf="...>
    <lego:concept rdf:about="...">
    <lego:hasCounterpart>
       <gold:LinguisticSign rdf:about="...">
          <gold:inLanguage>
            <gold:Language rdf:about="..."/>
          </gold:inLanguage>
          <gold:hasForm>
            <gold:formUnit>
              <gold:stringRep>-ak</gold:stringRep>
            </gold:formUnit>
          </gold:hasForm>
       </gold:LinguisticSign>
    </lego:hasCounterpart>
    </lego:concept>
</rdf:RDF>
```

Figure 1: Wordlist Entry RDF Fragment

An important feature of our RDF model, illustrated in Figure 1 is that the counterpart relation does not relate a meaning directly to a form but rather to a linguistic sign (`gold:LinguisticSign`) whose form feature then contains the relevant specification. This structure would allow additional information (e.g., part of speech, definition, example) about the lexical element specified by the given form to be added to the representation at the level of the linguistic sign, if it were to become available.

4. PROSPECTS

The data model described here was originally designed to promote lexical data interoperability for descriptive linguistic purposes. At the same time, it makes visible the similarities between a concepticon and an interlingua, thus opening up the possibility of straightforward exploitation of a data type produced in a descriptive linguistic context in NLP contexts. Furthermore, by expressing the model in the form of an RDF graph rather than a more parochial XML format, it can be more easily processed. Potential NLP applications for this datanet involve tasks where simple word-to-word mapping across languages may be useful. One such example is the PanImages[4] search of the PanLex project which facilitates cross-lingual image searching. More work could be done to promote interoperability, of course. For example, we could devise an LMF [2] expression of our model, though we leave this for the future.

5. REFERENCES

[1] In M. Haspelmath and U. Tadmor (eds.), *Loanwords in the world's languages: A comparative handbook*. 2009.

[2] G. Francopoulo, et al. Multilingual resources for NLP in the lexical markup framework (LMF). *Language Resources and Evaluation*, 43:57–70, 2009.

[3] `http://linguistics-ontology.org/`. Similar ontologies such as SKOS could also be used in lieu of GOLD.

[4] `http://www.panimages.org/`