Modeling and Encoding Traditional Wordlists for Machine Applications

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

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

Abstract

This paper describes work being done on the modeling and encoding of a legacy resource, the traditional descriptive wordlist, in ways that make its data accessible to NLP applications. We describe an abstract model for traditional wordlist entries and then provide an instantiation of the model in RDF/XML which makes clear the relationship between our wordlist database and interlingua approaches aimed towards machine translation, and which also allows for straightforward interoperation with data from full lexicons.

1 Introduction

When looking at the relationship between NLP and linguistics, it is typical to focus on the different approaches taken with respect to issues like parsing and generation of natural language data—for example, to compare statistical NLP approaches to those involving grammar engineering. Such comparison is undoubtedly important insofar as it helps us understand how computational methods that are derived from these two lines of research can complement each other. However, one thing that the two areas of work have in common is that they tend to focus on majority languages and majority language resources. Even where this is not the case (Bender et al., 2002; Alvarez et al., 2006; Palmer et al., 2009), the resulting products still cover relatively few languages from a worldwide perspective. This is in part because such work cannot easily make use of the extensive language resources produced by descriptive linguists, the group of researchers that are most actively involved in documenting the world's entire linguistic diversity. In fact, one particular descriptive linguistic product, the wordlist—which is the focus of this paper—can be found for at least a quarter of the world's languages.

Clearly, descriptive linguistic resources can be of potential value not just to traditional linguistics, but also to computational linguistics. The difficulty, however, is that the kinds of resources produced in the course of linguistic description are typically not easily exploitable in NLP applications. Nevertheless, in the last decade or so, it has become widely recognized that the development of new digital methods for encoding language data can, in principle, not only help descriptive linguists to work more effectively but also allow them, with relatively little extra effort, to produce resources which can be straightforwardly repurposed for, among other things, NLP (Simons et al., 2004; Farrar and Lewis, 2007).

Despite this, it has proven difficult to create significant electronic descriptive resources due to the complex and specific problems inevitably associated with the conversion of legacy data. One exception to this is found in the work done in the context of the ODIN project (Xia and Lewis, 2009), a significant database of interlinear glossed text (IGT), a standard descriptive linguistic data format (Palmer et al., 2009), compiled by searching the Web for legacy instances of IGT.

This paper describes another attempt to transform an existing legacy dataset into a more readily repurposable format. Our data consists of traditional descriptive wordlists originally collected for comparative and historical linguistic research.[1] Wordlists have been widely employed 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

[1] These wordlists were collected by Timothy Usher and Paul Whitehouse and represent an enormous effort without which the work described here would not have been possible. The RDF/XML implementations discussed in this paper will be made available at http://lego.linguistlist.org within the context of the Lexicon Enhancement via the GOLD Ontology project.

Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground, ACL 2010, pages 1–9,
Uppsala, Sweden, 16 July 2010.

are poorly documented). Because of this, they exist for many more languages than do full lexicons. While the lexical information that wordlists contain is quite sparse, they are relatively consistent in their structure across resources. This allows for the creation of a large-scale multilingual database consisting of rough translational equivalents which may lack precision but has coverage well-beyond what would otherwise be available.

2 The Data and Project Background

The data we are working with consists of 2,700 wordlists drawn from more than 1,500 languages (some wordlists represent dialects) and close to 500,000 forms. This is almost certainly the largest collection of wordlists in a standardized format. The average size of the individual wordlists is rather small, around 200 words, making them comparable in size to the resources found in a project like NEDO (Takenobu, 2006), though smaller than in other related projects like those discussed in section 4. While the work described here was originally conceived to support descriptive and comparative linguistics, our data model and choice of encoding technologies has had the additional effect of making these resources readily exploitable in other domains, in particular NLP. We have approached the data initially as traditional, not computational, linguists, and our first goal has been to encode the available materials not with any new information but rather to transfer the information they originally contained in a more exploitable way.

By way of introduction, the hypothetical example in (1) illustrates a traditional presentation format of a wordlist, with English as the source language and French as the target language.

(1) MAN *homme*
WOMAN *femme*

As we will describe in more detail in section 5, they key features of a wordlist entry are an index to a concept assumed to be of general provenance (e.g., MAN) and a form drawn from a specific language (e.g. *homme*) determined to be the counterpart for that concept within that language. Most typically, the elements indexing the relevant concepts are words drawn from languages of wider communication (e.g., English or Spanish).

3 Related Work in Descriptive Linguistics

Recent years have seen a fair amount of attention paid to the modeling of traditional linguistic data types, including lexicons, glossed texts, and grammars (Bell and Bird, 2000; Good, 2004; Palmer and Erk, 2007; Nordhoff, 2008). The data type of focus here, wordlists, has not seen serious treatment. Superficially, wordlists resemble lexicons and, of course, they can be considered a kind of lexical resource. However, as will be shown in section 5, there are important differences between lexicons and wordlists which have implications for how they should be modeled.

Most of the work on modeling descriptive linguistic data types has proceeded without special consideration for possible NLP applications for the data being encoded. This is largely because the work was initially a response to issues relating to the longevity of digital descriptive data which was, otherwise, quite often being encoded solely in (often proprietary) presentation formats (Bird and Simons, 2003). However, the possibility for fruitful interaction between computational linguistics and descriptive linguistics is apparent and has been the subject of some work (Palmer et al., 2009).

The work described here is also interested in this possibility. In particular, we address the question of how to model and encode a large-scale dataset that was originally intended to be used for descriptive purposes in ways that not only allow us to faithfully represent the intention of the original creator but also permit the data to be straightforwardly exploitable for new uses, including NLP. To the best of our knowledge, our work is innovative both because of the data type being explored and because the data modeling is being done parallel with the transformation of a legacy resource with significant coverage of the world's languages. This stands in contrast to most other work (again, with the exception of work done within ODIN (Xia and Lewis, 2009)) whose data, while representative, is not of the same scale.

4 Related Work on Lexicon Interoperability in NLP

The relevant related work in NLP is that focused on interoperation among lexical resources. One way to achieve this is to make use of language independent ontologies (or comparable objects) for word meanings which can serve as pivots for mul-

tilingual applications (Ide et al., 1998; Vossen, 2004; Nirenburg et al., 2004; Ronzano et al., 2010). The word senses provided by WordNet, for example, have been used for this purpose (O'Hara et al., 1998).

A recognized data modeling standard for lexical interoperation is the Lexical Markup Framework (LMF), which provides standardized framework for the description and representation of lexicons (Francopoulo et al., 2009). Instantiations of LMF have also been extended to represent WordNets, e.g., Wordnet-LMF (Soria et al., 2009), in ways which facilitate interoperation.

While we do not attempt to express the data model we develop here in LMF, doing so should be relatively straightforward. The key conceptual observation is to recognize that the sets of meaning labels found in wordlists (see section 2) can be treated either as a shared language-neutral ontology or as a kind of interlingua, both of which have already been the subject of LMF modeling (Vossen, 2004). As such, they are also comparable to language-independent ontologies of word meaning, bringing them in line with the work on multilingual NLP mentioned above.

These similarities should not be too surprising. After all, one of the functions of wordlists has been to facilitate language comparison, something which is also at the heart of multilingual NLP. An important development, however, is that new data encoding technologies can allow us to encode word list data in ways that facilitate its repurposing for NLP applications much more easily than would have been possible previously. We will come back to this in section 6.

5 Modeling Wordlists

5.1 Wordlist Entries as Defective Signs

A common linguistic conceptualization of a lexical item is to treat it as a *sign* triple: an association of a *form* with *meaning* and *grammar*. Lexical items in a lexicon generally contain information on all three aspects of this triple. Wordlists do not, and the information they encode is quite sparse. In general, they give no indication of grammatical information (e.g., part of speech), nor of language-specific semantics.

In addition, from a descriptive standpoint, lexicons and wordlists differ in the direction of the form-meaning mapping. As the example in (1) suggests, in order to create or interpret a wordlist, one begins with an abstract meaning, for example MAN, and then tries to find the word in the target language which represents the best semantic fit for that meaning. Lexicons, on the other hand, prototypically map in the opposite direction from form to meaning. Furthermore, as will be elaborated in section 5.3, the meanings employed in wordlists are not intended to refer to meanings of lexical items in specific languages. In this way, they are quite distinct from bilingual dictionaries.

We can therefore view a wordlist as a set of defective signs—containing information on the form and meaning parts of the triple, but not the grammar. The meaning information is not directly associated with the specific form but, rather, is a kind of "tag" indicating that the entire sign that a given form is associated with is the best counterpart in the language for a general concept.

Figure 1 compares the kind of information associated with signs in a lexicon to those in a wordlist. The box on the left gives a schematic form-grammar-meaning triple for the Spanish word *perro* 'dog', containing the sort of information that might be found in a simple bilingual dictionary. The box on the right schematizes the content of a parallel French wordlist entry for *chien* 'dog'. Here, no grammatical or semantic information is associated with the form, but there is an indication that in French, this lexical item is the closest counterpart to the general concept DOG. Of course, in this case, the word *chien* is not only the counterpart of DOG in French, but can be translated as *dog* in English. The semantic connection between a concept label and a lexical item may not always be so straightforward, as we will see in section 5.2.

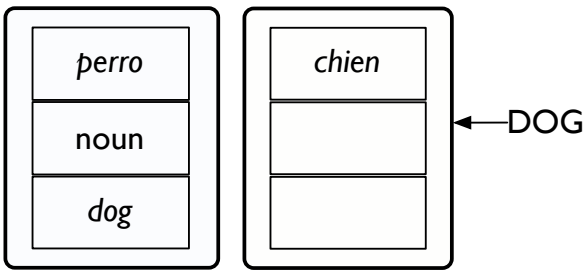

Figure 1: Lexicon sign versus wordlist sign

5.2 Mapping between Form and Concept

A challenge in comparing lexical data among numerous languages is that a complete match between a word's meaning and a general concept rarely occurs within a single language, let alone

across languages (Haspelmath and Tadmor, 2009). Therefore, in order to describe the relationship between form and meaning in a wordlist, we use the term *counterpart*, in the sense developed by Haspelmath and Tadmor (2009). This is in contrast to related notions like *definition* or *translation*. While the meanings found in wordlists could, in some cases, be interpreted as definitions or translations, this is not how they are conceived of in their core function. Rather, they are intended to refer to language-independent concepts which have been determined to be a useful way to begin to explore the lexicon of a language.

A key property of the counterpart relationship is that that even if one particular language (e.g., English or Spanish) is used to refer to a particular concept (e.g., MAN), it is not the idiosyncratic semantics of the word in that language that is used to determine the relevant wordlist entry in the target language. For instance, the meaning of the English word MAN is ambiguous between *human* and *male human* but the term in (1) only refers to *human*. In using a language of wider communication, the goal is to find the closest counterpart in the target language for a general concept, not to translate.

We therefore distinguish between the meanings associated with words in a given language from the more general meanings found in wordlists by using the term *concept* for the latter. Thus, a wordlist entry can be schematized as in (2) where a concept and a lexical item are related by the `hasCounterpart` relation. In attested wordlist entries, the concept is, as discussed, most typically indexed via a language of wider communication and a lexical item is indexed via a transcription representing the lexical item's form.

(2) CONCEPT `hasCounterpart` *lexicalItem*

The counterpart relation is, by design, a relatively imprecise one since a lack of precision facilitates the relatively rapid data collection that is considered an important feature of wordlist creation. The meaning of a given counterpart could be broader or narrower than that of the relevant concept, for example (Haspelmath and Tadmor, 2009, p. 9). In principle, the counterpart relation could be made more precise by specifying, for example, that the relevant relation is *sub-counterpart* for cases where a word in a target language refers to a concept narrower than the one referred to in the word list, as illustrated in (3) for English as the target language. There are other logical kinds of counterpart relationships as well (e.g., *super-counterpart*), and the example is primarily for illustrative purposes. In our database, we only employ the counterpart relation since that was the level of precision found in the original data.

(3) PARENT'S SIBLING `hasSubCounterpart` *aunt*, *uncle*

Though the canonical case for the counterpart relation is that there will be one counterpart for a given concept, this is often not the case in languages and in our data. To take an example from a familiar language, the English counterpart for MOVIE could reasonably be *film* or *movie*, and it is quite easy to imagine a wordlist for English containing both words. The entry in (4) from the dataset we are working with gives an example of this from a wordlist of North Asmat, a language spoken in Indonesia. The concept GRANDFATHER has two counterparts, whose relationship to each other has not been specified in our source.

(4) GRANDFATHER `hasCounterpart` *-ak*, *afak*

Data like that in (4) has led us to add an additional layer in our model for the mapping between concept and form allowing for the possibility that the mapping may actually refer to a group of forms. With more information, of course, one may be able to avoid mapping to a group of forms by, for example, determining that each member of the group is a sub-counterpart of the relevant concept. However, this information is not available to us in our dataset.

5.3 The Concepticon

The concepts found in wordlists have generally been grouped into informally standardized lists. Within our model, we treat these lists as an object to be modeled in their own right and refer to them as *concepticons* (i.e., "concept lexicon"). As will be discussed in section 6, a concepticon is similar to an interlingua, though this connection has rarely, if ever, been explicitly made.

As understood here, concepticons are simply curated sets of concepts, minimally indexed via one or more words from a language of wider communication but, perhaps, also more elaborately described using multiple languages (e.g., English and Spanish) and illustrative example sentences. Concepticons may include terms for concepts of

such general provenance that counterpart words would be expected to occur in almost all languages, such as TO EAT, as well as terms that may occur commonly in only a certain region or language family. For instance, Amazonian languages do not have words for SNOWSHOE or MOSQUE, and Siberian languages do not have a term for TOUCAN (Haspelmath and Tadmor, 2009, p. 5–6).

The concepticon we are employing has been based on three different concept lists. Of these, the most precise and recently published list is the Loanword Typology (LWT) concepticon (Haspelmath and Tadmor, 2009), which consists of 1,460 entries and was developed from the Intercontinental Dictionary Series[2] (IDS) concepticon (1,200 entries). The LWT concepticon often offers more precision for the same concept than the IDS list. For instance, the same concept in both LWT and IDS is described in the LWT list by labeling an English noun with the article *the* (5) in order to clearly distinguish it from a homophonous verb.

(5) **LWT**: THE DUST
IDS: DUST

In addition, certain concepts in the IDS concepticon have been expanded in the LWT list to make it clearer what kinds of words might be treatable as counterparts.

(6) **IDS**: THE LOUSE
LWT: THE LOUSE, HEAD LOUSE, BODY LOUSE

The concepts in LWT and IDS concepticons refer to a wide range of topics but, for historical reasons, they are biased towards the geographical and cultural settings of Europe, southwest Asia, and (native) South America (Haspelmath and Tadmor, 2009, p. 6). The unpublished Usher-Whitehouse concepticon (2,656 entries), used to collect the bulk of the data used in the work described here, includes LWT and IDS concepticons but also adds new concepts, such as WILDEBEEST or WATTLE, in order to facilitate the collection of terms in languages from regions like Africa and Papua New Guinea. Furthermore, certain concepts in the LWT and IDS lists are subdivided in the Usher-Whitehouse concepticon, as shown in (7).

[2] http://lingweb.eva.mpg.de/ids/

(7) 1. **LWT**: TO BREAK
2. **IDS**: BREAK, TR
3. **Usher-Whitehouse**:
   (a) BREAK, INTO PIECES
   (b) BREAK, BY IMPACT
   (c) BREAK, BY MANIPULATION
   (d) BREAK, STRINGS ETC.
   (e) BREAK, LONG OBJECTS
   (f) BREAK, BRITTLE SURFACES

Our unified concepticon combines information from the LWT, IDS, and Usher-Whitehouse lists. This allow us to leverage the advantages of the different lists (e.g., the expanded term list in Usher-Whitehouse against the more detailed concept descriptions of LWT). No wordlist in our database has entries corresponding to all of the concepts in our concepticon. Nonetheless, we now have a dataset with several thousand wordlists whose entries, where present, are linked to the same concepticon, thereby facilitating certain multilingual and cross-lingual applications.

5.4 The Overall Structure of a Wordlist

We schematize our abstract wordlist model in Figure 2. The oval on the left represents the language being described, from which the word forms are drawn (see section 5.1). On the right, the box represents a concepticon (see section 5.3) where the concepts are listed as a set of identifiers (e.g., 1.PERSON) that are associated with labels and related to their best English counterpart. Of course, the labels could be drawn from languages other than English, and other indexing devices, such as pictures, could also be used.

Counterparts from the language being described for the relevant concepts are mapped to blocks of defective signs (most typically containing just one sign, but not always—see section 5.2) which are, in turn, associated with a concept. The schematization further illustrates a possibility not yet explicitly discussed that, due to the relatively imprecise nature of the counterpart relation, one group of forms may be the counterpart for multiple concepts. In short, the mapping between forms and concepts is not necessarily particularly simple.

6 Implementing the Model

We have used the conceptual model for wordlists developed in section 5 to create a wordlist

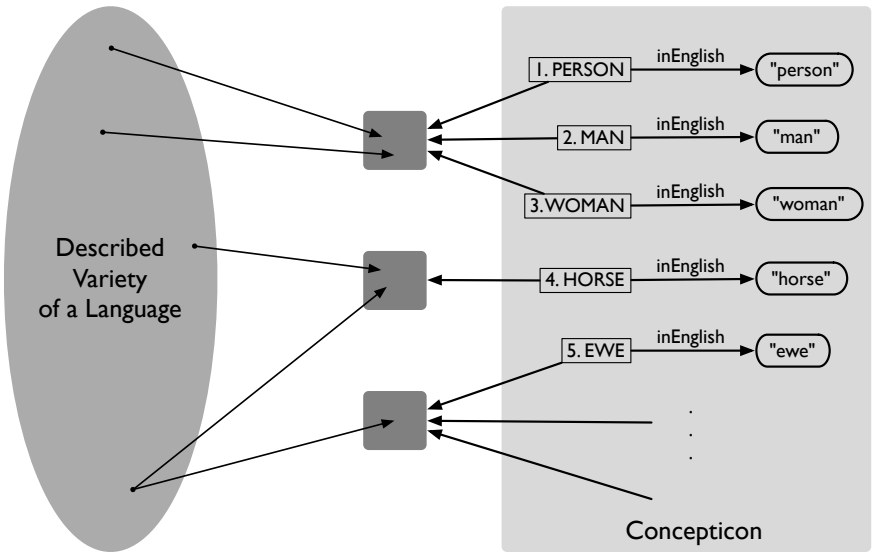

Figure 2: Wordlist modeled as a mapping between a language and a concepticon via blocks of signs

database using Semantic Web technologies, in particular RDF/XML, which we expect to have both research and practical applications.

Each wordlist in our database consists of two components: a set of metadata and a set of entries. The metadata gives the various identifying names and codes for the wordlist e.g., a unique identifier, the ISO 639-3 code, the related Ethnologue language name[3], alternate language names, reference(s), the compilers of the wordlist, etc. All forms in the wordlist are expressed as a sequence of Unicode characters and annotated with appropriate contextual information. In cases where there is more than one form attached to a concept, we create multiple concept-form mappings. We do not explicitly model form groups (see section 2) in our RDF at present since the data we are working with is not sufficiently detailed for us to need to attach information to any particular form group.

Expressing the data encoded in our wordlist database as RDF triples ensures Semantic Web compatibility and allows our work to build on more general work that facilitates sharing and interoperating on linguistic data in a Semantic Web context (Farrar and Lewis, 2007). An RDF fragment describing the wordlist entry in (6) is given in Figure 3 for illustrative purposes. In addition to drawing on standard RDF constructs, we also make use of descriptive linguistic concepts from GOLD[4] (General Ontology for Linguistic Description), which is intended be a sharable OWL ontology for language documentation and description (Farrar and Lewis, 2007). The key data encoded by our RDF representation is the counterpart mapping between a particular wordlist concept (`lego:concept`) drawn from our concepticon and a form (`gold:formUnit`) found in a given wordlist. (The "lego" prefix refers to our internal project namespace.)

An important feature of our RDF encoding is that the counterpart relation does not relate a concept directly to a form but rather to a linguistic sign (`gold:LinguisticSign`) whose form feature contains the relevant specification. This would allow for additional information about the lexical element specified by the given form (e.g., part of speech, definition) to be added to the representation without modification of the model.

Our RDF encoding, at present, is inspired by the traditional understanding of wordlists, building directly on work done by linguists (Haspelmath and Tadmor, 2009). While our use of RDF and an OWL ontology brings the data into a format allowing for much greater interoperability than would otherwise be possible, in order to achieve maximal integration with current efforts in NLP more could be done. For example, we could devise an RDF expression of our model compatible with LMF (Francopoulo et al., 2009) (see section 3).

The most difficult aspect of our model to encode in LMF would appear to be the counterpart relation since core LMF assumes that meanings will be expressed primarily as language-specific *senses*. However, there is work in LMF encod-

[3] `http://ethnologue.com/`

[4] `http://linguistics-ontology.org/`

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:lego="http://purl.org/linguistics/lego/"
   xmlns:gold="http://purl.org/linguistics/gold/">
    <lego:concept rdf:about= "http://www.purl.org/linguistics/lego/concept/106">
        <lego:hasConceptID>106</lego:hasConceptID>
        <lego:hasConceptLabel>the grandfather</lego:hasConceptLabel>
        <lego:hasSource>LEGO Project Unified Concepticon</lego:hasSource>

        <lego:hasCounterpart>
           <gold:LinguisticSign rdf:about=
            "http://www.purl.org/linguistics/North_Asmat_Voorhoeve/12">
             <gold:inLanguage>
              <gold:Language rdf:about=
              "http://www.sil.org/ISO639-3/documentation.asp?id=nks"/>
             </gold:inLanguage>
             <gold:hasForm>
              <gold:formUnit>
                <gold:stringRep>-ak</gold:stringRep>
              </gold:formUnit>
             </gold:hasForm>
             <lego:hasSource>Voorhoeve 1980</lego:hasSource>
           </gold:LinguisticSign>
        </lego:hasCounterpart>

        <lego:hasCounterpart>
           <gold:LinguisticSign rdf:about=
            "http://www.purl.org/linguistics/North_Asmat_Voorhoeve/13">
             <gold:inLanguage>
              <gold:Language rdf:about=
              "http://www.sil.org/ISO639-3/documentation.asp?id=nks"/>
             </gold:inLanguage>
             <gold:hasForm>
              <gold:formUnit>
                <gold:stringRep>afak</gold:stringRep>
              </gold:formUnit>
             </gold:hasForm>
             <lego:hasSource>Voorhoeve 1980</lego:hasSource>
           </gold:LinguisticSign>
        </lego:hasCounterpart>
    </lego:concept>
</rdf:RDF>
```

Figure 3: Wordlist Entry RDF Fragment

ing something quite comparable to our notion of counterpart, namely a *SenseAxis*, intended to support interlingual pivots for multilingual resources (Soria et al., 2009).

As discussed in section 3, the concept labels used in traditional wordlists can be understood as a kind of interlingua. Therefore, it seems that a promising approach for adapting our model to an LMF model would involve making use of the SenseAxes. Because of this we believe that it would be relatively straightforward to adapt our database in a way which would make it even more accessible for NLP applications than it is in its present form, though we leave this as a task for future work.

7 Evaluation

We have identified the following dimensions across which it seems relevant to evaluate our work against the state of the art: (i) the extent to which it can be applied generally to wordlist data, (ii) how it compares to existing wordlist databases, (iii) how it compares to other work which develops data models intended to serve as targets for migration of legacy linguistic data, and (iv) the extent to which our model can create lexical data that can straightforwardly interoperate with other lexical data. We discuss each of these dimensions of evaluation in turn.

(i) The RDF/XML model described here has been successfully used to represent the entire core

dataset being used for this project (see section 2). This represents around 2,700 wordlists and half a million forms, suggesting the model is reasonable, at least as a first attempt. Further testing will require attempting to incorporate wordlist data from other sources into our model.

(ii) Wordlists databases have been constructed for comparative linguistic work for decades. However, there have not been extensive systematic attempts to encode them in interoperable formats to the best of our knowledge, and certainly not involving a dataset of the size explored here. The only comparable project is found in the World Loanword Database (Haspelmath and Tadmor, 2010) (WOLD) which includes, as a possibility, an RDF/XML export. This feature of the database is not explicitly documented, making a direct comparison difficult. An examination of the data produced makes it appear largely similar to the model proposed here. The database itself covers many fewer languages (around 40) but has much more data for each of its entries. In any event, we believe our project and WOLD are roughly similar regarding the extent to which the produced resources can be used for multiple purposes, though it is difficult to examine this in detail at this time in the absence of better documentation of WOLD.

(iii) As discussed in section 3, most work on designing data models to facilitate migration of legacy descriptive data to more modern formats has used representative data rather than producing a substantial new resource in its own right. Furthermore, while the data models have been general in nature, the data encoding has often been in parochial XML formats. By producing a substantial resource in a Semantic Web encoding in parallel with the data modeling, we believe our results exceed most of the comparable work on legacy linguistic data, with the exception of ODIN (Xia and Lewis, 2009) which has also produced a substantial resource.

(iv) Finally, by building our wordlist model around the abstract notion of the linguistic sign, and explicitly referring to the concept of sign through an OWL ontology, we believe we have produced a wordlist data model which can produce data which can straightforwardly interoperate with data from full lexicons since lexicon entries, too, can be modeled as signs, as in Figure 1.

Therefore, while our work cannot be straightforwardly evaluated with quantitative metrics, we believe that on a qualitative level it can be evaluated at or above the state of the art across several key dimensions.

8 Applications

Unlike typical research in NLP, our dataset covers thousands of minority languages that are otherwise poorly represented. Therefore, while our data is sparse in many ways, it has a coverage well-beyond what is normally found.

Crucially, our data model makes visible the similarities between a concepticon and an interlingua, thus opening up a data type produced for descriptive linguistics for use in NLP contexts. In particular, we have created a resource that we believe could be exploited for NLP applications where simple word-to-word mapping across languages is useful, as in the PanImages[5] search of the PanLex project, which facilitates cross-lingual image searching. Such a database can also be readily exploited for machine identification of cognates and recurrent sound correspondences to test algorithms for language family reconstruction (Kondrak et al., 2007; Nerbonne et al., 2007) or to assist in the automatic identification of phonemic systems and, thereby, enhance relevant existing work (Moran and Wright, 2009). We, therefore, think it represents a useful example of using data modeling and legacy data conversion to find common ground between descriptive linguistics and NLP.

Acknowledgments

Funding for the work described here was provided by NSF grant BCS-0753321, and the work is being done in the context of the larger Lexicon Enhancement via the GOLD Ontology project, headed by researchers at the Institute for Language Information and Technology at Eastern Michigan University. Partial funding for the collection and curation of the wordlists was provided by the Rosetta Project (NSF DUE-0333727), along with the Max Planck Institute for Evolutionary Anthropology.

References

Alison Alvarez, Lori Levin, Robert Frederking, Simon Fung, Donna Gates, and Jeff Good. 2006. The MILE corpus for less commonly taught languages. In *NAACL '06: Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume*, pages 5–8. ACL.

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

John Bell and Steven Bird. 2000. A preliminary study of the structure of lexicon entries. In *Proceedings from the Workshop on Web-Based Language Documentation and Description*. Philadelphia, December 12–15, 2000.

Emily M. Bender, Dan Flickinger, and Stephan Oepen. 2002. The Grammar Matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In John Carroll, Nelleke Oostdijk, and Richard Sutcliffe, editors, *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*, pages 8–14, Taipei, Taiwan.

Steven Bird and Gary Simons. 2003. Seven dimensions of portability for language documention and description. *Language*, 79:557–582.

Scott Farrar and William D. Lewis. 2007. The GOLD Community of Practice: An infrastructure for linguistic data on the Web. *Language Resources and Evaluation*, 41:45–60.

Gil Francopoulo, Nuria Bel, Monte George, Nicoletta Calzolari, Monica Monachini, Mandy Pet, and Claudia Soria. 2009. Multilingual resources for NLP in the lexical markup framework (LMF). *Language Resources and Evaluation*, 43:57–70.

Jeff Good. 2004. The descriptive grammar as a (meta)database. In *Proceedings of the E-MELD Workshop on Linguistic Databases and Best Practice*. Detroit, Michigan.

Martin Haspelmath and Uri Tadmor. 2009. The Loanword Typology Project and the World Loanword Database. *Loanwords in the world's languages: A comparative handbook*, pages 1–33. Berlin: De Gruyter.

Martina Haspelmath and Uri Tadmor, editors. 2010. *World Loanword Database*. Munich: Max Planck Digital Library. http://wold.livingsources.org.

Nancy Ide, Daniel Greenstein, and Piek Vossen. 1998. Introduction to EuroWordNet. *Computers and the Humanities*, 32:73–89.

Grzegorz Kondrak, David Beck, and Philip Dilts. 2007. Creating a comparative dictionary of Totonac-Tepehua. In *Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology*, pages 134–141. ACL.

Steven Moran and Richard Wright. 2009. *Phonetics Information Base and Lexicon (PHOIBLE)*. http://phoible.org.

John Nerbonne, T. Mark Ellison, and Grzegorz Kondrak. 2007. Computing and historical phonology. In *Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology*, pages 1–5. ACL.

Sergei Nirenburg, Marge McShane, and Steve Beale. 2004. The rationale for building resources expressly for NLP. In *4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal.

Sebastian Nordhoff. 2008. Electronic reference grammars for typology: Challenges and solutions. *Language Documentation & Conservation*, 2:296–324.

Tom O'Hara, Kavi Mahesh, and Sergei Nirenburg. 1998. Lexical Acquisition with WordNet and the Mikrokosmos Ontology. *In Proceedings of the ACL Workshop on the Use of WordNet in NLP*, pages 94–101.

Alexis Palmer and Katrin Erk. 2007. IGT-XML: An XML format for interlinearized glossed text. In *Proceedings of the Linguistic Annotation Workshop*, pages 176–183. ACL.

Alexis Palmer, Taesun Moon, and Jason Baldridge. 2009. Evaluating automation strategies in language documentation. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 36–44. ACL.

Francesco Ronzano, Maurizio Tesconi, Salvatore Minutoli, Andrea Marchetti. 2010. Collaborative management of KYOTO Multilingual Knowledge Base: The Wikyoto Knowledge Editor. In *Proceedings of the 5th International Conference of the Global WordNet Association (GWC-2010)*. Mumbai, India.

Gary Simons, Brian Fitzsimons, Terence Langendoen, William Lewis, Scott Farrar, Alexis Lanham, Ruby Basham, and Hector Gonzalez. 2004. The descriptive grammar as a (meta)database. In *Proceedings of the E-MELD Workshop on Linguistic Databases and Best Practice*. Detroit, Michigan.

Claudia Soria, Monica Monachini, and Piek Vossen. 2009. Wordnet-LMF: Fleshing out a Standardized Format for Wordnet Interoperability. In *International Workshop on Intercultural Collaboration (IWIC)*, pages 139–146. ACM.

Tokunaga Takenobu, Nicoletta Calzolari, Chu-Ren Huang, Laurent Prevot, Virach Sornlertlamvanich, Monica Monachini, Xia YingJu, Shirai Kiyoaki, Thatsanee Charoenporn, Claudia Soria, and Hao, Yu. 2006. Infrastructure for standardization of Asian language resources In *Proceedings of the COLING/ACL Main Conference Poster Sessions*, pages 827–834. ACL.

Piek Vossen. 2004. EuroWordNet: A multilingual database of autonomous and language-specific wordnets connected via an Inter-Lingual-Index. *International Journal of Lexicography*, 17:161–173.

Fei Xia and William D. Lewis. 2009. Applying NLP technologies to the collection and enrichment of language data on the Web to aid linguistic research. In *LaTeCH-SHELT&R '09: Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*, pages 51–59. ACL.