# DBlexipedia: A nucleus for a multilingual lexical Semantic Web

Sebastian Walter, Christina Unger, and Philipp Cimiano

Semantic Computing Group, CITEC, Bielefeld University

**Abstract.** A huge amount of datasets on the Semantic Web are linked to a few datahubs, the most prominent of which is DBpedia. What makes the exploitation of DBpedia challenging for natural language-based applications, however, is that such NLP applications require knowledge about how the ontology elements are verbalized in natural language. In order to provide such knowledge at the required scale and thereby leverage the use of DBpedia in different applications, we construct a lexicon for the DBpedia 2014 ontology by means of existing automatic methods for lexicon induction. It contains 11,998 lexical entries for 574 different properties in three languages: English, German, and Spanish. Just like DBpedia provides a hub for Semantic Web datasets, this lexicon can provide a hub for the lexical Semantic Web, an ecosystem in which ontology lexica are published, linked, and re-used across applications.

**Keywords:** Ontology lexicalization, DBpedia, *lemon*, M-ATOLL

## 1 Introduction

The amount of datasets on the Semantic Web is evermore increasing, and large part of these datasets are linked to central hubs. The biggest and arguably most important of those hubs is DBpedia [3], a general-purpose, multi-domain dataset extracted from Wikipedia, which is, for example, heavily used by systems that employ structured data for applications like web-based information retrieval or search. But what makes the exploitation of DBpedia challenging for a variety of natural language-based applications (such as question answering [4] and natural language generation [1]) is that they usually require knowledge about how the ontology elements are verbalized in natural language, including lexical variants across different languages. In order to provide such knowledge at the required scale and thereby leverage the use of DBpedia and all connected datasets in applications, we construct a lexicon for the DBpedia 2014 ontology by means of existing automatic methods for lexicon induction. Just like DBpedia provides a hub for Semantic Web datasets, this lexicon can provide a hub for the lexical Semantic Web, an ecosystem part of the linguistic linked data cloud[1] in which ontology lexica are published, linked, and re-used across applications.

For an example of a lexical entry consider the property `http://dbpedia.org/ontology/spouse` from the DBpedia 2014 ontology. This property expresses

[1] `http://linguistic-lod.org/llod-cloud`

that two persons are married to each other. The property can be expressed in natural language by the following expressions (lexical entries).

- *X is the husband of Y* e.g. *Barack Obama is the husband of Michelle Obama.*
- *X is the wife of Y* e.g. *Michelle Obama is the wife of Barack Obama.*
- *X is married to Y* e.g. *Barack Obama is married to Michelle Obama.*

All these lexical entries can be interpreted as expressing the spouse property. We say that the property spouse is the *reference* of the lexical entry.

To our knowledge, DBlexipedia is the first wide-coverage lexicon of DBpedia. It contains 11,998 lexical entries for 574 different RDF-properties from the DBpedia 2014 ontology in three languages: English, German, and Spanish.

In contrast to an earlier manually crafted DBpedia lexicon [9], DBlexipedia is automatically constructed and therefore can easily be updated as DBpedia evolves.

The remainder of the paper is structured as follows. In the next section, we present our approach to generating ontology lexica, followed by a description of the resulting lexicon in Section 3. We conclude with perspectives for future work.

## 2 Approach

The lexicon published on `http://dblexipedia.org` is the result of applying M-ATOLL[2] [11, 10] to the DBpedia ontology and a Wikipedia text corpus. M-ATOLL creates ontology lexica in *lemon* [6] format, and it is designed to be employed in a semi-automatic fashion, i.e. automatically constructing lexical entries that are then manually checked and corrected by a human.

In the following we briefly explain the main corpus-based approach as well as an extension of it that deals with a special case of adjective entries. We call this second approach label-based, as we use the label of a property in combination with machine learning techniques to generate the entries.

### 2.1 Corpus-based Approach

M-ATOLL takes as input an ontology and a dependency parsed text corpus in the target language. For DBlexipedia, the input was the DBpedia 2014 ontology and a Wikipedia text corpus parsed for three target languages: English, German, and Spanish. As dependency parser we use the MaltParser [7] for English, the ParZu parser [8] for German, and an online service with an instance of the Spanish MaltParser [5] for Spanish.

M-ATOLL performs three steps in order to find lexicalizations of ontology properties in the accompanying text corpus:

1. Retrieving all triples for a given property from the ontology. For example, the results for the property `spouse` include the triple `<Barack_Obama,spouse, Michelle_Obama>`.

[2] `https://github.com/ag-sc/matoll`

2. Retrieving all sentences from the parsed text corpus which contain mentions of the subject and object of the triples discovered in Step 1.
3. Searching for predefined patterns in those sentences, in order to extract candidate lexicalizations of the property.

So far, M-ATOLL covers entries that describe transitive verbs (e.g. *to cross*), intransitive verbs with a prepositional object (e.g. *to live in*), relational nouns with prepositional object (e.g. *capital of*), and relational adjectives (e.g. *similar to*) in all three languages: English, German, and Spanish. Important to note is that one entry can have multiple references, e.g. the relational noun entry *village in*[3] can refer to `hometown`, `birthplace`, and `location`, among others.

### 2.2 Label-based Approach

Examining the lexical entries created by the above approach and the kind of lexicalizations needed for question answering, for example, it is clear that M-ATOLL does not yet cover all relevant patterns. Consider the request *Give me all female Danish politicians*. In this case a lexical entry for *female* is needed, which w.r.t. DBpedia refers to all individuals that are related to the resource `Female` by means of the property `gender`. Similarly, *Danish* refers to all individuals that are related to the resource `Denmark` by means of the property `country` or `birthPlace`.

As these adjective lexicalizations are not handled by the standard corpus-based approach, we extended M-ATOLL with a dedicated module, extending [13]. This approach is based on the observation that a lot of those adjectives actually occur in the labels of the objects, e.g. *female* in `<_,gender,Female>` or *Catholic* in `<_,religion,Catholic_Church>`. We therefore check for adjectives occurring in object labels, training an SVM in order to decide whether those adjectives are valid laxicalizations of the restriction class in question.

Resulting entries are, for example, the above mentioned *female*[4], and *blue*[5]. For more information about this approach and the implemented features, see [12].

This approach currently works only for English, but will be soon adapted to German and Spanish.

## 3 Dataset

The dataset we present in this paper is the result of the above approaches and contains entries in three languages: English, German, and Spanish. Table 1 shows how many properties were lexicalized for each language. Overall, 574 different properties were lexicalized, but note that not every property is covered for each language. This is due to the fact that for some properties no sentences are found

[3] `http://dblexipedia.org/LexicalEntry_village_as_Noun_withPrep_in`
[4] `http://dblexipedia.org/LexicalEntry_female_as_AdjectiveRestriction`
[5] `http://dblexipedia.org/LexicalEntry_blue_as_AdjectiveRestriction`

in the text corpus that match the predefined lexicalization patterns. For English, 567 properties were lexicalized, 224 by the corpus-based approach and 445 by the label-based approach. For German and Spanish, 145 and 50 properties, respectively, were lexicalized using the corpus-based approach (the label-based approach is currently not implemented for these languages).

Table 2 shows the number of entries generated by the different approaches for the different languages. Overall our dataset contains 11,998 entries. It is important to mention that one lexical entry can lexicalize multiple properties.

| | Corpus-based | Label-based | Total |
|---|---|---|---|
| English | 224 | 445 | 567 |
| German | 145 | n.a. | 145 |
| Spanish | 50 | n.a. | 50 |

Table 1: Number of lexicalised properties per approach and language.

| | Corpus-based | Label-based | Total |
|---|---|---|---|
| English | 4456 | 4092 | 8548 |
| German | 3320 | n.a. | 3320 |
| Spanish | 130 | n.a. | 130 |

Table 2: Number of entries generated for each language for each approach.

Attached to the entries we also publish meta-data, in particular about provenance, specifying from which pattern an entry was created (for the corpus-based approach), with which frequency, and with which confidence (for the label-based approach). In the future we also intend to include example sentences for each entry (for the corpus-based approach) and the set of corresponding features which led to the entry (for the label-based approach).

Moreover, the lexical entries are linked to *dbnary*[6] and *lemon UBY*[7], considering the canonical form and the part of speech as relevant information for comparison.

The dataset is published at `http://dblexipedia.org`, which was generated by a modified version of *YUZU*[8]. The website enables a user to browse through the lexical entries and supports search over them. The whole dataset can also be downloaded at `http://dblexipedia.org/public/all.nt.gz` (as N-Triples). In

[6] `http://kaiko.getalp.org/about-dbnary/development/`
[7] `http://www.lemon-model.net/lexica/uby/`
[8] `https://github.com/jmccrae/yuzu`

addition to the dataset, the most current version of M-ATOLL is available at `http://dblexipedia.org/public/MATOLL.jar`.

We compared the published entries of the English corpus-based approach with those of the manually created DBpedia lexicon [9]. Out of the 224 properties lexicalised by M-ATOLL, 84 were also lexicalised in the manually created lexicon. We therefore evaluated on those 84 properties and found that 48 % of the automatically constructed entries were consistent with the corresponding manually created ones. This shows that on a larger number of properties, on average half of the generated entries are good, whereas the other half has to be manually corrected. For the example above with the property *spouse*, M-ATOLL creates the three lexical entries mentioned in the introduction. Additionally we find for this property the expressions *X is the widow of Y* and *X is reunited with Y*.

The main limitation of the presented dataset is the small coverage of the DBpedia 2014 ontology, containing around 2796 properties. However, for 1424 properties of this ontology, no data is available at the official DBpedia SPARQL endpoint. Considering only properties with at least one received data item, our very first release of the lexicon covers already 42% of the properties.

## 4 Conclusion

In this paper we presented the first multilingual, automatically generated lexicon for DBpedia, covering 574 properties from the DBpedia 2014 ontology. The evaluation showed that on average half of the generated entries are correct. In order to further improve the quality of the resulting lexica across languages, we will work on the adaptation of the label-based approach to German and Spanish, and in the future also Japanese. Moreover, we plan to evaluate to which extent an increase in coverage can improve tasks like question answering (see, e.g. [2]). We will also continue to publish updates of the dataset at `http://dblexipedia.org`, offering it to the community as a resource that can support natural language-based applications over the Semantic Web and that can serve as a hub for other lexical resources on the linguistic linked data cloud.

## Acknowledgment

This work was supported by the Cluster of Excellence Cognitive Interaction Technology *CITEC* (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG), and the FP7 European project LIDER (610782).

## References

1. Nadjet Bouayad-Agha, Gerard Casamayor, and Leo Wanner. Natural language generation in the context of the semantic web. *Semantic Web*, 5(6):493–513, 2014.

2. Sherzod Hakimov, Christina Unger, Sebastian Walter, and Philipp Cimiano. Applying semantic parsing to question answering over linked data: Addressing the lexical gap. In Chris Biemann, Siegfried Handschuh, André Freitas, Farid Meziane, and Elisabeth Métais, editors, *Natural Language Processing and Information Systems*, volume 9103 of *Lecture Notes in Computer Science*, pages 103–109. Springer International Publishing, 2015.
3. Jens Lehmann, Chris Bizer, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. DBpedia - a crystallization point for the web of data. *Journal of Web Semantics*, 7(3):154–165, 2009.
4. Vanessa Lopez, Victoria Uren, Marta Sabou, and Enrico Motta. Is question answering fit for the semantic web?: A survey. *Semantic Web*, 2(2):125–155, 2011.
5. Montserrat Marimon and Núria Bel. Dependency structure annotation in the IULA Spanish LSP Treebank. *Language Resources and Evaluation*, 49(2):433–454, 2015.
6. John McCrae, Dennis Spohr, and Philipp Cimiano. Linking lexical resources and ontologies on the semantic web with lemon. In *The Semantic Web: Research and Applications*, pages 245–259. Springer, 2011.
7. Joakim Nivre. An efficient algorithm for projective dependency parsing. In *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT*, pages 149–160, 2003.
8. Rico Sennrich. The UZH system combination system for WMT 2011. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, WMT '11, pages 166–170, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
9. Christina Unger, John McCrae, Sebastian Walter, Sara Winter, and Philipp Cimiano. A lemon lexicon for DBpedia. In *Proceedings of 1st International Workshop on NLP and DBpedia, co-located with the 12th International Semantic Web Conference (ISWC 2013), October 21-25, Sydney, Australia*, 2013.
10. Sebastian Walter, Christina Unger, and Philipp Cimiano. ATOLL – a framework for the automatic induction of ontology lexica. *Data & Knowledge Engineering*, 94, Part B(0):148–162, 2014. Special Issue following the 18th International Conference on Applications of Natural Language Processing to Information Systems (NLDB'13).
11. Sebastian Walter, Christina Unger, and Philipp Cimiano. M-ATOLL: a framework for the lexicalization of ontologies in multiple languages. In *The Semantic Web – ISWC 2014*, volume 8796 of *Lecture Notes in Computer Science*, pages 472–486. Springer International Publishing, 2014.
12. Sebastian Walter, Christina Unger, and Philipp Cimiano. Automatic acquisition of adjective lexicalizations of restriction classes: a machine learning approach. In *Journal on Data Semantics (to appear)*, 2015.
13. Sebastian Walter, Christina Unger, Philipp Cimiano, and Bettina Lanser. Automatic acquisition of adjective lexicalizations of restriction classes. In *Proceedings of 2st International Workshop on NLP and DBpedia, co-located with the 13th International Semantic Web Conference (ISWC 2014), October 19-23, Riva del Garda, Italy*, 2014.