Semantic Web 0 (2015) 1
IOS Press

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

Editor(s): Sebastian Hellmann, AKSW, University of Leipzig, Germany; Steven Moran, LMU Munich, Germany; Martin Brümmer, AKSW, University of Leipzig, Germany; John P. McCrae, CITEC, Bielefeld University, Germany
Solicited review(s): Jose Emilio Labra Gayo, University of Oviedo, Spain; John P. McCrae, CITEC, Bielefeld University, Germany; Menzo Windhouwer, Max Planck Institute for Psycholinguistics, The Netherlands

Gerard de Melo *
IIIS, FIT 1-208, Tsinghua University, Beijing 10084, China

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 published 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 information about numerous aspects of human language online in both human-readable and machine-readable form, contributing to the Web of Linked Data and the Semantic Web. Language is the basis of communication and the key to the tremendous body of written knowledge produced by humanity. Due to the ubiquity of textual data, the value of lexical and other linguistic information is increasingly being recognized in several communities, including the Database, Semantic Web, and Digital Library communities. At the same time, the value of making linguistic data interoperable has been receiving an increased amount of attention in linguistics and lexicography. Among others, the Open Linguistics Working Group [3] of the Open Knowledge Foundation has brought researchers working in areas 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 focusing on linguistic data. This article describes Lexvo.org[1] and its contribution to this emerging cloud of Linguistic Linked Data.

2. Language Information

One main focus of Lexvo.org is to provide descriptions of human languages. The term *languages*, here, is meant to be inclusive, encompassing specific language variants (such as dialects), and larger groups of language variants (e.g. macrolanguages and language families), given that the distinctions between such categories are largely conventional.

2.1. Language Identification

In numerous application settings, there is a need to reference a given human language. For example, one may wish to express that a book is written in a specific language or that a person prefers that the

*Gerard de Melo is supported in part by National Basic Research Program of China Grants 2011CBA00300, 2011CBA00301, and NSFC Grants 61033001, 61361136003.

[1] In particular, the 2015-03-01 version of the dataset.

1570-0844/15/$27.50 © 2015 – IOS Press and the authors. All rights reserved

user interface be set to a particular language. One of the central motivations for the Web of Linked Data is the idea of liberating data from traditional data silos by relying on shared global identifiers rather than database-dependent strings of characters. Thus, instead of establishing a `language` column in their relational database, with values such as `engl.`, `grk.`, `albn.` (or `en`, `el`, `sq`), data publishers and application developers are able to publish data on the Web using global identifiers (URIs) such as `http://lexvo.org/id/iso639-3/eng` and `http://lexvo.org/id/iso639-3/ell`. URIs of this sort are part of a shared global vocabulary common to a multitude of different datasets on the Web. This allows us to recognize that two databases are in fact referring to the same language, as opposed to when, for example, one uses a marker such as `grk.`, while the other one uses `el` to refer to the Greek language. Additionally, these URIs are also dereferenceable, meaning that humans may use them to open a corresponding page in their browser and software tools can download machine-readable data to obtain further details about what the URI identifies.

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 describes just a few hundred languages.

To address this situation, Lexvo.org, since 2008, has defined 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 [23], there is no good Linked Data alternative to Lexvo.org for ISO 639-3. Lexvo.org's language identifiers are used by the British Library[2], the Spanish National Library (`datos.bne.es`), and the French academic catalog Sudoc, among many 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 [25] serves significantly more fine-grained identifiers for language definitions proposed by individual linguists.

2.2. Language Descriptions

Lexvo.org delivers extensive descriptions of each language, extracted from sources such as Wikipedia and the Unicode CLDR. In the Lexvo.org data, these are often expressed using properties and classes from the Lexvo Ontology, a custom ontology focusing on general semantic classes and properties[3]. Examples include language names in many languages, relevant geographical regions, 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. Geographical regions are identified using URIs based on ISO 3166 / UN M.49, which have also been connected to the GeoNames dataset.

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 around 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 and language collections 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. the set of sign languages.

Lexvo.org draws information about language collections, including relationships between them, from Wikipedia, WordNet [12], and the ISO standard. This leads to an extensive language hierarchy that is fully integrated into a general-purpose WordNet-based word sense hierarchy (see Section 3.2). From Mandarin Chinese, for instance, one can easily navigate to general

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

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

Chinese, the Sinitic languages, and the Sino-Tibetan languages. While there is still considerable research and debate on certain putative language family relationships, Lexvo.org relies on a methodology that favours well-established relationships and to some extent eschews any forced distinctions between dialects, languages, and language families. Instead, it draws on official standards, WordNet, and Wikipedia to form a hierarchy of language systems of varying granularity, as described elsewhere in more detail [10].

2.4. Scripts and Characters

On Lexvo.org, languages are connected to the writing systems commonly used for them, such as for example Cyrillic, Indian Devanagari, or the Korean Hangul system. The identifiers for these writing systems are based on the ISO 15924 standard. By extracting Unicode Property Values from the Unicode specification, Lexvo.org also connects writing systems with the specific characters that are part of them.

URIs of the form `http://www.lexvo.org/id/char/5A34` are provided for each of the several thousand abstract characters defined by the international Unicode standard. The final part here, `5A34` in the example, refers to the hexadecimal notation for the respective Unicode code point of the character. Following the Unicode definitions, such characters are abstract units of information that do not have a particular concrete form and hence should not be confused with glyphs. 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, especially regarding their composition and variants.

2.5. Phones

Lexvo.org was recently extended to include phonetic information. For a given phone, it provides different representations (IPA, X-SAMPA, Arpabet, etc.) and properties (e.g. labiodental, plosive).[4]

[4]The phoible.org project provides a more extensive description of the phonetic properties of different languages.

3. Lexical Information

3.1. Identifiers for Terms

The second major focus of Lexvo.org is to identify and describe terms (or words) and their properties. In the RDF standard that much of the Semantic Web and Linked Data cloud is based on, string literals cannot serve as subjects of a statement. The same applies to string literals with language tags, and thus it is non-trivial to describe terms in RDF. Some ontologies 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 representation of WordNet [27] defined URIs for the words covered by the WordNet lexical database [12].

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 any term in a given language. Data publishers and developers can obtain and work with such URIs by using a simple Java API (see Section 5) or by following the specifications described below.

3.1.1. Formal Semantics

Formally, different levels of abstraction could be chosen to refer to terms. Lexvo.org's term URIs are intended to serve as interchange URIs that can easily be created from any word-segmented natural language text. When a term is encountered in a text document, a computer system initially only sees the surface form. Typically, one wishes to look up terms in a given knowledge source (e.g. in a thesaurus or a dictionary) without already knowing what meanings they have.

Lexvo.org's notion of term is thus based exclusively on the surface form, at an abstraction level that does not reflect any semantic or historic differences between words sharing the same form. This notion of terms is hence at a higher level of abstraction than the standard linguistic notion of words, which typically considers the animal noun "*bear*" different from the verb "*bear*", and typically distinguishes "*bass*" as in music from "*bass*" as a fish. Lexvo.org's notion of terms explicitly avoids distinguishing the meanings of polysemous or homonymic words within a specific language, because in many settings all one has is the surface form without any knowledge of which specific homonymic variant one is dealing with.

Instead, Lexvo.org makes such distinctions only at the word sense level, described later on in Section 3.2.

Lexvo.org thereby avoids the often rather subjective decisions about whether two word senses have a distinct enough history to count as separate words or not. However, data publishers are free to use Lexvo.org in conjunction with word entities that group together different word senses.

In contrast, Lexvo.org does, nonetheless, 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 terms in different languages using simple RDF triples. If one instead used URIs based on pure string literals without language information, some rather cumbersome form of reification would be required to properly reference the languages of those terms.

Different word forms are treated as distinct terms. Here, however, there are a few minor subtleties of term identity regarding their string encodings. 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 [4] is applied to avoid duplicate entities. Formally, given a term t in a language L, the URI is constructed as follows:

1. The term t is encoded using Unicode, and the NFC normalization procedure [4] 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. Normalization ensures that there is only one unique form.
2. The resulting Unicode code point string is encoded in UTF-8 to obtain a sequence of bytes.
3. This byte sequence is converted to an ASCII path segment by applying percent-encoding as per the RFC 3986 standard. Essentially, this means that certain characters need to be escaped: Such 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.
4. 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. For Mandarin Chinese, for instance, the string `http://lexvo.org/id/term/cmn/` is prepended to the RFC 3986-encoded UTF-8 byte sequence of the original term string.

Fortunately, Lexvo.org's Java API (Section 5) 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`.

In order to make the meaning of URIs more explicit, Lexvo.org proposes linking URIs to term URIs of one or more natural languages using a lexicalization property, whenever appropriate. Such a property 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. For instance, a URI of the form `http://.../PasspNo`, which in principle could be arbitrary, can be connected to an English lexicalization such as "*passport number*". The semantics of this URI thereby becomes better grounded due to the newly established correspondence with the meaning of the term "*passport number*" in the English language.

Unlike SKOS [23], which aims at expressing descriptions of individual vocabularies and their properties, Lexvo.org describes generic properties of words and other expressions in a language as a whole. For instance, Lexvo.org's properties do not make any normative claims about which labels are to be preferred, but merely describes the fact that a certain term can be used in some specific sense. The focus on describing language per se also means that the Lexvo Ontology does not try to describe specific vocabularies, lexicons, or lexical entries in them. Given that Lexvo.org predates the lemon model [21] and has different goals, it relies on a somewhat different conceptual framework. Fortunately, the two models can be made interoperable to a considerable extent. While both SKOS and lemon focus on providing a rich vocabulary that can be used to express terminological or lexical knowledge,

Lexvo.org's primary focus, in any case, has always been to actually provide certain kinds of language-related knowledge as well as reusable identifiers for the involved entities.

Much of the multilingual information about words that Lexvo.org supplies is procured from Wiktionary, a well-known effort to collaboratively create dictionaries on the Web. Lexvo.org currently provides links from terms to human-readable web pages in the English, Catalan, French, German, Greek, Portuguese, Spanish, and Swedish Wiktionary versions. Lexvo.org also extracts translations as well as part-of-speech tag information, explaining whether a word functions as a noun or an adjective, for instance. Lexvo.org also publishes etymological relationships between words as triples (see Section 4.2). Lexvo.org was the first web site to publish Linked Data based on Wiktionary on the Web back in 2008. More recently, other sites have emerged that aim at establishing a more direct mapping of Wiktionary's data to RDF [16].

Lexvo.org's term entities are also linked to corresponding 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 provided 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 [13]. In sources like Wiktionary, sense definitions can be quite ephemeral and vary over time as site visitors 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 appropriate senses, as automatic disambiguation is known to be rather 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 datasets often do not align in a clean way [7]. Some even go as far as arguing that word senses are not necessarily a useful notion at all [17]. 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 [12]. 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, especially in UWN, the Universal WordNet project [9]. WordNet's identifiers have been linked to YAGO, SUMO, OpenCyc, VerbNet, and numerous other datasets (some of which are discussed later on). Lexvo.org was the first service to put WordNet 3.0 online as Linked Data. It links English terms to their respective WordNet synsets and provides links between synsets.

4. Towards a Linguistic Linked Data Cloud

Lexvo.org first went online in 2008, serving information from lexical resources like Wiktionary as well as language information data. 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. Due to these developments, a number of data publishers have recognized the value of Lexvo.org's language descriptions.

Recently, many third parties have created linguistic datasets, leading to the emergence of a cloud of Linguistic Linked Data [3]. In order to strengthen and accelerate these efforts, Lexvo.org has begun publishing several separate new datasets on its download page[5]. All of these are linked to Lexvo.org, with predicates 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 the American 1911 edition of Roget's Thesaurus [20,2]. The textual data is parsed using a rule-based top-down approach [8], which also needs to handle various formatting problems (including errors) in

[5] `http://www.lexvo.org/linkeddata/resources.html`

the original data. The RDF conversion includes both the hierarchy of topics as well as the terms associated with individual headword-level topics.

The WordNet Evocation dataset [1] provides data about associations between word senses, e.g. between senses of "*car*" and "*road*". We obtain WordNet URIs for the respective senses, and, following the original paper, determine the median score of each evocation relationship. In order to express these scores, RDF Reification is used.

WordNet Domains delivers thematic domain markers such as `astronomy` or `engineering` 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 [15]. WordNet synsets are linked to topics and topics form a shallow hierarchy.

4.2. Cross-Linguistic Data

Etymological WordNet [6] contributes links between words that are not semantic but etymological or derivational in nature. For instance, the English word "*salary*" ultimately goes back to the Latin word "*sal*" (salt). The RDF conversion straightforwardly maps the original relationships between words extracted from Wiktionary to triples that use Lexvo.org term URIs and predicates from the Lexvo Ontology.

4.3. Semantic Frames and Roles

Lexvo.org maintains RDF datasets for FrameNet [14], the PropBank lexicon [26] (not the corpus), and NomBank [22], all resources that model the phrase- and sentence-level semantic frames and roles that words express. The RDF conversions are currently quite incomplete, as they only contain a subset of information that can easily be expressed using the Lexvo Ontology and other well-known RDF predicates. However, they will be extended over time, as lemon and other models finalize their representations of such data [21].

Lexvo.org also publishes an RDF version of VerbNet [18], a resource providing a Levin-style [19] classification of verbs based on their syntactic properties. The RDF conversion includes word senses, class memberships, and class hierarchy relationships, as well as mappings to WordNet.

4.4. Sentiment Analysis Data

Lexvo.org hosts a Linked Data version of the MPQA Subjectivity Lexicon [28], which supplies subjectivity and sentiment polarity labels.

Additionally, the AFINN dataset has been converted [24], offering more fine-grained numeric sentiment valency scores. The Lexvo Ontology assumes scores in the range $[0, 1]$, so all values are converted or normalized to this range.

4.5. Speech-Related Data

An RDF version of the CMU Pronunciation Dictionary[6] has been produced. We convert the original ArpaBet encoding to IPA, which is what the Lexvo Ontology relies on, and guess the most likely capitalization of words using dictionary lookups.

5. Architecture and Service

5.1. Processing Framework

Lexvo.org relies on an automatic data processing architecture that allows for simple updates whenever any of the original data sources are updated. Lexvo.org is updated regularly to ensure that its data is always fairly up-to-date. For deprecated ISO 639-3 language codes, the system points to relevant alternative language identifiers.

The information that Lexvo.org serves is processed in a fairly sophisticated workflow system. In a first step, data is extracted from numerous sources, including official registries, Wikipedia, and Wiktionary.[7] Then, manually and automatically created identity links between the entity identifiers from different data sources are processed using a data integration algorithm relying on combinatorial optimization techniques [11,5]. This algorithm creates equivalence classes of the original entity identifiers and thus allows us to aggregate information from different sources. Additional algorithms are invoked to create the language hierarchy (Section 2.3) and to transform and filter RDF triples in several different ways. Other external resources are read in order to create links to them.

[6] `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`

[7] Full list at `http://www.lexvo.org/linkeddata/references.html`

For instance, Lexvo.org provides links to several thesauri but first needs to rely on dictionary heuristics to guess the true capitalization of the words, since thesauri often capitalize all headwords. Finally, the data is exported to an RDF dump as well as to custom binary databases used by the online server.

5.2. Publishing

Lexvo.org is backed by a custom Linked Data server infrastructure that makes its URIs dereferenceable online and part of the Linked Data cloud. This infrastructure also provides machine-readable dumps and mapping tables (see `http://www.lexvo.org/linkeddata/resources.html` for more information). The RDF dump is available under a Creative Commons CC-BY-SA license.[8] Certain kinds of statements are generated on-the-fly by the server, e.g. RDF descriptions for arbitrary term URIs. Additionally, the server also enhances the human-readable versions of the dumps, for instance by serving links to online maps in the case of geographic locations.

5.3. Use Cases

Data publishers typically use Lexvo.org as a provider of URIs for language-related entities. The aforementioned Java API provides an easy way of constructing Lexvo.org-based URIs. One can provide an ISO language code as input and obtain the corresponding Lexvo.org language URI. One can also easily construct term URIs simply by supplying a string and a language. The API automatically carries out the steps detailed earlier in Section 3. Further information about this API is available online.[9]

Data publishers may also want to consult the Lexvo Ontology[10], which provides numerous language-related properties and classes, including properties for notions of identity and near-identity aimed at mitigating long-standing problems in the Linked Data world [5].

Finally, data consumers use Lexvo.org to retrieve the various kinds of information elaborated earlier in Sections 2, 3, and 4. For example, they can retrieve senses and translations of a term or the geographical locations associated with a language.

[8] Details at `http://www.lexvo.org/legal.html`

[9] `http://www.lexvo.org/linkeddata/tutorial.html`

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

6. Conclusion

In summary, Lexvo.org is a valuable service that defines standard identifiers for languages and language families, words and word senses, scripts, characters, and other language-related entities. These are being used by a multitude of data publishers in several different communities. Additionally, Lexvo.org publishes a broad spectrum of language-related information about these entities that is extensively used by numerous third-party data consumers. Lexvo.org provides online access as well as machine-readable dumps to ensure widespread availability of this information. This ecosystem of data constitutes a useful basis for applications in linguistics, natural language processing, and other areas that benefit from the considerably interlinked and interoperable nature of the resources. We believe that this provides tremendous incentives for third parties to contribute to the growing Linguistic Linked Data cloud.

References

[1] Jordan Boyd-Graber, Christiane Fellbaum, Daniel Osherson, and Robert Schapire. Adding dense, weighted, connections to WordNet. In Petr Sojka, Key-Sun Choi, Christine Fellbaum, and Piek Vossen, editors, *Proceedings of the 3rd International WordNet Conference (GWC)*, pages 29–35, Brno, 2006. Masaryk University.

[2] Patrick Cassidy. An investigation of the semantic relations in the roget's thesaurus: Preliminary results. In Alexander Gelbukh, editor, *Proceedings of CICLing-2000, Conference on Intelligent Text Processing and Computational Linguistics*, pages 181–204, Mexico, 2000. IPN Publishing House.

[3] 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 Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the 8th Language Resources and Evaluation Conference (LREC 2012)*, pages 3603–3610. European Language Resources Association (ELRA), 2012.

[4] Mark Davis, Ken Whistler, and Martin Dürst. Unicode normalization forms. Unicode Standard Annex 15. Technical report, Unicode Consortium, 2009.

[5] Gerard de Melo. Not quite the same: Identity constraints for the Web of Linked Data. In Marie desJardins and Michael L. Littman, editors, *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, pages 1092–1098, Menlo Park, CA, USA, 2013. AAAI Press.

[6] Gerard de Melo. Etymological Wordnet: Tracing the history of words. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani,

Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014)*, pages 1148–1154, Paris, France, 2014. European Language Resources Association (ELRA).

[7] Gerard de Melo, Collin F. Baker, Nancy Ide, Rebecca Passonneau, and Christiane Fellbaum. Empirical comparisons of MASC word sense annotations. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the 8th Language Resources and Evaluation Conference (LREC 2012)*, pages 3036–3043, Paris, France, 2012. European Language Resources Association (ELRA).

[8] Gerard de Melo and Gerhard Weikum. Mapping Roget's Thesaurus and WordNet to French. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008)*, pages 3306–3313, Paris, France, May 2008. European Language Resources Association (ELRA).

[9] Gerard de Melo and Gerhard Weikum. Towards a universal wordnet by learning from combined evidence. In David Wai-Lok Cheung, Il-Yeol Song, Wesley W. Chu, Xiaohua Hu, and Jimmy J. Lin, editors, *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, pages 513–522, New York, NY, USA, 2009. ACM.

[10] Gerard de Melo and Gerhard Weikum. Towards universal multilingual knowledge bases. In Pushpak Bhattacharyya, Christiane Fellbaum, and Piek Vossen, editors, *Principles, Construction, and Applications of Multilingual Wordnets. Proceedings of the 5th Global WordNet Conference (GWC 2010)*, pages 149–156, New Delhi, India, 2010. Narosa Publishing.

[11] Gerard de Melo and Gerhard Weikum. Untangling the cross-lingual link structure of Wikipedia. In Jan Hajič, Sandra Carberry, Stephen Clark, and Joakim Nivre, editors, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 844–853, Stroudsburg, PA, USA, July 2010. Association for Computational Linguistics.

[12] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA, USA, 1998.

[13] Charles J. Fillmore and Beryl T. Atkins. Describing polysemy: The case of 'crawl'. In Yael Ravin and Claudia Leacock, editors, *Polysemy: Theoretical and Computational Approaches*, pages 91–110. Oxford University Press, 2000.

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

[15] Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. Multilingual Central Repository version 3.0. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the 8th Language Resources and Evaluation Conference (LREC 2012)*, pages 2525–2529, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA).

[16] Sebastian Hellmann, Jonas Brekle, and Sören Auer. Leveraging the crowdsourcing of lexical resources for bootstrapping a linguistic data cloud. In Hideaki Takeda, Yuzhong Qu, Riichiro Mizoguchi, and Yoshinobu Kitamura, editors, *Semantic Technology*, volume 7774 of *Lecture Notes in Computer Science*, pages 191–206. Springer, Berlin/Heidelberg, 2013.

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

[18] Karen Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. Extending VerbNet with novel verb classes. In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk, and Daniel Tapias, editors, *Proceedings of the 5th Language Resources and Evaluation Conference (LREC 2006)*, pages 1027–1032, 2006.

[19] Beth Levin. *English verb classes and alternations: A preliminary investigation*. University of Chicago Press, Chicago, 1993.

[20] C.O. Sylvester Mawson, editor. *Roget's Thesaurus of English Words and Phrases Classified and Arranged so as to Facilitate the Expression of Ideas and Assist in Literary Composition*. McDevitt-Wilson's, Inc., New York, NY, USA, 1911.

[21] John McCrae, Guadalupe Aguado-de Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr, and Tobias Wunner. Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 46(4):701–719, 2012.

[22] Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. The NomBank Project: An interim report. In Adam Meyers, editor, *Proceedings of the HLT-NAACL 2004 Workshop on Frontiers in Corpus Annotation*, pages 24–31, Boston, MA, USA, 2004. Association for Computational Linguistics.

[23] Alistair Miles and José R. Pérez-Agüera. SKOS: Simple knowledge organisation for the web. *Cataloging & Classification Quarterly*, 43(3):69–83, 2007.

[24] Finn Årup Nielsen. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In Matthew Rowe, Milan Stankovic, Aba-Sah Dadzie, and Mariann Hardey, editors, *Proceedings of the ESWC 2011 Workshop on Making Sense of Microposts: Big things come in small packages*, volume 718 of *CEUR Workshop Proceedings*, pages 93–98, May 2011.

[25] 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.

[26] Martha Palmer, Daniel Gildea, and Paul Kingsbury. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106, 2005.

[27] Mark van Assem, Aldo Gangemi, and Guus Schreiber. RDF/OWL representation of WordNet. Technical report, W3C, June 2006. http://www.w3.org/TR/wordnet-rdf/.

[28] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399–433, September 2009.