Lexvo.org: Language-Related Information for the Linguistic Linked Data Cloud

Gerard de Melo
ICSI, Berkeley, CA, USA

Abstract Lexvo.org brings information about languages, words, and other linguistic entities to the Web of Linked Data. It defines URIs for terms, languages, scripts, and characters, which are not only highly interconnected but also linked to a variety of resources on the Web. Additionally, new datasets are being publishing to contribute to the emerging Linked Data Cloud of Language-Related information.

Keywords: languages, lexical information

1. Introduction

Lexvo.org is a service that publishes and provides information about various aspects of human language to the Linked Data cloud and the Semantic Web. Language is the basis of human communication and the key to the tremendous body of written knowledge available on the Web. Due to the ubiquity of textual data, the value of lexical and other linguistic data is increasingly being recognized in the Semantic Web and Digital Library communities, among others. More recently, the value of interoperable linguistic data has finally also been receiving an increased amount of attention in linguistics and lexicography. The Open Linguistics Working Group [2] of the Open Knowledge Foundation has brought researchers working in this area together and has begun to proselytize and educate by organizing workshops and meetings. These developments are leading to the emergence of a significant new part of the Linked Data Cloud that focuses on linguistic data. This article describes Lexvo.org and its contribution to this emerging cloud of Linguistic Linked Data[1].

2. Language Information

2.1. Language Identification

In many different application scenarios, it is important to be able to specify a given human language. For example, one might want to state that a book is written in a particular language, or that a user prefers a particular language.

One of the main motivations for the Web of Linked Data is the idea of liberating data from traditional data silos by using shared global identifiers rather than database-dependent strings of characters. For instance, instead of having a "language" column in a database that might contains values like “engl.”, “grk.”, “albn.” (or “en”, “el”, “sq”), data publishers and application developers can publish data on the Web using global identifiers (URIs) like `http://lexvo.org/id/iso639-3/eng` and `http://lexvo.org/id/iso639-3/ell`. Such URIs are part of a common global vocabulary that many different data sets on the Web share. This makes it much easier to see that two databases are referring to the same thing as opposed to when one uses “el” and the other uses “grk.”. Additionally, these URIs are also dereferenceable, meaning that humans can open them in their browser and software tools can download machine-readable data to find out more about what the URI identifies.

[1] This article describes the 2013-01-02 version of the data set.

The ubiquitous two-letter ISO 639-1 codes for languages ("en", "fr", etc.) are defined for no more than around 180 languages. While the slightly more recent ISO 639-2 standard provides around 500 three-letter codes and hence covers the major languages of the world, it cannot by any means be considered complete, lacking codes for Ancient Greek, American Sign Language, and of course thousands of rare minority languages spoken around the world. The same holds for URIs derived from the English Wikipedia, which merely describes a few hundred languages.

To address this situation, Lexvo.org defines URIs of the form `http://www.lexvo.org/id/iso639-3/eng` for all of the 7 000 languages covered by the ISO 639-3 standard. While the Library of Congress has published ISO 639-2 as a controlled vocabulary (based on SKOS), there is no good Linked Data alternative to Lexvo.org for ISO 639-3. Lexvo.org's language identifiers are used by British Library[2], the Spanish National Library (*datos.bne.es*), and the French academic catalog Sudoc, among others.

Obviously, even ISO 639-3 cannot be complete in the sense of covering every possible dialect. However, the standard has well-defined procedures for adding new identifiers and is regularly updated. It thus serves as a good practical solution for most language identification needs. The Glottolog project [15] provides a solution for those that require more fine-grained ways to identify language definitions by individual linguists.

2.2. Language Descriptions

Lexvo.org provides extensive descriptions of each language, based on sources like Wikipedia and the Unicode CLDR. These are often expressed using properties and classes from the Lexvo Ontology[3]. Examples include language names in many languages, geographical regions[4], identification codes, relationships between languages, etc. Information about ancient and constructed languages come from the Linguist List, which officially maintains inventories of them for use in ISO 639-3.

In order to facilitate linking to Lexvo.org, the site provides mapping tables for MARC 21 / USMARC language codes and also defines an alternative set of IDs based on the commonly used 2-letter ISO 639-1 language codes.

Lexvo.org's language identifiers are connected to DBpedia, YAGO, and other existing sites. Additonally, for nearly 400 languages, the service now provides links to text samples (specifically, the UN Declaration of Human Rights).

2.3. Language Families and Collections

Language families (or collections) and their relationships are described using URIs based on ISO 639-5, e.g. `http://www.lexvo.org/id/iso639-5/sit` for the Sino-Tibetan languages. Some of the identifiers refer not to language families per se, but to other types of collections (e.g. sign languages). Lexvo.org draws information about these language families from Wikipedia and WordNet.

An extensive language family hierarchy is provided, fully integrated into a general-purpose WordNet-based word sense hierarchy (see Section 3.2). From Mandarin Chinese, for instance, one can thus navigate to general Chinese, the Sinitic languages, and the Sino-Tibetan languages.

2.4. Scripts and Characters

The language identifiers are linked to identifiers that have been set up for the scripts defined by the ISO 15924 standard. Examples include Cyrillic, Indian Devanagari, and the Korean Hangul system. By extracting Unicode Property Values from the Unicode specification, these script URIs have also been connected with the specific characters that are part of the respective scripts.

URIs of the form `http://www.lexvo.org/id/char/5A34` are provided for each of the several thousand characters defined by the Unicode standard. A large number of Unicode code points represent Han characters used in East Asian languages. Additional data from the Unihan database and other sources has been extracted to provide semantic information about such characters.

2.5. Phones

Lexvo.org was recently extended to include phonetic information. For a given phone, it provides dif-

[2] `http://www.bl.uk/bibliographic/datafree.html`

[3] `http://lexvo.org/ontology`

[4] Geographical regions are identified using URIs based on ISO 3166 / UN M.49, which have also been connected to the GeoNames dataset.

ferent representations (IPA, X-SAMPA, Arpabet, etc.) and properties (e.g. labiodental, plosive, etc.)[5].

3. Lexical Information

3.1. *Identifiers for Words*

From the start, Lexvo.org also focused on describing words (or terms) and their properties. String literals cannot serve as subjects of an RDF triple. In order to express knowledge about words, some ontologies have defined OWL classes that represent words or other terms in a language. However, data publishers still needed to create the URIs for individual terms on an ad hoc basis. For instance, the W3C draft RDF/OWL Representation of WordNet [16] has defined URIs for the words covered by the WordNet lexical database [7].

In order to provide data publishers with a simple way of identifying any word using an URI, Lexvo.org proposed a standard, uniform scheme for referring to terms in a specific language. Data publishers and developers can obtain and work with such URIs by using a simple Java API.

3.1.1. *Formal Semantics*

Formally, different levels of abstraction could be chosen to refer to words. For practical reasons, the service focuses mostly on the pure surface form.

Within a specific language, its term URIs do not distinguish the meanings of polysemous or homonymic words, e.g. the verb and noun meanings of the English term “call”, or the animal noun “bear” from the verb “bear”. This is because, typically, one wishes to look up terms in a given knowledge base (e.g. in a thesaurus or a dictionary) without already knowing what word senses exist. Lexvo.org thus treat two words with identical surface forms in the same language as one single term. Such distinctions are instead only made at the word sense level, as described later on.

In contrast, Lexvo.org does, however, consider the language of a term relevant to its identity. Thus, the Spanish term “con”, which means “with”, is treated as distinct from the French term “con”, which means “idiot”. This level of abstraction allows us to model relationships between words in different languages using simple RDF triples. If one instead used URIs based on pure string literals without language information, it would be necessary to specify the two respective languages using reification for each original statement.

Different word forms are treated as distinct terms. Here, however, there are a few minor subtleties of term identity regarding string encoding. For multilingual applications, the ISO 10646 / Unicode standards offer an appropriate set of characters for encoding strings. Since Unicode allows encoding a character such as “à” in either a composed or in a decomposed form, NFC normalization [3] is applied to avoid duplicate entities. Formally, given a term t in a language L, the URI is constructed as follows:

- The term t is encoded using Unicode, and the NFC normalization procedure [3] is applied to ensure a unique representation. Conventional unnormalized Unicode allows encoding a character such as “à” in either a composed or in a decomposed form.
- The resulting Unicode code point string is encoded in UTF-8 to obtain a sequence of octets.
- These octet values are converted to an ASCII path segment by applying percent-encoding as per RFC 3986. Unacceptable characters as well as the “%” character are encoded as triplets of the form “%4D” with the respective octet value stored as two upper-case hexadecimal digits.
- The base address `http://www.lexvo.org/id/term/` and the ISO 639-3 code for the language L followed by the “/” character are prepended to this path segment to obtain a complete URI.

Fortunately, Lexvo.org’s Java API hides most of these details from data publishers, instead providing a very simple interface to obtain a term URI given a string and its language.

3.1.2. *Term Descriptions and Links*

Capturing links to terms is particularly significant in light of the important role of natural language for the Semantic Web. In general, a non-information resource URI string itself does not convey reliable information about its intended meaning, because a URI (including class or property names) can be chosen quite arbitrarily. Oftentimes the meaning is specified using natural language definitions or characteristic labels. From a semantic perspective, however, RDFS `label` is merely an annotation property that provides human-readable display labels, which can be identifier strings such as “minCardinality”.

[5] The *phoible.org* project will provide a more extensive description of the phonetic properties of different languages.

In order to make the meaning of URIs more formal, Lexvo.org proposes explicitly linking to term URIs of one or more natural languages using a lexicalization property, whenever appropriate. Such a property formally captures the semantic relationship between a concept and its natural language lexicalizations or between an arbitrary entity and natural language terms that refer to it.

Much of the multilingual information that Lexvo.org provides about words comes from Wiktionary, a well-known effort to collaboratively create dictionaries on the Web. Links to Wiktionary are now extracted from the English, Catalan, French, German, Greek, Portuguese, Spanish, and Swedish Wiktionary versions. Lexvo.org also provides part-of-speech tag information extracted from Wiktionary, explaining whether a word functions as a noun or an adjective, for instance.

Term entities are linked to the respective concepts in external resources, such as the GEneral Multilingual Environmental Thesaurus (GEMET), the United Nations FAO AGROVOC thesaurus, the US National Agricultural Library Thesaurus, EuroVoc, and the RAMEAU subject headings. Links to upper ontologies such as OpenCyc are present as well.

3.2. Word Senses

Ideally, there would be a universal registry of word meanings that could serve as a hub in the Linked Data world that anyone could link to. Unfortunately, there are several challenges: (1) There is no obvious universal inventory of word senses. Even authoritative dictionaries differ significantly in the senses they enumerate for a given word [9]. In sources like Wiktionary, the senses vary over time as editors make changes to the pages. (2) Even if we had an adequate registry of senses, most existing linguistic resources do not make sense distinctions, so we cannot easily link them to the right senses, as automatic disambiguation is known to be very error-prone. (3) Even when a resource does distinguish senses, these are unlikely to be compatible with the chosen inventory. Empirical studies show that senses in different data sets often do not align in a clean way [4]. Some even go as far as arguing that word senses are not necessarily a useful notion at all [11]. For these reasons, Lexvo.org mainly focuses on term-based URIs that do not distinguish word senses.

The service does, however, also include word sense-specific URIs based on Princeton's WordNet lexical database [7]. WordNet is the most widely used sense inventory in natural language processing and thus the closest we have to a universal word sense inventory. While designed for English, WordNet's senses have also been used for many other languages [5]. Its identifiers have been linked to YAGO, SUMO, OpenCyc, VerbNet, and numerous other data sets (some of which will be discussed later on). Lexvo.org links English terms to their respective WordNet synsets, based on information from WordNet 3.0.

4. Towards a Linguistic Linked Data Cloud

Lexvo.org is backed by a Linked Data server infrastructure that makes its URIs dereferenceable and part of the Linked Data cloud. In 2010, Bernard Vatant decided to deprecate the lingvoj.org service, which had been publishing language identifiers based on Wikipedia, instead redirecting users to Lexvo.org, which provides richer descriptions of over an order of magnitude more languages. With his help, a number of data publishers recognized the value of Lexvo.org's language descriptions.

More recently, several third parties have created new linguistic datasets, leading to the beginnings of a cloud of Linguistic Linked Data [2]. In order to strengthen and accelerate these efforts, Lexvo.org is publishing several new datasets. All of these are linked to Lexvo.org at the word level and in some cases also at other levels. Predicates come from the Lexvo Ontology as well as other existing ontologies. The datasets fall into several categories.

4.1. Semantic Information

Roget's Thesaurus is the most well-known English thesaurus, but the standard distribution comes in a text format that is hard to parse. Lexvo.org hosts an RDF version of Roget's Thesaurus.

The WordNet Evocation dataset [1] provides data about associations between words, e.g. between "car" and "road".

WordNet Domains delivers thematic domain markers for WordNet synsets. Lexvo.org publishes an RDF conversion of the extended WordNet 3.0-aligned version produced as part of the Multilingual Central Repository 3.0 [10].

4.2. Cross-Linguistic Data

Etymological WordNet [6] contributes links between words that are not semantic but etymological or derivational in nature.

4.3. Semantic Frames and Roles

Lexvo.org maintains RDF datasets for FrameNet [8], PropBank [12] and NomBank [13], three resources that model the phrase- and sentence-level semantic frames and roles that words express.

4.4. Sentiment Analysis Data

Lexvo.org hosts a Linked Data version of the MPQA Subjectivity Lexicon [17], which supplies subjectivity and sentiment polarity labels for words. Additionally, the AFINN dataset has been converted [14], offering more fine-grained numeric sentiment valency scores.

4.5. Speech Data

An RDF version of the CMU Pronunciation Dictionary has been produced, in which the original encoding has been converted to IPA.

5. Conclusion

In summary, Lexvo.org defines standard identifiers for languages and language families, words and word senses, scripts, characters, etc. Additionally, it publishes a broad spectrum of language-related information. The service is being used by numerous third parties. This ecosystem of data constitutes a useful basis for applications in linguistics, natural language processing, and other areas that benefit from the more interlinked and interoperable nature of the resources. We believe that this provides significant incentives for third parties to contribute to the Linguistic Linked Data cloud.

References

[1] Jordan Boyd-Graber, Christiane Fellbaum, Daniel Osherson, and Robert Schapire. Adding dense, weighted connections to WordNet. In *Proc. GWC 2006*, 2006.

[2] Christian Chiarcos, Sebastian Hellmann, Sebastian Nordhoff, Steven Moran, Richard Littauer, Judith Eckle-Kohler, Iryna Gurevych, Silvana Hartmann, Michael Matuschek, and Christian M. Meyer. The Open Linguistics Working Group. In *Proc. LREC 2012*, 2012.

[3] Mark Davis and Martin Dürst. Unicode normalization forms, rev. 29. Technical report, Unicode, 2008.

[4] Gerard de Melo, Collin F. Baker, Nancy Ide, Rebecca Passonneau, and Christiane Fellbaum. Empirical comparisons of MASC word sense annotations. In *Proc. LREC 2012*, Paris, France, 2012.

[5] Gerard de Melo and Gerhard Weikum. Towards a universal wordnet by learning from combined evidence. In *Proc. CIKM 2009), publisher = ACM, year = 2009, isbn = 978-1-60558-512-3, pages = 513–522, location = Hong Kong, China, doi = http://doi.acm.org/10.1145/1645953.1646020, address = New York, NY, USA,*.

[6] Gerard de Melo and Gerhard Weikum. Towards universal multilingual knowledge bases. In *Proc. GWC 2010*, pages 149–156, 2010.

[7] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, May 1998.

[8] Charles J. Fillmore, Christopher R. Johnson, and Miriam R.L. Petruck. Background to FrameNet. *International Journal of Lexicography*, 16.3:235–250, 2003.

[9] C.J. Fillmore and B.T.S. Atkins. Describing polysemy: The case of 'crawl'. *Polysemy: Theoretical and computational approaches*, pages 91–110, 2000.

[10] Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. Multilingual Central Repository version 3.0. In *Proc. LREC 2012*, 2012.

[11] A. Kilgarriff. I don't believe in word senses. *Computers and the Humanities*, 31(2):91–113, 1999.

[12] Paul Kingsbury and Martha Palmer. Propbank: The next level of treebank. In *Proceedings of Treebanks and Lexical Theories*, Växjö, Sweden, 2003.

[13] Adam Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. The nombank project: An interim report. In A. Meyers, editor, *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 24–31, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics.

[14] Finn Årup Nielsen. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *CoRR*, abs/1103.2903, 2011.

[15] Sebastian Nordhoff. Linked data for linguistic diversity research: Glottolog/langdoc and asjp online. In Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann, editors, *Linked Data in Linguistics*, pages 191–200. Springer Berlin Heidelberg, 2012.

[16] Mark van Assem, Aldo Gangemi, and Guus Schreiber. RDF/OWL Representation of WordNet. W3C Working Draft, World Wide Web Consortium, June 2006. http://www.w3.org/TR/wordnet-rdf/.

[17] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proc. HLT 2005*, HLT '05, pages 347–354, 2005.