A Quick Tour of BabelNet 1.1

Roberto Navigli

Dipartimento di Informatica
Sapienza Università di Roma
Viale Regina Elena, 295 – Roma, Italy
navigli@di.uniroma1.it
http://lcl.uniroma1.it

Abstract. In this paper we present BabelNet 1.1, a brand-new release of the largest "encyclopedic dictionary", obtained from the automatic integration of the most popular computational lexicon of English, i.e. WordNet, and the largest multilingual Web encyclopedia, i.e. Wikipedia. BabelNet 1.1 covers 6 languages and comes with a renewed Web interface, graph explorer and programmatic API. BabelNet is available online at http://www.babelnet.org.

Keywords: BabelNet, knowledge acquisition, semantic networks, multilingual ontologies.

1 Introduction

In the information society, lexical knowledge is a key skill for understanding and decoding an ever-changing world. Indeed, lexical knowledge is not only an essential component for human understanding of text, it is also indispensable for Natural Language Processing tasks. Unfortunately, however, building such lexical knowledge resources manually is an onerous task requiring dozens of years – and what is more it has to be repeated from scratch for each new language.

The multilinguality aspect is key to this vision, in that it enables Natural Language Processing tasks which are not only cross-lingual, but are also independent of the language of the user input and of the other data utilized to perform the task.

In this paper we present BabelNet 1.1 (http://www.babelnet.org), a brand-new version of a very large multilingual ontology and semantic network, obtained as a result of a novel integration and enrichment methodology [11]. This resource is created by linking the largest multilingual Web encyclopedia – i.e., Wikipedia – to the most popular computational lexicon – i.e., WordNet [6]. The integration is performed via an automatic mapping and by filling in lexical gaps in resource-poor languages with the aid of Machine Translation (MT). The result is an "encyclopedic dictionary" that provides babel synsets, i.e., concepts and named entities lexicalized in many languages and connected with large amounts of semantic relations.

BabelNet complements existing resources, such as DBPedia [1], YAGO [4] or WikiNet [8]. In fact, these resources just provide coverage of Named Entities and

A. Gelbukh (Ed.): CICLing 2013, Part I, LNCS 7816, pp. 25–37, 2013.

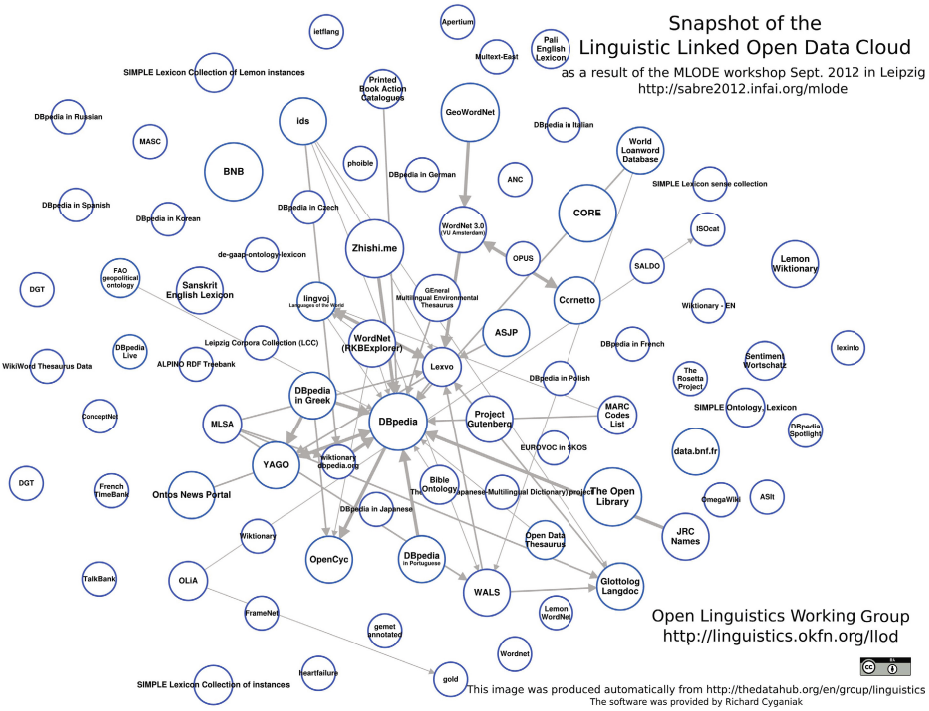

Fig. 1. Open Linguistics Working Group (2012), The Linguistic Linked Open Data cloud diagram (draft), version of September 2012, `http://linguistics.okfn.org/llod`.

more or less explicitly target the Linked Open Data (LOD) cloud, i.e. a vision of the (Semantic) Web in which related data that was not previously linked is connected. BabelNet, instead, focuses both on word senses and on Named Entities in many languages. Its aim, therefore, is to provide full lexicographic and encyclopedic coverage. BabelNet can be viewed both as a multilingual ontology, a large machine-readable encyclopedic dictionary and a multilingual semantic network. Compared to YAGO [16], which links Wikipedia categories to WordNet synsets using the first-sense heuristic, BabelNet integrates the two resources by means of a mapping strategy based on a Word Sense Disambiguation algorithm [9], and provides additional lexicalizations resulting from the application of Statistical Machine Translation. This is also in contrast to Wikipedia-based resources like WikiNet, which utilizes the lexicalizations available in Wikipedia articles and categories, or the Universal WordNet (UWN) [5], which is based on the lexicographic coverage of words.

Recently, BabelNet has also been linked to the LOD cloud [10], with the objective of contributing to the so-called Linguistic Linked Open Data (LLOD, see Figure 1), a vision fostered by the Open Linguistic Working Group (OWLG)[1]

[1] `http://linguistics.okfn.org`

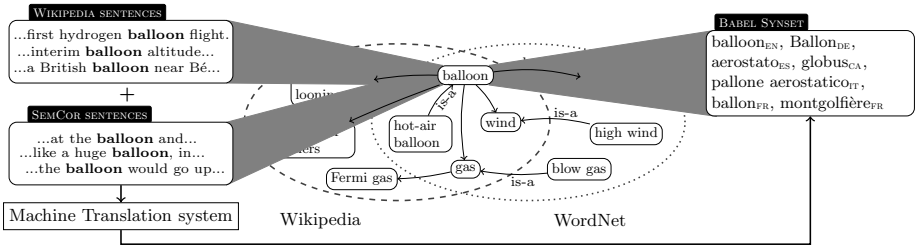

Fig. 2. An overview of BabelNet (nodes are labeled with English lexicalizations only): unlabeled edges are extracted from Wikipages (e.g., BALLOON (AIRCRAFT) links to MONTGOLFIER BROTHERS), labeled edges come from WordNet (e.g., balloon^1_n *has-part* gasbag^1_n).

in which part of the Linked Open Data cloud is made up of interlinked linguistic resources [2], aimed at fostering integration, interoperability and reuse on the Semantic Web [3].

In the next Section we introduce BabelNet 1.1 and briefly illustrate its features. Then, in Section 3 we provide statistics, in Section 4 we take a tour of the new Web interface and in Section 5 we show how to get started with the BabelNet API. Finally, we give some conclusions in Section 6.

2 BabelNet

BabelNet [11] encodes knowledge as a labeled directed graph $G = (V, E)$ where V is the set of nodes – i.e., concepts such as balloon and named entities such as Montgolfier brothers – and $E \subseteq V \times R \times V$ is the set of edges connecting pairs of concepts (e.g., balloon *is-a* lighter-than-air craft). Each edge is labeled with a semantic relation from R, e.g., $\{$*is-a*, *part-of*, ..., $\epsilon\}$, where ϵ denotes an unspecified semantic relation. Each node $v \in V$ contains a set of lexicalizations of the concept for different languages, e.g., { balloon$_{EN}$, Ballon$_{DE}$, pallone aerostatico$_{IT}$, ..., montgolfière$_{FR}$ }. We call such multilingually lexicalized concepts *Babel synsets*. Concepts and relations in BabelNet are harvested from the largest available semantic lexicon of English, WordNet, and a wide-coverage collaboratively-edited encyclopedia, Wikipedia. In order to build the BabelNet graph, we collect at different stages:

a. from WordNet, all available word senses (as concepts) and all the lexical and semantic pointers between synsets (as relations);
b. from Wikipedia, all encyclopedic entries (i.e., Wikipages, as concepts) and semantically unspecified relations from hyperlinked text.

An overview of BabelNet is given in Figure 2. The excerpt highlights that WordNet and Wikipedia can overlap both in terms of concepts and relations: accordingly, in order to provide a *unified resource*, we merge the intersection of these

two knowledge sources. Next, to enable multilinguality, we collect the lexical realizations of the available concepts in different languages. Finally, we connect the multilingual Babel synsets by establishing semantic relations between them. Thus, our methodology consists of three main steps:

1. We **combine WordNet and Wikipedia** by automatically acquiring a mapping between WordNet senses and Wikipages. This avoids duplicate concepts and allows their inventories of concepts to complement each other.
2. We **harvest multilingual lexicalizations** of the available concepts (i.e., Babel synsets) by using (a) the human-generated translations provided by Wikipedia (the so-called *inter-language* links), as well as (b) a machine translation system to translate occurrences of the concepts within sense-tagged corpora.
3. We **establish relations between Babel synsets** by collecting all relations found in WordNet, as well as all wikipedias in the languages of interest: in order to encode the strength of association between synsets we compute their degree of correlation using a measure of relatedness based on the Dice coefficient.

3 Statistics

In this section we provide statistics for BabelNet 1.1, obtained by applying the construction methodology briefly described in the previous Section and detailed in [11].[2]

3.1 Lexicon

BabelNet currently covers 6 languages, namely: English, Catalan, French, German, Italian and Spanish. Its lexicon includes lemmas which denote both concepts (e.g., balloon) and named entities (e.g., Montgolfier brothers). The second column of Table 1 shows the number of lemmas for each language. The lexicons have the same order of magnitude for the 5 non-English languages, whereas English displays larger numbers due to the lack of inter-language links and annotated sentences for many terms, which prevents our construction approach from providing translations.

In Table 2 we report the number of monosemous and polysemous words divided by part of speech. Given that we work with nominal synsets only, the numbers for verbs, adjectives and adverbs are the same as in WordNet 3.0. As for nouns, we observe a very large number of monosemous words (almost 19 million),

[2] However, note that some additional heuristics, especially concerning Wikipedia redirections, have been applied in version 1.1 in order to improve the quality of the Wikipedia-WordNet mapping.

Table 1. Number of lemmas, synsets and word senses in the 6 languages currently covered by BabelNet

Language	Lemmas	Synsets	Senses
English	8,108,298	3,898,579	8,891,049
Catalan	1,319,089	754,187	1,483,131
French	2,901,230	1,530,017	3,222,211
German	2,846,656	1,591,536	3,100,059
Italian	1,989,959	1,264,333	2,212,186
Spanish	2,678,657	1,236,250	3,038,745
Total	19,843,889	5,581,954	21,947,381

Table 2. Number of monosemous and polysemous words by part of speech (verbs, adjectives and adverbs are the same as in WordNet 3.0)

POS	Monosemous words	Polysemous words
Noun	18,722,263	1,084,076
Verb	6,280	5,251
Adjective	16,591	4,947
Adverb	3,748	733
Total	18,748,882	1,095,007

but also a large number of polysemous words (more than 1 million). Both numbers are considerably larger than in WordNet, because – as remarked above – words here denote both concepts (mainly from WordNet) and named entities (mainly from Wikipedia).

3.2 Concepts

BabelNet 1.1 contains more than 5.5 million concepts, i.e., Babel synsets, and almost 22 million word senses (regardless of their language). In Table 1 we report the number of synsets covered for each language (third column) and the number of word senses lexicalized in each language (fourth column). The overall number of word senses in English is much higher than those in the other languages (owing to the high number of synonyms and redirections in English).

In Table 3 we show for each language the number of word senses obtained directly from WordNet, Wikipedia pages and redirections, as well as Wikipedia and WordNet translations.

3.3 Relations

We now turn to relations in BabelNet. Relations come either from Wikipedia hyperlinks (in any of the covered languages) or WordNet. All our relations are

Table 3. Composition of Babel synsets: number of synonyms from the English WordNet, Wikipedia pages and translations, as well as translations of WordNet's monosemous words and SemCor's sense annotations

		English	Catalan	French	German	Italian	Spanish	Total
English WordNet		206,978	-	-	-	-	-	206,978
Wikipedia	pages	3,829,110	375,046	1,231,374	1,287,134	932,536	915,046	8,570,246
	redirections	4,854,961	250,681	1,145,075	954,949	438,856	1,272,184	8,916,706
	translations	-	749,337	738,128	749,381	732,342	743,389	3,712,577
WordNet	translations	-	108,067	107,634	108,595	108,452	108,126	540,874
Total		8,891,049	1,483,131	3,222,211	3,100,059	2,212,186	3,038,745	21,947,381

Table 4. Number of lexico-semantic relations harvested from WordNet, WordNet glosses and the 6 wikipedias

	English	Catalan	French	German	Italian	Spanish	Total
WordNet	364,552	-	-	-	-	-	364,552
WordNet glosses	617,785	-	-	-	-	-	617,785
Wikipedia	68,626,079	4,553,922	18,390,507	21,942,058	14,443,937	12,758,598	140,715,101
Total	69,608,416	4,553,922	18,390,507	21,942,058	14,443,937	12,758,598	141,697,438

Table 5. Glosses for the Babel synset referring to the concept of balloon as aircraft

English	WordNet	Large tough nonrigid bag filled with gas or heated air.
	Wikipedia	A balloon is a type of aircraft that remains aloft due to its buoyancy.
German		Ein Ballon ist eine nicht selbsttragende, gasdichte Hülle, die mit Gas gefüllt ist und über keinen Eigenantrieb verfügt.
Italian		Un pallone aerostatico è un tipo di aeromobile, un aerostato che si solleva da terra grazie al principio di Archimede.
Spanish		Un aerostato, o globo aerostático, es una aeronave no propulsada que se sirve del principio de los fluidos de Arquímedes para volar, entendiendo el aire como un fluido.

semantic, in that they connect Babel synsets (rather than senses), however the relations obtained from Wikipedia are unlabeled.[3] In Table 4 we show the number of lexico-semantic relations from WordNet, WordNet glosses and the 6 wikipedias used in our work. We can see that the major contribution comes from the English Wikipedia (68 million relations) and Wikipedias in other languages (a few million relations, depending on their size in terms of number of articles and links therein).

3.4 Glosses

Each Babel synset comes with one or more glosses (possibly available in many languages). In fact, WordNet provides a textual definition for each English synset, while in Wikipedia a textual definition can be reliably obtained from

[3] In a future release of the resource we plan to perform an automatic labeling based on work in the literature. See [7] for recent work on the topic.

the first sentence of each Wikipage[4]. Overall, BabelNet 1.1 includes 8,439,497 glosses (3,905,508 of which are in English). In Table 5 we show the glosses for the Babel synset which refer to the concept of balloon as 'aircraft'.

4 The Online Interface

The new online interface, shown in Figure 3 and available from `http://www.babelnet.org`, has 2 interactive options: search and explore. We describe the two options in the following subsections.

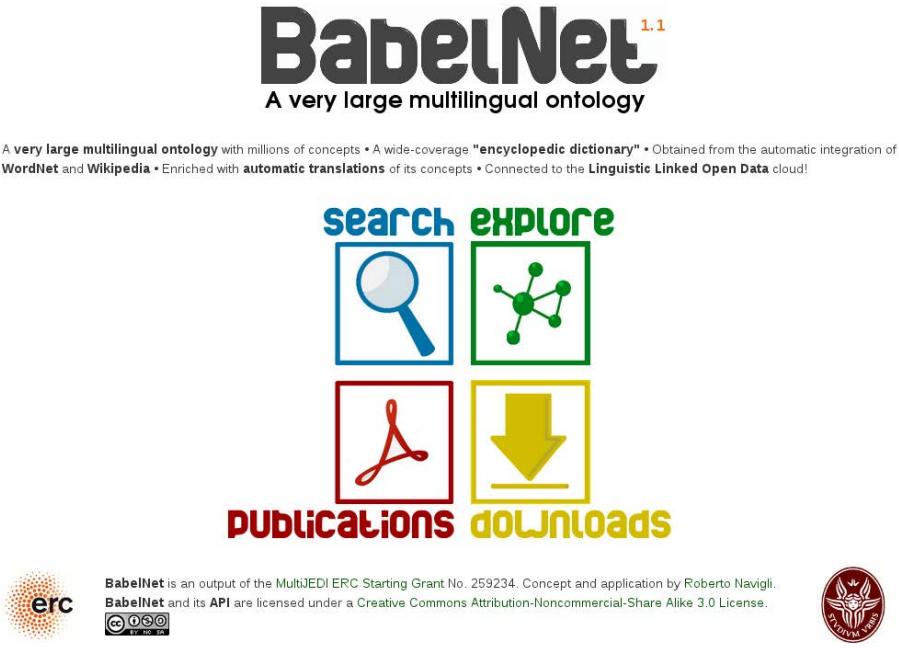

Fig. 3. The BabelNet homepage

4.1 Searching BabelNet

The search interface allows the user to look up a word in BabelNet by specifying its lemma and the language of interest (among the 6 available). In Figure 4 we

[4] "The article should begin with a short declarative sentence, answering two questions for the nonspecialist reader: *What (or who) is the subject?* and *Why is this subject notable?*", extracted from `http://en.wikipedia.org/wiki/Wikipedia:Writing_better_articles`. This simple, albeit powerful, heuristic has previously been used successfully to construct a corpus of definitional sentences [15] and learn a definition and hypernym extraction model [14].

show the first two results when *balloon* is searched in the English language. Each entry in the search result page shows several kinds of information:

- **Basic information**, such as the main title (e.g. *balloon*[1] for the first nominal sense of balloon, following the notation of [9]), its unique ID (e.g. bn:00008187n) and its type (i.e. Concept vs. Named Entity).
- **the WordNet synset:** this includes all the synonyms in the synset which, if a mapping is available, denotes the target concept in WordNet.
- **Wikipedia titles:** i.e. the Wikipedia page titles in the various languages for the target concept. For instance, for the aircraft sense of balloon, we have Globus aerostàtic for Catalan, Balloon (aircraft) for English, Aérostation for French, etc.
- **Wikipedia redirections:** this includes the Wikipedia redirections to the corresponding Wikipage page titles. For instance, Hydrogen balloon redirects to Balloon (aircraft).
- **Automatic translations**, i.e. the majority translations obtained as a result of the application of Statistical Machine Translation to sense-tagged sentences. For instance, globo in Spanish was a majority translation of the aircraft sense of balloon.
- **Glosses:** we show the textual definitions available in the covered languages. For English, we might have two different definitions, one from WordNet, one from Wikipedia, if a link could be established.
- **Wikipedia categories:** the categories provided for the above Wikipedia page titles.
- **A link to DBPedia:** a hyperlink to the corresponding DBPedia entry is provided if available.

For instance, the lack of translations for the second sense of balloon (Toy balloon) in Figure 4 implies either that there were not enough annotated sentences with the sense of interest or that no majority translation could be found.

4.2 Exploring BabelNet

Beside each sense of a given lemma an "explore" link is available which allows the user to move to the exploration interface. Alternatively, the user can click on the link with the same name on top of the page and specify the lemma she wants to explore. In either case, a graph representation of the selected Babel synset is shown.

The graph shows up to 50 neighbours of the selected Babel synset. For example in Figure 5 we show the graph obtained from the interface for the first sense of balloon. The user can click on any node in the graph and follow that concept, thus moving along the graph edges. Note that for presentation purposes only the main title of each Babel synset is shown for each node.

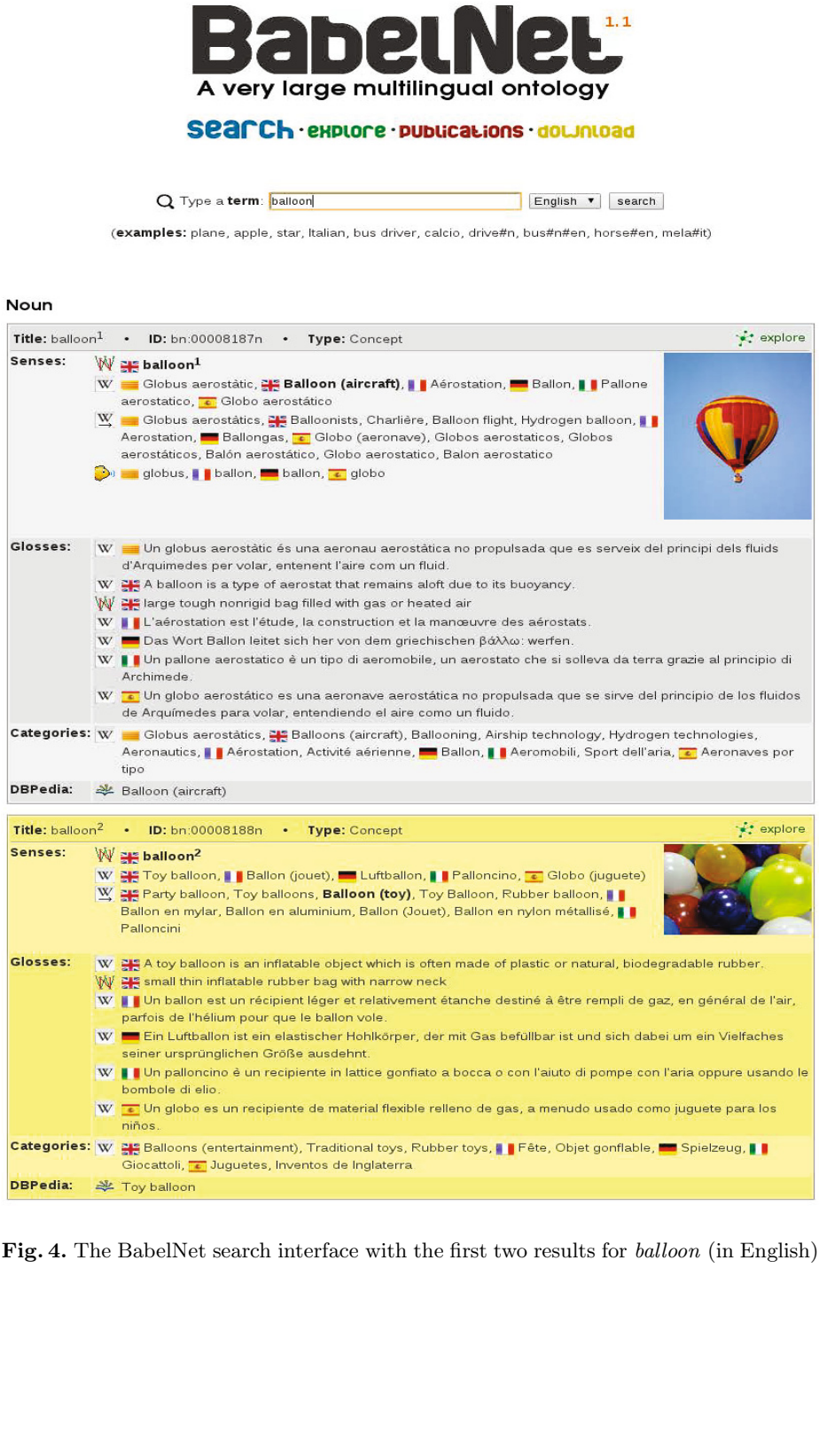

Fig. 4. The BabelNet search interface with the first two results for *balloon* (in English)

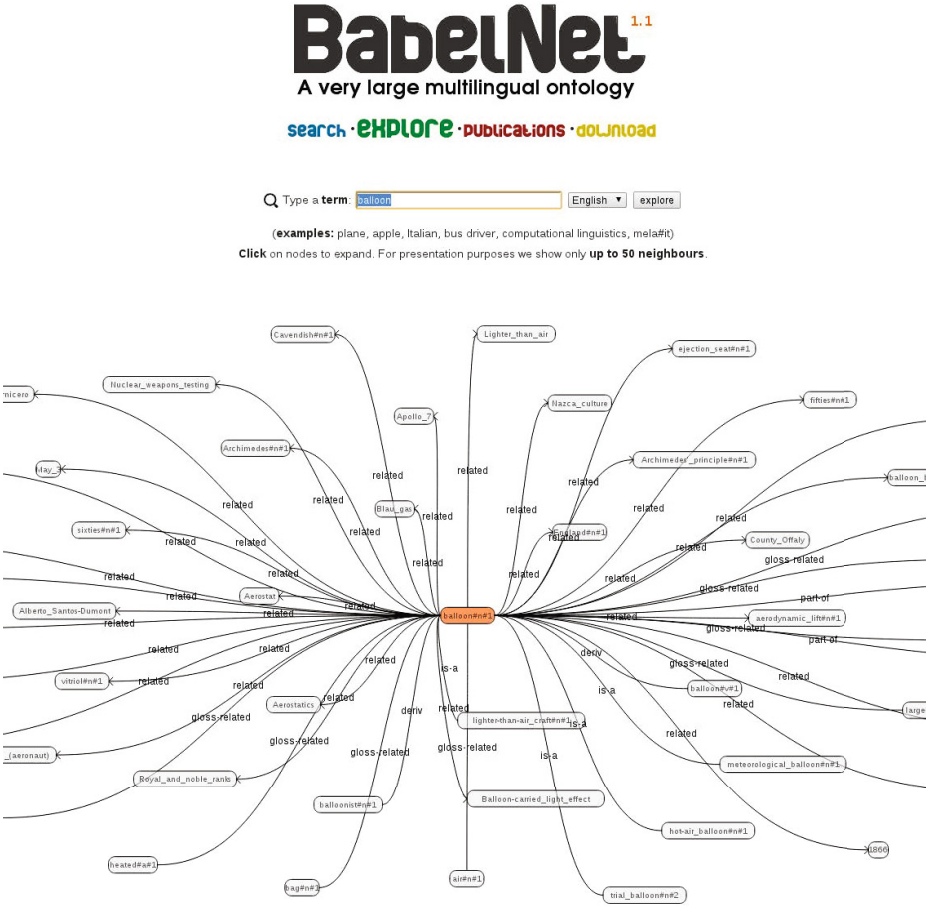

Fig. 5. The BabelNet explore interface with the first sense of *balloon* (in English)

5 The BabelNet API

Access to the BabelNet Java API can be obtained from the homepage (bottom-right icon of Figure 3). We will now take a quick tour of the main features of the API.

5.1 Creating a BabelNet Instance

An instance of BabelNet can be obtained with just one line of code:

```
BabelNet bn = BabelNet.getInstance();
```

5.2 Obtaining the Senses of a Lemma

We can obtain the senses of a given lemma as follows:

```
List<BabelSense> senses =
    bn.getSenses(languageToSearch, lemma, pos, includeRedirects);
```

where `languageToSearch` is any of the supported languages in BabelNet (i.e. `Language.CA`, `.EN`, `.ES`, `.FR`, `.DE`, `.IT`), `pos` is any of the four open-class part-of-speech tags (`POS.NOUN`, `.ADJECTIVE`, `.ADVERB`, `.VERB`) and `includeRedirects` is a boolean specifying whether we also want senses matching the lemma in the Wikipedia redirections. The `getSenses` method returns a list of `BabelSense` instances.

5.3 Obtaining the Babel Synsets Containing a Lemma

Similarly to the above method for obtaining senses, we can query BabelNet for synsets containing the input lemma:

```
List<BabelSynset> synsets =
    bn.getSynsets(languageToSearch, lemma, pos, includeRedirects);
```

5.4 Sorting Senses and Synsets

Senses can be sorted using the `BabelSenseComparator` according to the following criteria: WordNet senses first, sorted by sense number, next, Wikipedia senses in lexicographic order:

```
Collections.sort(senses, new BabelSenseComparator());
```

Similarly, one can sort synsets as follows:

```
Collections.sort(synsets, new BabelSynsetComparator());
```

using the `BabelSynsetComparator`.

5.5 Getting a String Representation for a Sense

To obtain a string representation for a given `BabelSense`:

```
System.out.println(sense.getSenseString());
```

where `sense` is a `BabelSense`. What is returned by the `getSenseString` method depends on the information available for the input sense:

- a lemma-pos-sense representation if it is a WordNet sense (e.g. balloon#n#1);
- the Wikipedia page title if it corresponds to a Wikipedia sense (e.g. Balloon (aircraft));
- the lemma if it is a translation (e.g. Ballon).

5.6 Important Information in a Babel Synset

Babel synsets contain many different kinds of information available by means of its methods, among which we have:

- `getId`, which returns the synset id;
- `getPOS`, which provides the synset's part-of-speech tag;
- `getSynsetSource`, which returns the synset source, i.e. whether it belongs to WordNet only, Wikipedia only or both (i.e. a mapping between a Wikipedia page and a WordNet sense could be established);
- `getSynsetType`: does the synset denote a `NAMED_ENTITY` or a `CONCEPT`?
- `getSenses`, which returns the list of senses in the synset;
- `getTranslations`, which returns a source-to-target map of the translations of Babel senses;
- `getRelatedMap`, which returns a map of all Babel synsets connected to the given synset by means of lexical-semantic relations;
- `getCategories`, which returns a list of Wikipedia categories for the given synset;
- `getImages`, which returns a list of image URIs for the given synset;
- `getWordNetOffsets`, which returns a list of WordNet offsets to which the synset is linked;
- `getMainSense`, which returns the main sense of the synset.

6 Conclusions

In this paper we have taken a quick tour of BabelNet 1.1, a very large multilingual ontology available at `http://www.babelnet.org`. We first introduced BabelNet, provided statistics for the new version 1.1, overviewed the new interface for search and graph exploration, and surveyed the Java programmatic API. This resource has already proven useful in tasks such as multilingual Word Sense Disambiguation [13] and semantic relatedness [12], but many horizons are still to be explored.

Acknowledgments.

The author gratefully acknowledges the support of the ERC Starting Grant MultiJEDI No. 259234. Thanks go to Daniele Vannella for his help on calculating statistics.

References

1. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: Dbpedia - a crystallization point for the web of data. Journal of Web Semantics 7(3), 154–165 (2009)
2. Chiarcos, C., Hellmann, S., Nordhoff, S.: Towards a linguistic linked open data cloud: The Open Linguistics Working Group. TAL 52(3), 245–275 (2011)

3. Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Gómez-Pérez, A., Buitelaar, P., McCrae, J.: Challenges for the multilingual web of data. J. Web Sem. 11, 63–71 (2012)
4. Hoffart, J., Suchanek, F.M., Berberich, K., Weikum, G.: Yago2: A spatially and temporally enhanced knowledge base from wikipedia. Artificial Intelligence 194, 28–61 (2013)
5. de Melo, G., Weikum, G.: Towards a universal wordnet by learning from combined evidence. In: Proceedings of the Eighteenth ACM Conference on Information and Knowledge Management, Hong Kong, China, pp. 513–522 (2009)
6. Miller, G.A., Beckwith, R., Fellbaum, C.D., Gross, D., Miller, K.: WordNet: an online lexical database. International Journal of Lexicography 3(4), 235–244 (1990)
7. Moro, A., Navigli, R.: WiSeNet: Building a Wikipedia-based semantic network with ontologized relations. In: Proceedings of the 21st ACM Conference on Information and Knowledge Management (CIKM 2012), Maui, HI, USA (2012)
8. Nastase, V., Strube, M.: Transforming wikipedia into a large scale multilingual concept network. Artificial Intelligence 194, 62–85 (2013)
9. Navigli, R.: Word Sense Disambiguation: A survey. ACM Computing Surveys 41(2), 1–69 (2009)
10. Navigli, R.: BabelNet goes to the (Multilingual) Semantic Web. In: ISWC 2012 Workshop on Multilingual Semantic Web (2012)
11. Navigli, R., Ponzetto, S.P.: BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence 193, 217–250 (2012)
12. Navigli, R., Ponzetto, S.P.: BabelRelate! a joint multilingual approach to computing semantic relatedness. In: Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, AAAI (2012)
13. Navigli, R., Ponzetto, S.P.: Joining forces pays off: Multilingual joint word sense disambiguation. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp. 1399–1410 (2012)
14. Navigli, R., Velardi, P.: Learning Word-Class Lattices for definition and hypernym extraction. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL), Uppsala, Sweden, pp. 1318–1327 (2010)
15. Navigli, R., Velardi, P., Ruiz-Martínez, J.M.: An annotated dataset for extracting definitions and hypernyms from the web. In: Proceedings of the 7th International Conference on Language Resources and Evaluation, Valletta, Malta, May 19-21 (2010)
16. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: A large ontology from Wikipedia and WordNet. Journal of Web Semantics 6(3), 203–217 (2008)