Representing Multilingual Data as Linked Data: the Case of BabelNet 2.0

Maud Ehrmann[1], Francesco Cecconi[1], Daniele Vannella[1], John M^{c}Crae[2], Philipp Cimiano[2], Roberto Navigli[1]

[1] Department of Computer Science, Sapienza University of Rome, Italy
[2] Semantic Computing Group, CITEC, University of Bielefeld, Germany
{ehrmann|vannella|navigli@di.uniroma1.it}, francesco.cecconi@gmail.com, {jmccrae|cimiano}@cit-ec.uni-bielefeld.de

Abstract

Recent years have witnessed a surge in the amount of semantic information published on the Web. Indeed, the Web of Data, a subset of the Semantic Web, has been increasing steadily in both volume and variety, transforming the Web into a 'global database' in which resources are linked across sites. Linguistic fields – in a broad sense – have not been left behind, and we observe a similar trend with the growth of linguistic data collections on the so-called 'Linguistic Linked Open Data (LLOD) cloud'. While both Semantic Web and Natural Language Processing communities can obviously take advantage of this growing and distributed linguistic knowledge base, they are today faced with a new challenge, i.e., that of facilitating multilingual access to the Web of data. In this paper we present the publication of BabelNet 2.0, a wide-coverage multilingual encyclopedic dictionary and ontology, as Linked Data. The conversion made use of *lemon*, a lexicon model for ontologies particularly well-suited for this enterprise. The result is an interlinked multilingual (lexical) resource which can not only be accessed on the LOD, but also be used to enrich existing datasets with linguistic information, or to support the process of mapping datasets across languages.

Keywords: Linguistic Linked Data, Multilingual Semantic Web, lexical-semantic resource, semantic network

1. Introduction

On the strength of the widespread adoption of the Linked Data paradigm (Bizer et al., 2009), along with the maturation of techniques and methodologies for the publication of Linked Data (Auer et al., 2011; Heath and Bizer, 2011), recent years have witnessed a surge in the amount of semantic information published on the Web. As evidenced by the Linked Open Data (LOD) cloud[1], the Web of Data, a subset of the Semantic Web, has been increasing steadily in both volume and variety, transforming the Web into a 'global database' in which resources are linked across sites.

A grassroots effort by members of the Natural Language Processing (NLP) and Semantic Web communities, in particular the Open Linguistics subgroup[2] of the Open Knowledge Foundation[3], has initiated the development of a LOD sub-cloud: the *Linguistic* Linked Open Data (LLOD) cloud. Indeed, stimulated by initiatives such as the W3C Ontology-Lexica community group[4], the publication of linguistic data collections on the Web is progressing encouragingly. As defined by Chiarcos et al. (2013), the challenge is to "store, to connect and to exploit the wealth of language data", with the key issues of (linguistic) resource interoperability, i.e., the ability to syntactically process and semantically interpret resources in a seamless way (Ide and Pustejovsky, 2010), and information integration, i.e., the ability to combine information across resources. The adoption of linked data principles along with the use of the Resource Description framework (RDF), a generic model for data representation, is proving to be an effective and successful approach towards achieving this goal. All types of linguistic resource are eligible for the LLOD cloud, ranging across lexical-semantic resources (such as machine-readable dictionaries, semantic knowledge bases, ontologies) to annotated linguistic corpora, and including repositories of linguistic terminologies and meta-data repositories (Chiarcos et al., 2011).

The benefits of such a 'Web of Linguistic Data' are diverse and lie on both Semantic Web and NLP sides. On the one hand, ontologies and linked data sets can be augmented with rich linguistic information, thereby enhancing Web-based information processing. On the other hand, NLP algorithms can take advantage of the availability of a vast, interoperable and federated set of linguistic resources, as well as benefit from a rich ecosystem of formalisms and technologies. In the medium term, a Web-based integration of NLP tools and applications is foreseeable; a few steps have already been taken in this direction with the recent definition of the NLP Interchange Format (NIF) (Hellmann et al., 2013). De facto, common initiatives between SW and NLP are multiplying[5].

The publication of a large amount of data on the Web, coupled with the emerging LLOD cloud, thus represents a great opportunity. Nevertheless, if data sharing is to become common practice, a new challenge has to be faced: overcoming language barriers. Indeed, if the Semantic Web can be assumed to be inherently language-independent (Gracia et al., 2011), the question arises of how to enable the interaction between users, who are operating

[1] http://lod-cloud.net/state/
[2] http://linguistics.okfn.org/2011/05/20/the-open-linguistics-working-group/
[3] http://okfn.org/
[4] http://www.w3.org/community/ontolex/
[5] See for example the Multilingual Web Linked Open Data and DBpedia&NLP workshops (http://www.multilingualweb.eu/en/documents/dublin-workshop and http://iswc2013.semanticweb.org/content/dbpedia-nlp-2013) respectively).

in their own natural languages, and language-independent data representations. So far, there has been a clear bias in the vocabularies used in the Web of Data towards the English language, however, we can observe a growing trend towards the publication of non-English data sources. Thus, the problem remains of how to connect data expressed in different languages. Indeed, despite having recognized the necessity of overcoming 'data silos'[6], we are currently facing the danger of creating confined 'monolingual islands' of data that do not interoperate. Multilinguality is thus a crucial concern for the Semantic Web, and addressing it could act as a significant lever towards attaining full access to knowledge and data.

This paper presents a contribution for the Multilingual Web of Data, with the publication of BabelNet 2.0 as linked data. BabelNet (Navigli and Ponzetto, 2012) is a very large multilingual encyclopedic dictionary and ontology whose version 2.0 covers 50 languages. Based on the integration of lexicographic and encyclopedic knowledge, BabelNet 2.0 offers a large network of concepts and named entities along with an extensive multilingual lexical coverage. Its conversion to linked data was carried out using the *lemon* model (**Le**xicon **M**odel for **On**tology) (McCrae et al., 2012b), a lexicon model for representing and sharing ontology lexica on the Semantic Web. It is our hope that the publication as linked data of such a semantic network lexicalized in an ample set of languages will support the Semantic Web in its effort to scale to further languages, as well as enhance knowledge and linked data-based NLP applications.

The remainder of the paper is organized as follows. In Section 2 we specify the needs of the Semantic Web in terms of multilingual and cross-lingual information access and integration, and expose how this need can be partially addressed by the emergence of a multilingual Web of linguistic data. Next, after the introduction of the BabelNet resource (Section 3), we detail its conversion to linked data and present its interconnections with other datasets on the Web (Section 4). Section 5 provides an account of statistics and aspects related to publication; finally, after the discussion of related work in Section 6, we conclude in Section 7.

2. Multilingual Linguistic Linked Data

The Semantic Web and NLP stand in a symbiotic relationship, in that the Semantic Web is in need of NLP and content analytic solutions in order to embrace the Web of unstructured documents, while NLP can benefit from the plethora of semantic resources and knowledge bases available on the Semantic Web, e.g., as linked data. In this section, we explain and outline the different aspects of this symbiosis and discuss some of the issues involved therein.

In contrast to the traditional Web, where information can be found in different languages if, and only if, corresponding translations exist on Web sites, the Semantic Web can be considered as inherently language-independent (Gracia et al., 2011). By virtue of the first linked data principle, each 'thing' (or element of a resource) is indeed assigned a unique identifier (URI) through which it can be uniquely and unambiguously recognized. However, while Semantic Web data models are to a large extent language-agnostic, the primary means of human communication remains natural language, which poses several problems. One of these problems concerns information access and raises the question of the interaction between users, whose information needs are expressed in natural language, and language-agnostic data representations. A key point in this respect is, for example, the automatic generation of SPARQL queries from questions expressed in different natural languages, and, vice versa, the production of human understandable descriptions of RDF resources. Another problem pertains to data description, with the question of the different types of information associated to data sources, which are, to a large extent, language-dependent. Therefore, information can be found in different languages on the Semantic Web if, and only if, mediation mechanisms between users and data are in place, and if the data itself is lexicalized and interlinked on a multilingual basis. Cross-language access to information depends on addressing the above mentioned obstacles which, if fully dealt with, could lead to a "level playing field for [semantic Web] users with different cultural backgrounds, native languages, and originating from different geo-political environments" (Buitelaar et al., 2012).

Considering the challenges lying ahead for a truly multilingual Web, Gracia et al. (2011) envision the multilingual Web of data as the current LOD, enriched with an additional layer of resources and services centred on multilingualism. This emphasizes the following points: i) linguistic information for different natural languages used in the content description of linked data, ii) mapping across linked datasets, regardless of the language in which they are expressed and iii) services to dynamically exploit the multilingual linked data cloud.

With regard to the aspect of linguistic information attached to linked datasets (i.e., how resources are described), a recent study by Gómez-Pérez et al. (2013) provides useful insights on the current state of the LOD, giving particular attention to the multilingual dimension[7]. The most striking point of the study concerns the distribution of natural languages, with a large proportion (approx. 80%) consisting of monolingual RDF datasets. The authors observe, however, that the number of multilingual datasets has as much as doubled during the period under consideration. This suggests, first, that the dominance of English, if confirmed, might be receding and, second, that the need for cross-lingual mappings between linked datasets will become even more critical. Another analysis of this study concludes that the usage of language tags (on literals) is generally very low, with around 21% of literals having a language tag, confirming once again the predominance of English (85% of the tagged literals). This study considers the LOD, where the published information is primarily of a factual nature, and demonstrates its lack of multilingual linguistic information. Consequently, we now turn our attention on its lin-

[6] http://blog.ted.com/2009/03/13/tim_berners_lee_web/

[7] The study was based on three snapshots of DyLDO corpora (containing about 42 million literals) over the year 2012.

guistic counterpart, namely, the LLOD cloud.

Linguistic linked-data resources convey information about language(s) and are of the greatest importance for the Semantic Web in the sense that they can, apart from answering the above mentioned needs, enable linked data-based NLP applications and content analytics. Linguistic resources in the linked data format can be used to extend the linguistic information of linked datasets by providing lexicalizations in several languages and/or richer linguistic descriptions, and to support cross-lingual mapping between datasets. Furthermore, beyond the needs of the Semantic Web in terms of multilingual and linguistic linked data, it is worth mentioning that NLP, which has been struggling for several years with the definition of standards for linguistic resources as well as accessibility issues, can benefit greatly from the LLOD. Chiarcos et al. (2013), while advocating the publication of language resources as linked data, identify several advantages which can be summarized as follows: resource interoperability, both at a structural (same format) and conceptual (same vocabulary) level, resource integration (*via* interlinking) and resource maintenance (*via* a rich ecosystem of technologies allowing, among other things, a continuous updating). The recent advent of the linked data paradigm could thus, at the level of natural language processing, help overcoming several 'bottlenecks' related to (linguistic) knowledge.

In its September 2013 edition, the Linguistic Linked Open Data cloud diagram (Chiarcos et al., 2012) published by the Open Linguistics Working Group and displayed here in Figure 1 provides language data, such as corpora and lexical-semantic resources, and meta-data, such as linguistic typologies and terminologies. With respect to lexical-semantic resources we should note that, with the exception of resources built from multilingual wikipedias such as DBpedia or YAGO, English language prevails once again; this is particularly true when dealing with pure lexical knowledge (WordNet, FrameNet, VerbNet), as opposed to encyclopedic knowledge.

In light of all this, it appears that potential benefits for the Semantic Web and for NLP with regard to linguistic linked data are highly intertwined and that there exists a common need, i.e., to handle content in multiple languages. Recently, various efforts have been made in order to enable the multilingual capacities of both LOD and LLOD, with the definition of guidelines and best practices (Gómez-Pérez et al., 2013) and the design of principled models (McCrae et al., 2012b; Montiel-Ponsoda et al., 2011). Our work builds on such efforts and aims at filling the multilingual gap of the (linguistic) linked open data cloud.

3. BabelNet 2.0

BabelNet[8] is a lexico-semantic resource whose aim is to provide wide-coverage encyclopedic and lexicographic knowledge in many languages. More precisely, BabelNet is both a multilingual encyclopedic dictionary, with lexicographic and encyclopedic coverage of terms, and an ontology which connects concepts and named entities in a very large network of semantic relations, made up of more than 9

[8] http://www.babelnet.org

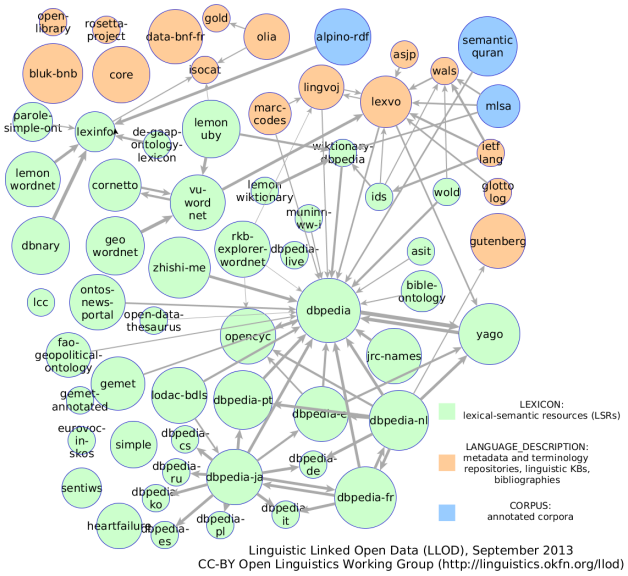

Figure 1: The Linguistic Linked Open Data cloud diagram (draft version of September 2013) by the Open Linguistics Working Group.

million entries, called *Babel synsets*. Adopting a structure similar to that of WordNet (Fellbaum, 1998), each Babel synset represents a given meaning and contains all the synonyms, called *Babel senses*, which, in different languages, express that meaning. The resource provides, for example, lexical knowledge about the concept *apple* as a fruit, with its part of speech, its definitions and its set of synonyms in multiple languages, as well as encyclopedic knowledge about, among other entities, the *Apple Inc.* company, anew along with definitions in multiple languages. Thanks to the semantic relations it is furthermore possible to learn that *apple* is an *edible fruit* (or a *fruit comestible*, a *frutta*, an *essbare Früchte*) and that *Apple Inc.* is related to *server* and *Mountain View California*. While 6 languages were covered in the prime version 1.0, BabelNet 2.0 makes giant strides in this respect and covers 50 languages. This new version is obtained from the automatic integration of:

- *WordNet*, a popular computational lexicon of English (version 3.0),
- *Open Multilingual WordNet* (OMWN), a collection of wordnets available in different languages,
- *Wikipedia*, the largest collaborative multilingual Web encyclopedia, and
- *OmegaWiki*, a large collaborative multilingual dictionary.

As its starting point, BabelNet's core is grounded on the mapping of Wikipedia to WordNet. As described in (Navigli and Ponzetto, 2012), the acquisition methodology consists of three main steps. The first step consists in mapping resources. The second step involves harvesting multilingual lexicalizations. The third step consists in the establishment of semantic relations. In this work, the integration of the resources is based on the estimation of mapping

probabilities between English Wikipedia pages and WordNet synsets. To this end, two scoring functions were experimented, one based on a simple bag-of-words sense representation, and the other on more advanced structural (i.e., graph-based) representations (Navigli and Ponzetto, 2012); subsequently, the correctness of the WordNet-Wikipedia mapping in BabelNet 1.1.1 was estimated around 91% on open-text words (Navigli et al., 2013).

BabelNet 2.0 builds naturally on this core and further expands its lexical knowledge and multilingual coverage with the integration of Open Multilingual WordNet[9] (OMWN) and OmegaWiki[10]. OMWN (Bond and Foster, 2013) is a collection of wordnets in multiple languages with minimal license restrictions. Though simple in essence, the setting up of such a collection called upon its authors to extract (and thus first find) the resources, to normalize their disparate formats and, finally, to link them to a unique Princeton WordNet version. This resulted in a database of wordnets in 26 languages, 16 of which were integrated into BabelNet 2.0 through a simple matching between synsets.

In contrast to these fully-structured wordnets, OmegaWiki belongs to the family of collaboratively built semi-structured resources. Standing on a WordNet-like structure, this multilingual dictionary is organized around concepts called *Defined Meanings* subsuming sets of multilingual *expressions* considered as translations (and synonyms) of each other. Each expression in each language holds a definition, and DefinedMeanings are further characterized by means of semantic relations (*Braeburn* is a hyponym of *apple*) and semantic classes (in our case, *fruit*). The integration of this resource in BabelNet thus amounted to a mapping between DefinedMeanings and Babel synsets. This was achieved, in a nutshell, by combining several bag-of-words similarity measures over English glosses, multilingual glosses (in 5 languages), and multilingual senses (Babel senses and OmegaWiki expressions). The mapping was evaluated against a set of manually associated concepts (140 in total), with a performance of .86 in terms of F-measure.

Bringing together these multilingual resources, as well as including Wikipedia redirections and translation pages (as further detailed in (Navigli and Ponzetto, 2012)), has notably increased BabelNet's coverage. However, in the case of resource-isolated concepts, i.e., concepts existing in only one resource, it is nearly impossible for senses to be covered in all languages, as can also be the case with multi-resource concepts. To overcome such language coverage discontinuities, translations obtained from the application of a state-of-the-art machine translation system[11] over sense-annotated sentences were added to the Babel synsets, as was done in the first version.

BabelNet 2.0 covers 50 languages belonging to diverse language families such as Indo-European, Indo-Iranian, Balto-Slavic, Uralic and Celtic. Overall, the resource contains about 9.3 million concepts. These concepts cover around 50 million senses, are interconnected through more than 260 million lexico-semantic relations, and are described by almost 18 million glosses. Further statistics about coverage per language, composition of Babel synsets and polysemy are available on BabelNet's website[12].

[9]http://www.casta-net.jp/ kuribayashi/multi/

[10]http://www.omegawiki.org/

[11]Google Translate API.

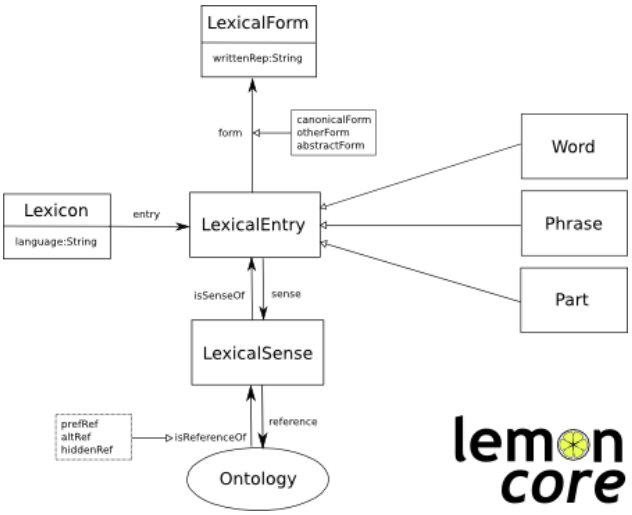

Figure 2: The core of the *lemon* model.

The characteristics of BabelNet, as both a dictionary and an ontology, naturally led to the choice of the *lemon* model for achieving its conversion as linked data.

4. Rendering BabelNet as Linked Data with Lemon

4.1. The *lemon* Model

lemon (McCrae et al., 2011) is a model developed for the representation of lexica regarding ontologies in RDF format. In line with the principle of semantics by reference (Buitelaar, 2010), the model maintains a clean separation between the lexical and semantic layers, enabling lexica to be easily reused to describe different ontologies. As outlined in Figure 2, the core of the lemon model consists of the following elements:

- *Lexical entry*, which comprises all syntactic forms of an entry,
- *Lexical form*, which represents a single inflection of a word, with its *representation*(s), i.e., the actual string(s) used for the word, and
- *Lexical sense*, which represents the usage of a word as a *reference* to a concept in the ontology.

As such the model has already been used for the representation of a number of lexica (Villegas and Bel, 2013; Eckle-Kohler et al., 2014) and proposals have been made to extend the model in new ways (Khan et al., 2013). Specifically designed as an interface between lexical and ontological knowledge and allowing the expression of linguistic information, *lemon* perfectly meets the needs of BabelNet as a candidate for the Linked Data Cloud.

[12]http://babelnet.org/stats.jsp

4.2. BabelNet as Linked Data

BabelNet contains a lot of information; yet, its conversion into RDF mainly involves the consideration of its two core elements, namely Babel senses and Babel synsets. As advocated above, ontological and lexical layers should be kept separated. Therefore, while *lemon* provided us with the means of representing lexical information, i.e., Babel senses, we chose to represent collections of equivalent senses, i.e., Babel synsets, using the class *Concept* of the SKOS (Simple Knowledge Organization System) model[13]. We additionally reused the existing vocabulary of *LexInfo 2* (Buitelaar et al., 2009; McCrae et al., 2012c) to encode some of the semantic relations between Babel synsets. Finally, when no existing vocabulary was answering our needs, we defined our own classes and properties. At the lexical level, Babel sense lemmas are encoded as *lemon* lexical entries. Each lexical entry receives a language tag *rdfs:label*, the indication of its part of speech (*lexinfo:partOfSpeech*) and is further described by means of a lexical form encoding the Babel sense lemma as written representation of the entry. According to their language, these entries are assembled into different *lemon* lexicons (51 in total). In accordance with the principle of semantics by reference modelled by *lemon*, possible meanings of lexical entries are expressed by way of lexical senses pointing to adequate Babel synsets encoded as SKOS concepts. Besides pointing to a referent, lexical senses[14] encode meta-data information with, first, the source of the sense (WordNet, OMWN, Wikipedia or OmegaWiki) and, when relevant, the way it was obtained: *via* automatic translation or thanks to a Wikipedia redirection page (boolean properties). Additionally, these *lemon* senses support the expression of translation variants between Babel senses; indeed, translations pertain to *lemon* sense relations as they should be stated between disambiguated words (i.e., the lexical senses of lexical entries), which do not necessarily refer to the same concept. As an illustration of the encoding of these lexical elements, Figure 3 depicts the *lemon* representation of the Italian Babel sense 'Web semantico' (in Turtle format[15]). Encoded as a *lemon:LexicalEntry* (bn:Web_semantico_n_IT) this entry is part of the Italian *lemon:Lexicon* (bn:lexicon_IT), it holds a *lemon:Form* (bn:Web_semantico_n_IT/canonicalForm), as well as a *lemon:LexicalSense* (bn:Web_semantico_IT/s02276858n).

From the ontological perspective, we used *skos:Concept(s)* to represent our 'units of thought', i.e., Babel synsets. These Babel SKOS concepts encode two types of information: regarding the concept itself, and regarding its semantic relations with other concepts. As a base, Babel SKOS concepts are linked back to the entries of the *lemon* lexica thanks to the property *isReferenceOf*. Next, a BabelNet property (*bn-lemon:synsetType*) indicates whether the Babel synset is a concept or a named entity (NE). Most importantly, multilingual glosses which provide a description of the concept in up to 50 languages, are specified through a *bn-lemon:definition* property referring to a *bn-lemon:BabelGloss*. Although the *skos:definition* would have been the ideal candidate to represent this piece of information, it nevertheless does not enable the expression of additional (meta-data) information about the definition. We therefore defined a class, namely *BabelGloss*, so as to be able to specify the source of the definition (WordNet, OMWN, Wikipedia or OmegaWiki), as well as its license. This is the only BabelNet component for which we could not reuse an element of an existing vocabulary. As regards the semantic relations between Babel synsets, these are encoded as *skos:narrower* and *skow:broader* for hyponyms and hypernyms respectively, as *lexinfo* relations when adequate (member meronym, member holonym, participle, *etc.*), and as *skos:related* when less specified. Finally, Wikipedia categories (in dozens of languages) and their DBpedia twin (in English) are reported for each concept, *via* a dedicated property. Following up with the 'Web semantico' example, Figure 3 shows the concept to which this entry refers, i.e., the *skos:Concept* bn:s02276858n. It holds the above mentioned properties, and links to a *BabelGloss* (here the German one, bn:s02276858n_Gloss1_DE).

```
@prefix bn: <http://babelnet.org/2.0/> .
@prefix bn-lemon: <http://babelnet.org/model/babelnet#>
@prefix lemon: <http://www.lemon-model.net/lemon#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix wikipedia-da: <http://da.wikipedia.org/wiki/Kategori/> .
@prefix wikipedia-it: <http://it.wikipedia.org/wiki/Categorie/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
 ...

bn:lexicon_IT
      a   lemon:Lexicon;
      dc:source   <http://babelnet.org/>;
      lemon:entry   bn:Web_semantico_n_IT, ... ;
      lemon:language   "IT".

bn:Web_semantico_n_IT
      a   lemon:LexicalEntry;
      rdfs:label   "Web_semantico"@IT;
      lemon:canonicalForm   bn:Web_semantico_n_IT/canonicalForm;
      lemon:language   "IT";
      lemon:sense   bn:Web_semantico_IT/s02276858n ;
      lexinfo:partOfSpeech   lexinfo:noun.

bn:Web_semantico_n_IT/canonicalForm
      a   lemon:Form ;
      lemon:writtenRep   "Web_semantico"@IT.

bn:Web_semantico_IT/s02276858n
      a   lemon:LexicalSense ;
      dc:source   <http://wikipedia.org/>;
      dcterms:license  <http://creativecommons.org/licenses/by-sa/3.0/>;
      bn-lemon:wikipediaPage   wikipedia-it:Web_semantico;
      lemon:reference   bn:s02276858n .

bn:s02276858n
      a   skos:Concept;
      bn-lemon:synsetType   "NE";
      bn-lemon:synsetID   "bn:02276858n";
      bn-lemon:wikipediaCategory   wikipedia-da:Kategori:Internet;
      lemon:isReferenceOf   bn:Web_semantico_IT/s02276858n ...;
      skos:exactMatch   dbpedia:Semantic_Web;
      bn-lemon:definition   bn:s02276858n_Gloss1_DE ... ;
      dcterms:license  <http://creativecommons.org/licenses/by-nc-sa/3.0/>;
      skos:related   bn:s00076736n , bn:s03586460n ... .

bn:s02276858n_Gloss1_DE
      a   bn-lemon:BabelGloss;
      bn-lemon:gloss   "Das Semantische Web ist... "@DE ;
      lemon:language   "DE" ;
      dc:source   <http://wikipedia.org/>;
      dcterms:license  <http://creativecommons.org/licenses/by-sa/3.0/> .
```

Figure 3: An excerpt of BabelNet as RDF in Turtle format.

[13] http://www.w3.org/TR/skos-reference

[14] Lexical senses URIs are based on the 'full' lemma of Babel senses; when originating from Wikipedia, they are thus made up from the sense-tagged lemmas as in 'Apple_(Fruit)' and Apple_(Computer).

[15] http://www.w3.org/TeamSubmission/turtle/

Resource	
# SKOS concepts	9,348,287
# Babel glosses	17,961,157
# semantic relations	262,663,251
# lemon senses	50,282,542
# lemon lexical entries	44,486,335
# lemon lexicons	51
Outgoing links	
# Wikipedia pages	35,784,593
# Wikipedia categories	45,520,563
# DBpedia categories	15,381,861
# DBpedia pages	3,829,053
# lemon WordNet 3.0 links	117,657
# lemon OmegaWiki links (En)	15,140
Total number of outgoing links	100,648,867
Total number of triples	1,138,337,378

Table 1: Statistics concerning the *lemon*-BabelNet 2.0 RDF dataset.

Based on a *lemon*-SKOS model, the RDF edition of BabelNet is able to render most of the information contained in the stand-alone version, offering a large multi-domain and linguistic linked dataset, associated with an extensive multilingual lexical coverage. Yet, beyond its content, one of the key features of a linked dataset is to set connections to other datasets and to be accessible over the Web.

5. Interlinking and Publishing on the Web

5.1. Interlinking *lemon*-BabelNet

Generated from the integration of various existing resources, the most natural way of linking *lemon*-BabelNet is to consider the RDF versions, if available, of these resources. *lemon*-BabelNet includes in the first place links to encyclopedic resources: links to Wikipedia pages are established at the sense level (when originating from Wikipedia), and links to Wikipedia *category* pages at the SKOS concept level. These links are set up from the Wikipedia dump from which the resource is derived. Regarding DBpedia, links are set at the SKOS level only, with pointers to DBpedia English pages and English category pages. The URIs of these links are set up by swapping Wikipedia namespace for the DBpedia one[16]; no links are provided towards localized versions of DBpedia for now. Additionally, we provide links to lexical resources by setting connections to the *lemon* versions of WordNet 3.0[17] and OmegaWiki[18] (English version), both at the SKOS concept level. In both cases, URIs are taken from the RDF dumps of these datasets, using the synsets IDs to match the resources.

5.2. Statistics

The RDF version of BabelNet 2.0 features an overall number of 1.1 billion triples. Table 1 gives further details about the nature of these triples, which naturally reflect the standalone version, especially for SKOS concepts and lemon lexical senses. Most importantly, the resource contains a significant number of outgoing links, with around 80 million connections to either Wikipedia pages or categories, 19 million similar relations to DBpedia and, at the level of genuine lexical knowledge, a complete linkage to the *lemon* edition of Princeton WordNet 3.0 and 15k links to the English OmegaWiki edition of *lemon*-UBY. These connections to other *lemon* resources are of particular interest as they lay the foundations for further linked data-based integration of ontology lexica.

5.3. Publication on the Web

BabelNet 2.0 is published under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 License, as its linked data edition. Additionally, as it is based on a collection of independent resources, special attention must be paid to the licensing policies of these grounding works. *lemon*-BabelNet respects the copyrights of the original resources, and reproduces the different licenses under which they were issued, in two different ways: by releasing different RDF dump files according to groups of compatible licenses in the first place, by specifying a license property (dcterms:license) on triples in the second. As advocated by Rodríguez-Doncel et al. (2013), our aim is to achieve maximum transparency, which such explicit rights declarations should guarantee.

On a more concrete standpoint, BabelNet is served on the Web in three ways, *via*:

- a set of RDF dump files (URIs and IRIs) in n-triples format downloadable at the following URL: `http://babelnet.org/download.jsp`,
- a public SPARQL endpoint set up using the Virtuoso Universal Server[19] and accessible from the following URL: `http://babelnet.org:8084/sparql/`, and
- dereferenceable URIs, supported by the Pubby Web application, a Linked Data frontend for SPARQL endpoints[20] (`http://babelnet.org/2.0/`).

6. Related work

The work presented here relates to, first, the design of models for representing linguistic information on the Web and, second, to the publication of (multilingual) lexical resources according to Linked Data principles. Accordingly, we situate this work with respect to those areas and ask the reader to refer to (Navigli and Ponzetto, 2012) and (Hovy et al., 2013) for further information on knowledge acquisition and automatic construction of large-scale lexico-semantic resources.

lemon derives from a number of models for the representation of lexical data both on the Web and in off-line format. In particular, *lemon* was designed to combine the strengths of the LexInfo (Cimiano et al., 2011) and the Linguistic Information Repository (Montiel-Ponsoda et al., 2008), both of which were based on the Lexical Markup Framework (Francopoulo et al., 2006), which is an ISO standard

[16] http://dbpedia.org/resource/
[17] http://lemon-model.net/lexica/pwn/
[18] http://lemon-model.net/lexica/uby/ow_eng/
[19] http://virtuoso.openlinksw.com/
[20] http://wifo5-03.informatik.uni-mannheim.de/pubby/

for the representation of lexica. Meeting the challenge of connecting lexica to ontologies and representing them has become the goal of the OntoLex Community Group[21], who are recommending the next iteration of the *lemon* model as a W3C vocabulary.

We now turn to existing knowledge bases and lexical-semantic resources published as linked data. The largest "hub" of Linked Data is undeniably DBpedia, a large-scale, multilingual knowledge base extracted from Wikipedia (Lehmann et al., 2013; Auer et al., 2007). Started in 2006 with Wikipedia infobox information extraction processes, the DBpedia project has significantly expanded and matured over the years. It consists today of a large community committed to the building, expansion and dynamic upgrading of a knowledge base covering at present 111 languages. While DBpedia provides wide coverage of Named Entities, BabelNet focuses both on word senses and on Named Entities, which are, furthermore, cross-lingually interconnected in many languages. As a result, apparently different tasks such as Word Sense Disambiguation and Entity Linking can be performed jointly and with state-of-the-art performance in virtually any language of interest (Moro et al., 2014). With a similar focus on encyclopedic knowledge, the YAGO2 ontology (Hoffart et al., 2013) provides millions of facts and entities, some of which are spatially and temporally anchored. As for YAGO (Suchanek et al., 2008), it is based on the integration of Wikipedia and WordNet, whose mapping relies on the most frequent sense heuristic. Conversely, BabelNet integrates these two resources by means of a mapping strategy based on a disambiguation algorithm and, as mentioned above, provides additional lexicalizations resulting from the application of machine translation and the further integration of OmegaWiki and OMWN.

From a more lexically-oriented perspective, several resources have been published in the linked data format (de Melo and Weikum, 2008; Assem et al., 2006), some of which used the *lemon* model. It is for example the case of UBY (Gurevych et al., 2012), a large-scale lexical-semantic resource built from the integration of nine lexical resources[22] in two languages, English and German. Some of its lexica were integrated into the Semantic Web through their conversion to the ontology lexicon model, the result of which is interlinked linguistic datasets: *lemon*-UBY (Eckle-Kohler et al., 2014). This resource provides particularly rich lexical knowledge for verbs with information on their syntactic behaviour and semantic roles, a type of information which is complementary to the knowledge BabelNet provides. These two resources are linked through the English OmegaWiki *lemon* lexicon, as explained in Section 5.1. Finally, in addition to previous exports of other lexical resources (McCrae et al., 2012a), a recent work by Unger et al. (2013) proposes a *lemon* lexicon for the DBpedia ontology. The first version covers the most frequent classes and properties of the schema, and provides manually created lexical entries for English. We believe that the continuation of this work could benefit greatly from the availability of a resource such as *lemon*-BabelNet.

[21] http://www.w3.org/community/ontolex

[22] WordNet (en), GermaNet (de), VerbNet (en), FrameNet (en), Wikipedia (en, de), Wiktionary (en, de), and OmegaWiki (en, de).

7. Conclusion

In this paper we presented the publication of BabelNet 2.0 as Linked Data using *lemon*, a lexicon model for ontology. *lemon*-BabelNet features more than 1 billion triples which describe 9.3 million concepts with encyclopedic and lexical information in 50 languages. The resource is interlinked with several other datasets including DBpedia as nucleus of the LOD cloud. Our hope is that such a wide, multilingual and interconnected lexical-semantic dataset might support the Semantic Web in its ongoing maturation process towards, among others, multilinguality management. Furthermore, together with other newcomers in the LLOD, the linked data edition of BabelNet represents a major opportunity for NLP. Indeed, if carefully published and interlinked, these resources could, potentially, turn into a huge body of machine-readable knowledge. Future work naturally includes the upgrading of *lemon*-BabelNet to take account of any expansion of BabelNet itself, e.g., its full taxonomization (Flati et al., 2014) and validation (Vannella et al., 2014), as well as the diversification and integration of links to other resources (Pilehvar and Navigli, 2014).

Acknowledgments

Sapienza affiliated authors gratefully acknowledge the support of the ERC Starting Grant MultiJEDI No. 259234.

The authors also acknowledge support from the LIDER project (No. 610782), a support action funded by the European Commission under FP7. Warm thanks go to Victor Rodríguez-Doncel for the helpful discussion on (linked data) copyrights.

8. References

M. Van Assem, A. Gangemi, and G. Schreiber. 2006. Conversion of WordNet to a standard RDF/OWL representation. In *Proceedings of the* 5^{th} *International Conference on Language Resources and Evaluation, Genoa, Italy*, pages 237–242.

S. Auer, C. Bizer, G.Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ive. 2007. DBpedia: A Nucleus for a Web of Open Data. In *Proceedings of 6th International Semantic Web Conference joint with 2nd Asian Semantic Web Conference (ISWC+ASWC 2007)*, pages 722–735, Busan, Korea.

S. Auer, J. Lehmann, and A. N. Ngomo. 2011. Introduction to linked data and its lifecycle on the web. In *Reasoning Web. Semantic Technologies for the Web of Data*, pages 1–75. Springer.

C. Bizer, T. Heath, and T. Berners-Lee. 2009. Linked data-the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3):1–22.

F. Bond and R. Foster. 2013. Linking and extending an Open Multilingual Wordnet. In *Proc. of* 51^{st} *Annual Meeting of the Association for Computational Linguistics*, pages 1352–1362.

P. Buitelaar, P. Cimiano, P. Haase, and M. Sintek. 2009. Towards linguistically grounded ontologies. In *Proc. of the* 6^{th} *Annual European Semantic Web Conference*, pages 111–125.

P. Buitelaar, Key-Sun K.S. Choi, P. Cimiano, and E. Hovy. 2012. The Multilingual Semantic Web. Technical Report 12362, Report from the Dagstuhl Seminar.

P. Buitelaar. 2010. Ontology-based semantic lexicons: Mapping between terms and object descriptions. *Ontology and the Lexicon*, pages 212–223.

C. Chiarcos, S. Hellmann, and S. Nordhoff. 2011. Towards a linguistic linked open data cloud: The open linguistics working group. *TAL (Traitement automatique des langues)*, 52(3):245–275.

C. Chiarcos, S. Hellmann, and S. Nordhoff. 2012. Linking linguistic resources: Examples from the open linguistics working group. In C. Chiarcos, S. Nordhoff, and S. Hellmann, editors, *Linked Data in Linguistics. Representing Language Data and Metadata*, pages 201–216. Springer.

C. Chiarcos, J. McCrae, P. Cimiano, and C. Fellbaum. 2013. Towards open data for linguistics: Linguistic linked data. In A. Oltramari, P. Vossen, L. Qin, and E. Hovy, editors, *New Trends of Research in Ontologies and Lexical Resources*, pages 7–25. Springer.

P. Cimiano, P. Buitelaar, J. McCrae, and M. Sintek. 2011. Lexinfo: A declarative model for the lexicon-ontology interface. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(1):29–51, March 2011.

Gerard de Melo and Gerhard Weikum. 2008. Language as a foundation of the semantic web. In *International Semantic Web Conference (Posters & Demos)*.

J. Eckle-Kohler, J. McCrae, and C. Chiarcos. 2014. *lemon*Uby-a large, interlinked, syntactically-rich resource for ontologies. *Semantic Web Journal, Special issue on Multilingual Linked Open Data*.

C. Fellbaum, editor. 1998. *WordNet: An Electronic Database*. MIT Press, Cambridge, MA.

T. Flati, D. Vannella, T. Pasini, and R. Navigli. 2014. Two Is Bigger (and Better) Than One: the Wikipedia Bitaxonomy Project. In *Proc. of the* 52^{nd} *Annual Meeting of the Association for Computational Linguistics*.

G. Francopoulo, M. George, N. Calzolari, M. Monachini, N. Bel, M. Pet, C. Soria, et al. 2006. Lexical markup framework (lmf). In *International Conference on Language Resources and Evaluation*, pages 233–236.

A. Gómez-Pérez, D. Vila-Suero, E. Montiel-Ponsoda, J. Gracia, and G. Aguado de Cea. 2013. Guidelines for multilingual linked data. In *Proceedings of the* 3^{rd} *International Conference on Web Intelligence, Mining and Semantics*, page 3. ACM.

J. Gracia, E. Montiel-Ponsoda, P. Cimiano, A. Goméz-Pérez, P. Buitelaar, and J. McCrae. 2011. Challenges for the multilingual Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 11:63–71.

I. Gurevych, J. Eckle-Kohler, . Hartmann, M. Matuschek, C. M. Meyer, and C. Wirth. 2012. UBY - A large-scale unified lexical-semantic resource based on LMF. In *Proc. of the* 13^{th} *Conference of the European Chapter of the Association for Computational Linguistics*, pages 580–590, Avignon, France.

T. Heath and C. Bizer. 2011. Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology*, 1(1):1–136.

S. Hellmann, J. Lehmann, S. Auer, and M. Brümmer. 2013. Integrating NLP using linked data. In *Proceedings of the* 12^{th} *International Semantic Web Conference*, pages 97–112.

J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum. 2013. YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence*, 194:28–61.

E. Hovy, R. Navigli, and S. P. Ponzetto. 2013. Collaboratively built semi-structured content and Artificial Intelligence: The story so far. *Artificial Intelligence*, 194:2–27.

N. Ide and J. Pustejovsky. 2010. What does interoperability mean, anyway? Towards an operational definition of interoperability for language technology. In *Proc. of the* 2^{nd} *Conference on Global Interoperability for Language Resources*.

F. Khan, F. Frontini, R. Del Gratta, M. Monachini, and V. Quochi. 2013. Generative lexicon theory and linguistic linked open data. In *Proceedings of the* 6^{th} *International Conference on Generative Approaches to the Lexicon*, pages 62–69.

J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer. 2013. DBpedia - a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web Journal*.

J. McCrae, D. Spohr, and P. Cimiano. 2011. Linking lexical resources and ontologies on the semantic web with lemon. In *The Semantic Web: Research and Applications*, pages 245–259. Springer.

J. McCrae, P. Cimiano, and E. Montiel-Ponsoda. 2012a. Integrating Wordnet and Wiktionary with lemon. *Linked Data in Linguistics*, pages 25–34.

J. McCrae, G. Aguado de Cea, P. Buitelaar, P. Cimiano, T. Declerck, A. Gómez-Pérez, J. Gracia, L. Hollink, E. Montiel-Ponsoda, D. Spohr, et al. 2012b. Interchanging lexical resources on the semantic web. *Language Resources and Evaluation*, 46(4):701–719.

J. McCrae, E. Montiel-Ponsoda, and P. Cimiano. 2012c. Collaborative semantic editing of linked data lexica. In *Proceedings of the* 8^{th} *International Conference on Language Resource and Evaluation*, pages 2619–2625, Istanbul, Turkey.

E. Montiel-Ponsoda, G. Aguado de Cea, and A. Gómez Pérez. 2008. Modelling multilinguality in ontologies. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 67–70.

E. Montiel-Ponsoda, J. Gracia del Río, G. Aguado de Cea, and A. Gómez-Pérez. 2011. Representing translations on the semantic web. In *Proceedings of the* 2^{nd} *International Workshop on the Multilingual Semantic Web*, pages 30–42.

A. Moro, A. Raganato, and R. Navigli. 2014. Entity Linking meets Word Sense Disambiguation: A Unified Approach. *Transactions of the Association for Computational Linguistics*, 2.

R. Navigli and S. P. Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

R. Navigli, D. A. Jurgens, and D. Vannella. 2013. Semeval-2013 task 12: Multilingual word sense disambiguation. In *Proceedings of the* 7^{th} *International Workshop on Semantic Evaluation (SemEval-2013)*, pages 222–231, Atlanta, USA.

M. T. Pilehvar and R. Navigli. 2014. A Robust Approach to Aligning Heterogeneous Lexical Resources. In *Proceedings of the* 52^{nd} *Annual Meeting of the Association for Computational Linguistics*.

V. Rodríguez-Doncel, A. Gómez-Pérez, and N. Mihindukulasooriya. 2013. Rights declaration in linked data. In *Proc. of the* 3^{rd} *International Workshop on Consuming Linked Data*.

F. M. Suchanek, G. Kasneci, and G. Weikum. 2008. Yago: A large ontology from Wikipedia and WordNet. *Journal of Web Semantics*, 6(3):203–217.

C. Unger, J. McCrae, S. Walter, S. Winter, and P. Cimiano. 2013. A *lemon* lexicon for DBpedia. In S. Hellmann, A. Filipowska, C. Barriere, P. Mendes, and D. Kontokostas, editors, *Proc. of* 1^{st} *Int'l Workshop on NLP and DBpedia*, Sydney, Australia.

D. Vannella, D. Jurgens, D. Scarfini, D. Toscani, and R. Navigli. 2014. Validating and Extending Semantic Knowledge Bases using Video Games with a Purpose. In *Proc. of the* 52^{nd} *Annual Meeting of the Association for Computational Linguistics*.

M. Villegas and N. Bel. 2013. PAROLE/SIMPLE 'lemon' ontology and lexicons. *Semantic Web Journal*.