The LEGO Unified Concepticon

Jeff Good [a,*], Shakthi Poornima [b], Timothy Usher [c]
[a] *University at Buffalo, Department of Linguistics, 609 Baldy Hall, Buffalo, NY 14260, USA*
[b] *inome, 500 108th Avenue, Bellevue, WA 98004, USA*
[c] *Seattle, WA, USA*

Abstract. The most widely available kind of linguistic data from a cross-linguistic perspective are wordlists, where forms from a language are paired with generalized semantic concepts to indicate the best counterpart for those concepts in that language. Wordlists can be readily understood as a pre-digital form of linked data insofar as standardized concept lists have frequently been employed in their construction to facilitate cross-linguistic comparison, in particular to aid in efforts to ascertain patterns of relatedness among large sets of languages. This paper describes the LEGO Unified Concepticon, a resource which expresses which concepts in a number of widely used concept lists can be understood as the same in order to allow wordlists collected using different concept lists to be more readily compared. While the resource itself contains relatively limited information on the concepts it describes, it can nevertheless serve as a means to link together forms across wordlists and has already been employed to these ends in the creation of linked wordlists across more than a thousand languages. Moreover, it has the potential to serve as the foundation for further applications in cross-linguistic semantic comparison using linked data.

Keywords: lexicons, comparative linguistics, semantics, linked data

1. Introduction to the LEGO Unified Concepticon

The primary goal of the Lexical Enhancement via the GOLD Ontology project (LEGO) (http://lego.linguistlist.org) was to develop standards and tools to facilitate interoperation of lexical data, with a focus on lexicons and wordlists describing low-resource languages. In order to facilitate interoperation among wordlists specifically, the LEGO Unified Concepticon was developed, and the purpose of this paper is to introduce this resource, describe its role in facilitating wordlist interoperation, and suggest possible additional uses and potential refinements. We use the term *concepticon* to refer to a "concept lexicon", that is an object which is structurally comparable to a natural language lexicon except that, rather than describing lexical items in an attested language, it describes abstract concepts which are understood to be language-independent.

The LEGO Unified Concepticon is available in RDF/XML format at http://lego-wordlists.googlecode.com/files/Lego-Unified.rdf. It is released under the MIT license. The current version was most recently modified in August 2010. As a composite resource, its content has a number of creators. As a linked data resource it was compiled by the first two authors of this paper. The third author was one of the primary content contributors and collaborated especially closely with Paul Whitehouse whose work was instrumental in setting the stage for the creation of the linked data resource described here.

In the rest of this description, we first give a general overview of the structure of traditional wordlists (Section 2). We then discuss the provenance and technical details of the LEGO Unified Concepticon (Section 3)

[*]Funding for the work described here has been provided by NSF grant BCS-0753321, as part 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. Partial funding for the collection and curation of the resources described here was provided by the Rosetta Project (NSF DUE-0333727), along with the Max Planck Institute for Evolutionary Anthropology (MPI EVA). Numerous people have contributed to the development of the conceptual model underlying the resource described here, and we would like to especially single out Gary Simons in this regard.

and describe how it has been used to facilitate wordlist interoperation (Section 4). We conclude by discussing some of its current limitations and indicate ways in which it could be improved (Section 5).

Before moving on, we would like to emphasize that the work described here was primarily motivated by a desire to convert legacy datasets to contemporary interoperable formats. This has meant that we emphasized re-encoding of available information over adding clearly desirable new kinds of information, in particular relating the concepts of our concepticon to available semantic ontologies. We will return to this point in Section 5 and make relevant other remarks throughout. More detailed discussion of conceptual and practical considerations that led to the development of the resource described here can be found in [11].

2. Wordlists as a linguistic data type

Wordlists—i.e., simple lexical resources consisting of the pairing of a form with an abbreviated label expressing some meaning—have the greatest cross-linguistic coverage of any kind of descriptive resource, offering lexical data on perhaps a quarter or more of the world's languages. As such, there are clear motivations to making them available in interoperable formats. 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*

The key features of a canonical 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*) which has been determined to be the counterpart for that concept within that language. Most typically, the elements indexing the relevant concepts are themselves words drawn from languages of wider communication (e.g., English, Spanish. French, etc.), though this is done for convenience, and not because of any specific principle. Actual wordlists can deviate, in various ways, from the canonical presentation given in (1), for instance by including additional information such as part of speech or refinements to the meaning associated with a given concept label.

Wordlists differ from other kinds of lexical resources, such as dictionaries, in being constructed on the basis of a mapping from a set of meanings to a set of forms, rather than the reverse. Our use of the word *concept* is intended to distinguish between the general sorts of meanings that are involved in the construction of wordlists and the language-specific meanings that are associated with the actual words found in some language. For instance, the English word *man* can be used either to refer to male humans or humans in general. However, when used as a concept label, as in Figure 1, it will typically be understood to solely refer to a male human, while the label PERSON would be used for the more general concept.

If the core conceptual construct upon which lexicons and wordlists are built is the linguistic sign consisting of a triple associating form, grammatical information, and meaning, then wordlists can be understood as consisting of a set of defective signs in two ways. First, they contain information on the form and meaning parts of the triple, but not the grammatical part. Second, the meaning information they contain 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 dictionary 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 information or semantic information is associated with the form, but there is an indication that this lexical item is the closest counterpart to the general concept DOG in French.

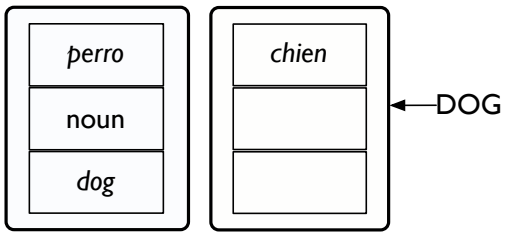

Fig. 1. Lexicon sign versus wordlist sign

In the case of a word like *chien*, it is not only the counterpart of DOG in French, but also the translational equivalent of the English word *dog*, which the concept label mnemonically refers to. However, this cannot be

assumed to be the case in general. As already discussed above, for example, the English concept label MAN is generally opposed to PERSON in concepticons to refer only to male humans. However, in actual English usage it will not always be the case that it should be translated with that sense, of course.

A traditional wordlist, based on the model described here, consists of both a set of forms and a set of concept labels, as well as statements mapping forms to concepts. We refer to the set of concept labels associated with a wordlist as a *concepticon* here, and discuss this object in further detail in the next section. As will be seen, our linked data representation of wordlists deviates from the traditional model in not directly containing concept labels but, rather, references to concepts described via labels in an external concepticon.

3. Towards a standard concepticon

3.1. Existing concepticons

Two important aspects of concept lists used in the construction of wordlists are that (i) they are often re-used, and, therefore, informally standardized and (ii) they can be curated to a greater or lesser extent. The most important form of curation is the selection of the concepts themselves, with the best known criterion of selection being concepts whose associated words across languages are less likely to be borrowed, making them a good basis for genealogical classification. This is the basis of the so-called Swadesh wordlists, and a motivating factor behind more recent efforts at concepticon construction [13]. Other forms of curation are possible as well, such as placing the concepts within a taxonomy or elaborating the way they are referred to, for instance by using labels from multiple languages, though the LEGO Unified Concepticon is not curated to that degree. In discussing concepticons here, we are referring only to curated sets of concepts used for linguistic purposes, which makes them distinct from efforts like DBpedia [1], which does allow reference to a range of concepts but was not designed specifically for linguistic resources (see also Section 5).

Understood in a broad sense, it is likely that hundreds of different concepticons have been used at one time or another, especially if one takes a "splitting" approach and considers variants of common concepticons to be distinct from one another. At the same time, there is often significant overlap among them, especially since relatively large concepticons (on the order of, say, 1000 entries) will often contain all of the terms found in smaller concepticons (on the order of, say, 100 entries), as seen in [13]. A clear example of this is work done in the context of the Automated Similarity Judgment Program which uses a forty-entry concepticon that is explicitly understood as a subset of a Swadesh concepticon of 100 terms [14]. In some cases, relatively large wordlists are intended to be of general use, just with an expanded set of concepts. In other cases, they are specifically designed to augment general concepts with concepts relevant to specific language families or parts of the world, for instance sub-Saharan Africa [12].

The goal of the construction of the LEGO Unified Concepticon was to create a new concepticon, using a contemporary interoperation format which incorporated concepts from three pre-existing general concepticons, those of the Intercontinental Dictionary Series [6], the Loanword Typology project [5], and one devised by the third author in collaboration with Paul Whitehouse. The choice of these three concepticons was purely practical in nature: Wordlists associated with these three projects were processed as part of the larger LEGO project, and a mechanism was needed to allow interoperation among the wordlists despite their use of different concepticons. The construction of the unified concepticon was greatly facilitated by the fact that the third author, in collaboration with Paul Whitehouse, had already put significant effort into constructing a table expressing mappings between the concepts of these concepticons, something which cannot be fully automated since the ideal mappings require knowledge of the meanings referred to by the labels associated with each concept. The (relatively modest) contribution of the LEGO project, in this regard, was devising a means to express these mappings using linked data.

3.2. The data model of the unified concepticon

The present LEGO Unified Concepticon is based on a very simple data model. It consists of a set of "container" concepts each associated with a unique identifier, with each container concept described by references to equivalent concepts in the three concepticons it unifies. (The container concepts do not always contain references to three other concepts since not all concepts are represented in all three concepticons.) In addition, each container is associated with a preferred label for the concept in English. This la-

bel draws on what we believe to be the most informative label in the concepticons that are being unified—though this was an implementation decision and the model itself allows for any preferred label to be specified. Finally, since one of the source concepticons [5] already was available in linked data form, our unified concepticon includes an explicit link between its concepts and ours stating that they should be interpreted as the same. The container concepts in our unified concepticon can readily be associated with further information if deemed desirable (see Section 5).

The structure of the concepticon is schematized in Figure 2. The concepticon is a container (associated with metadata not depicted in the figure), which consists of unified concepts which are themselves containers for concepts from the concepticons. For purposes of documentation (and potential legacy applications) the concepts from the legacy concepticons are associated with a string indicating their identifier (usually a number) in their original source. However, they are not otherwise given the same level of information as the unified concepts since they are not intended to be the basis for future interoperation. Our current concepticon only gives one preferred label for each concept, but additional labels could be added, for instance in languages other than English (and, of course, other kinds of information could be added as well).

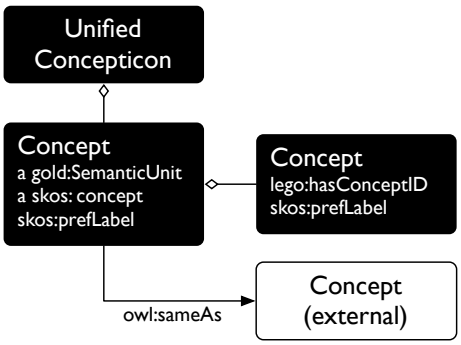

Fig. 2. Concepticon data structure

As indicated in Figure 2, we make use of concepts from the General Ontology for Linguistic Description (GOLD) [2] wherever possible as well as SKOS [9], in some cases.

4. Linking wordlists via the concepticon

The primary application of the LEGO Unified Concepticon has been to link together forms from legacy wordlists that have been associated with the same concept. Specifically, we express our wordlists as consisting of sets of signs associated only with information about their forms and a reference to the concept they are associated with in the unified concepticon. The structure of these word lists is depicted in Figure 3.

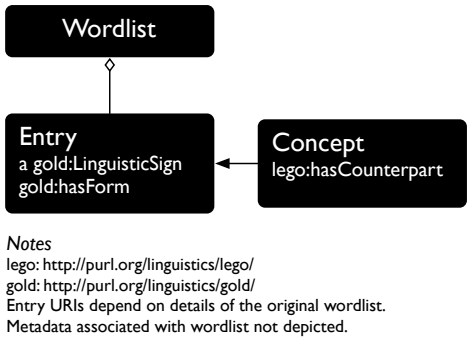

Fig. 3. Word list data structure

As can be seen, the information encoded in these wordlists is quite sparse—for instance, it only includes concept identifiers, not concept labels. Therefore, in order to reconstruct the information associated with traditional wordlists (as in (1)), the unified concepticon must be merged with the wordlist.

For purposes of illustration, we have made available three RDF/XML versions of the wordlists processed by the project at http://code.google.com/p/lego-wordlists/downloads/. Each of these is drawn from a different data set. They are (i) a wordlist for the Abar variety of the language associated with ISO 639-3 code [mij] as found in the wordlists collected by the third author and Paul Whitehouse, originally drawn from [4], (ii) a wordlist for Saramaccan [srm], based on an original drawn from the World Loanword Database [3], and (iii) a wordlist for Archi [aqc] prepared by Madzhid Khalilov for the Intercontinental Dictionary Series [7].

In our own project, we maintain a convention of linking all wordlists through the container concepts of our unified concepticon. However, it would also be possible to link to the concepts of our source concepticons and use the mapping implied by our container concepts to achieve interoperation, which, imaginably, could be more straightforward for some projects. All of our forms are represented by Unicode strings that

can be accessed through a `gold:stringRep` property of a `gold:FormUnit`.

While we use wordlists containing very sparse information for our own project, there is no technical reason why our model could not be extended for use with richer lexicons. Our `lego:hasCounterpart` relation is understood to map from concepts to linguistic signs, and, as such, the properties of the signs could, in principle, be specified in more detail than what our wordlists provide. It would, in fact, be quite straightforward to, in effect, embed a wordlist inside a full dictionary using an appropriate set of counterpart specifications. Moreover, while we have chosen to model our wordlists using the abstract notion of the linguistic sign rather than the more concrete notion of lexical entry, as adopted, for instance, by the Lexicon Model for Ontologies (lemon) [8], there is no inherent reason why the mapping could not be made to such entities.

Finally, we only make use of a simple counterpart relation, but our model could be straightforwardly extended to cases where one is not dealing with an exact counterpart but, rather narrower or broader matches, by creating more precise counterpart specifications such as *subcounterpart*, *supercounterpart*, etc. (see [11]).

5. Limitations and possible future directions

The fact that the LEGO Unified Concepticon was designed primarily as a means to allow legacy datasets to be expressed using a contemporary linked data model means that the information it contains is relatively limited since only limited information was required to re-encode the legacy data. There are relatively clear ways that it could be augmented, with the most obvious being adding ontological structure to the concepts found within it. Indeed, existing concepticons, such as that used for the Loanword Typology Project [5] (which was used in constructing our unified concepticon) already do something like this by grouping concepts into high-level categories such as *animals*, *food and drink*, etc. Another obvious improvement would be to add links from concepts in the unified concepticon to data sources that describe those concepts using more than a label, such as DBpedia [1]. In addition, of course, the set of concepts in the concepticon could be expanded to include those found in other wordlists and other concepticons using the concepts found in our concepticon could also be given appropriate reference in the container concepts on which our concepticon is built.

Given the limited information found in the LEGO Unified Concepticon, one may question the value it has over, for instance, an effort like DBpedia. From our perspective, the unified concepticon fulfills a quite distinct function. While DBpedia offers rich descriptions of concepts, our concepticon is sparser but more useful as a pivot through which different lexical resources can be interrelated. In particular, since our concepticon represents a "bottom-up" effort based on the kinds of data found in actual linguistic resources, it can efficiently allow the information found in those resources to be exploited in a Semantic Web context. Indeed, if many of the concepts in our concepticon could be linked to concept descriptions found in in DBpedia, then any wordlist whose forms can be readily linked to concepts in our concepticon will automatically link to DBpedia as well, and this should be a more straightforward task than linking directly to DBpedia since our concepticon is based on already common practice.

The concepts found in the LEGO Unified Concepticon, therefore, can be conceived as playing a role comparable to symbols drawn from the International Phonetic Alphabet (IPA) in the linked data system of the Phonetics Information Base and Lexicon [10]. They are not intended to be the "end point" of a semantic analysis but, rather, serve as a framework on which more detailed semantic specifications can be built. Furthermore, like the IPA, their value does not lie in having been devised using consistent ontological principles (or something comparably rigorous) but, rather, in the fact that they are widely used and understood by the community most actively collecting linguistic data—linguists themselves. For languages like English, there is sufficient interest to create custom-built semantic resources, but, for the vast majority of the world's languages, limited resources are likely to make the "universal pivot" approach of the LEGO Unified Concepticon considerably more viable in the long run.

References

[1] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. DBpedia–A crystallization point for the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7:154–165, 2009.

[2] Scott Farrar and D. Terence Langendoen. An OWL-DL implementation of GOLD: An ontology for the Semantic Web. In Andreas Witt and Dieter Metzing, editors, *Linguistic mod-*

eling of information and markup languages: Contributions to language technology, pages 45–66. Berlin: Springer, 2009.

[3] Jeff Good. Saramaccan vocabulary. In Martin Haspelmath and Uri Tadmor, editors, *World Loanword Database*. Munich: Max Planck Digital Library, 2009.

[4] Cameron Hamm, Jason Diller, Kari Jordan-Diller, and Ferdinand Assako a Tiati. *A rapid appraisal survey of Western Beboid languages (Menchum Division, Northwest Province)*. SIL Electronic Survey Reports: SILESR 2002-014, 2002.

[5] Martin Haspelmath and Uri Tadmor, editors. *World Loanword Database*. Munich: Max Planck Digital Library. http://wold.livingsources.org/, 2009.

[6] Mary Ritchie Key and Bernard Comrie, editors. *The Intercontinental Dictionary Series*. http://lingweb.eva.mpg.de/ids/, 2012.

[7] Madzhid Khalilov. Archi. In Bernard Comrie and Mary Ritchie Key, editors, *Intercontinental Dictionary Series*. http://lingweb.eva.mpg.de/ids/, 2008.

[8] John McCrae, Dennis Spohr, and Philipp Cimiano. Linking lexical resources and ontologies on the Semantic Web with lemon. In *The Semantic Web: Research and applications*, pages 245–259. Berlin: Springer, 2011.

[9] Alistair Miles and Sean Bechhofer, editors. *SKOS Simple Knowledge Organization System reference*. W3C Recommendation. http://www.w3.org/TR/skos-reference, 2009.

[10] Steven Moran. Using linked data to create a typological knowledge base. In Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann, editors, *Linked data in linguistics: Representing and connecting language data and language metadata*, pages 129–138. Berlin: Springer, 2012.

[11] Shakthi Poornima and Jeff Good. Modeling and encoding traditional wordlists for machine applications. In Fei Xia, William Lewis, and Lori Levin, editors, *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground (NLPLING '10)*, pages 1–9. Stroudsburg, PA: Association for Computational Linguistics, 2010.

[12] James Roberts and Keith Snider. *SIL Comparative African Wordlist (SILCAWL)*. SIL Electronic Working Papers: SILEWP 2006-005, 2006.

[13] Uri Tadmor, Martin Haspelmath, and Bradley Taylor. Borrowability and the notion of basic vocabulary. *Diachronica*, 27:226–246, 2010.

[14] Søren Wichmann, André Müller, and Viveka Velupillai. Homelands of the world's language families: A quantitative approach. *Diachronica*, 27:247–276, 2010.