# Linked Data for Language-Learning Applications

**Robyn Loughnane, Kate McCurdy, Peter Kolb, Stefan Selent**
Babbel (Lesson Nine GmbH)
{rloughnane, kmccurdy, pkolb, sselent}@babbel.com

## Abstract

The use of linked data within language-learning applications is an open research question. A research prototype is presented that applies linked-data principles to store linguistic annotation generated from language-learning content using a variety of NLP tools. The result is a database that links learning content, linguistic annotation and open-source resources, on top of which a diverse range of tools for language-learning applications can be built.

## 1 Introduction

Since Berners-Lee (2001) presented his vision of a Semantic Web at the turn of the century, there has been an explosion of technologies and tools made available to implement it[1]. The core idea of the Semantic Web is *linked data*, where data forms a giant graph spread across the internet, known as the Giant Global Graph or Web 3.0. In Berners-Lee's original vision, this linked data should be open source and the resulting graph is freely available over the internet. Of course, the same principles and technologies can be applied to create a private graph database used for commercial purposes, for applications like a social network or knowledge base.

Use of linked data in linguistics in general is a burgeoning research topic (Section 2). In this paper, linked-data technology is applied in the context of a language-learning application, in order to create a prototype database of linguistic annotation for learning content (Section 3). The database further links learning content and linguistic annotation with resources from the Linguistic Linked Open Data (LLOD) cloud and other open-source linguistic resources. The resulting database is flexible enough to allow a variety of useful applications for the language learner to be built on top of it.

Although NLP tools for creating linguistic annotation on the fly are becoming more and more accurate[2] and are adequate for many purposes, this prototype tests storage of linguistic annotation with the future aim of storing high-quality, curated linguistic annotation. This linguistic annotation, to be derived from a combination of various NLP tools and human expertise, could then be updated or expanded as new technology becomes available. The result would be a database of linguistic annotation that is more accurate than the output of any single tool and can be used for a variety of purposes related to language-learning applications.

There are already a number of approaches available for automatically generating exercises for language learning, such as using Google *n*-grams (Hill and Simha, 2016) or a mix of techniques including crowdsourcing, measuring WordNet distance, and machine learning (Kumar et al., 2015). Although it is the focus of the evaluation of the prototype (Section 4), automatic generation of exercises is only one possible use of the database discussed here. Linking between learning content, linguistic annotation and the LLOD cloud creates a resource that can be used for a variety of purposes, for example assessing the number of lemmas seen in exercises completed by a user up to a certain point in time, or showing the user grammatical information for a particular exercise.

[1] https://www.w3.org/standards/semanticweb/

[2] The state of the art in automatic syntax parsing reports models with an upper limit of around to 95% accuracy for certain types of input (Andor et al., 2016). For part-of-speech tagging, the state of the art is around 97%, depending on the type of input. Accuracy rates can be much lower for low-frequency tokens, out-of-context text, and data that differs significantly from the training set.

*Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 44–51
Copenhagen, Denmark, September 8, 2017.

## 2 Linked Data in Linguistics

Recently, applications of linked-data technology in the field of linguistics in general have been gaining in popularity, as witnessed by the large amount of resources in the LLOD cloud (Section 2.1) and the growing number of linguistic ontologies (Section 2.2). In addition to being able to link to the LLOD cloud, Semantic Web has the advantage of a native graph-based data model (Section 2.3), namely the Resource Description Framework[3] (RDF).

The use of linked-data technology in applications for language learning has, however, been limited, meaning that the potential of the LLOD cloud has yet be fully exploited in this area. A notable exception is El Maarouf et al. (2015), who created a multilingual network of linguistic resources by using sense linking to bridge the language gap with the goal of facilitating the creation of language-learning content.

### 2.1 LLOD

The LLOD cloud diagram[4] (McCrae et al., 2016; Chiarcos et al., 2012) shows that there is already a wealth of free and open-source linguistic linked data available to use. Major resources are each represented by a single node in the LLOD cloud diagram. These include DBpedia (Mendes et al., 2012), consisting of structured information extracted from Wikipedia; WordNet RDF (McCrae et al., 2014), an RDF translation of Princeton's WordNet lexical database project; and DBnary (Sérasset, 2015), derived from Wiktionary.

### 2.2 Ontologies

An ontology is a document that specifies the structure of a system through entities and relations (Guarino et al., 2009). Complex abstract models can be specified precisely via ontologies in the Web Ontology Language[5] (OWL). A variety of ontologies have been proposed to describe the components of language analysis, each developed with a different purpose in mind.

ISOcat (Windhouwer and Wright, 2012) and GOLD (Farrar and Langendoen, 2003) were created with the aim of covering a large range of linguistic terminological categories. Ontologies of Linguistic Annotation (OLiA), an intermediate level of representation between ISOcat and GOLD, addresses conceptual interoperability (Chiarcos, 2012; Chiarcos and Sukhareva, 2015).

POWLA (Chiarcos, 2012) represents any kind of linguistic annotation in a theory independent way. It is an adaptation of the PAULA XML exchange format (Zeldes et al., 2013).

Lemon (McCrae et al., 2012) is an ontology for exchanging lexical information on the Semantic Web. It is used, for example, in the DBnary project (Sérasset, 2015) and WordNet RDF (McCrae et al., 2014).

### 2.3 Linguistic Annotation as a Graph

Representing linguistic annotation as a graph has the advantage of avoiding undue influence from the data serialization format (e.g. XML) or the database type (e.g. relational). For example, Zipser (2009) describes how, when a format for exchanging linguistic annotation is specified without an abstract model being explicitly specified, it can lead to the format's implicit abstract model being influenced or limited by the data serialization format used. An example would be XML-based formats being influenced by the tree-based structure of XML to the extent that the implicit abstract model of the linguistic annotation format becomes tree based.

Semantic Web technology largely allows this problem to be avoided. RDF-based linguistic exchange formats are inherently graph based, so are only limited in structure to the extent that a labelled, directed multigraph is limited. Further, OWL is designed specifically for ontology specification, and allows complex models to be specified in a precise way. Although, of course, the XML syntax for RDF (Gandon and Schreiber, 2014) shows that a graph may be specified in the XML format, so the pitfall of influence from the data serialization format can also be avoided with clear specification of the abstract model independent of the data serialization format, e.g. in the Unified Modeling Language (UML).

The graph-based SALT model (Zipser and Romary, 2010) further shows that a graph structure preserves the abstract model for a wide range of linguistic annotation formats, including PAULA, ELAN, ANNIS and more.

Chiarcos (2012) likewise argues that a representation of linguistic annotation as a labelled, directed graph represented in OWL and RDF can

[3] https://www.w3.org/RDF/

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

[5] https://www.w3.org/OWL/

| Resource | Type |
|---|---|
| Stanford CoreNLP | Language analysis |
| FreeLing | Language analysis |
| WebLicht | Annotation framework |
| WordNet RDF | Lexical database |
| DBnary | Lexical database |
| Specialist lexicon | Lexical database |
| Lemon | Ontology |

Table 1: External Resources

solve interoperability issues and enables connection to the LLOD cloud.

Bird and Liberman (2001) also argued that it is of greatest importance to have a well-defined common conceptual framework and that the standardization of file formats is of secondary importance. They present an annotation graph as a common conceptual framework for a number of annotation formats.

## 3 Design of the Database

The starting point for the database was Babbel's learning content (Section 3.1). Linguistic annotation for the content was then created via NLP pipelines (Section 3.2). The learning content and its annotation was then converted to RDF and linked with LLOD resources and other open-source linguistic resources (Section 3.3). Table 1 summarizes the external dependencies.

### 3.1 Learning Content

Babbel is a language-learning application with over 1 million active subscribers and has been shown to be an effective way to learn a foreign language (Vesselinov and Grego, 2016). The language application is based on a large corpus of language exercises created by a team of didactic experts. There are a range of types of exercises, testing users' reading, writing, listening and speaking skills.

YAML files containing the exercises were used as the starting point for the database. Additionally, a variety of metadata for the learning content was available in an XML format.

### 3.2 Linguistic Annotation

Linguistic annotation was derived from NLP pipelines set up for each of the two learning languages, English and Spanish. These NLP pipelines used a combination of custom implementations and open-source tools, including Stanford CoreNLP (Manning et al., 2014) and FreeLing (Padró and Stanilovsky, 2012). As the pipelines are used for a variety of research purposes, the resulting linguistic annotation was stored in WebLicht's Text Corpus Format (TCF) (Heid et al., 2010) in XML files, rather than directly in RDF. The NLP pipeline produces the following layers: text, tokens, sentences, lemmas, part-of-speech tags, morphological features, and dependency parsing.

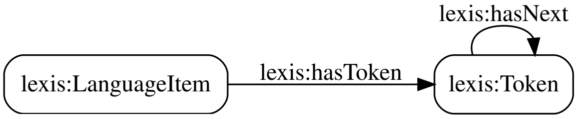

Figure 1: Lexis Language Item and Token

### 3.3 Linking the Data

The learning content and linguistic annotation were converted to RDF (Section 3.3.1) and then linked to existing LLOD resources (Section 3.3.2), and other open-source linguistic resources converted to RDF (Section 3.3.2).

#### 3.3.1 Linking Learning Content

Three ontologies were created with OWL to model the learning content from the three different sources: the Graph ontology for the XML metadata files; the Lesson ontology for the learning content YAML files; and the Lexis[6] ontology for the TCF XML files. A Java program was then created to convert the XML and YAML structures to RDF triples.

The Graph ontology models a variety of metadata, including the order of lessons within a learning module. The Lesson ontology models information within a lesson, like the parts of the language item that the user interacts with e.g. a gap in a sentence that the user fills in. Given that the learning content and metadata already had a well-defined underlying structure, a parallel structure was created in the Graph and Lesson ontologies.

The following OWL classes were defined within the Lexis ontology: `LanguageItem`, `Token`, `Dependency`, `Feature` and `Sense`. Figures 1 to 5 show the main OWL object property relations between the classes.

[6] From the Ancient Greek λέξης meaning 'word'

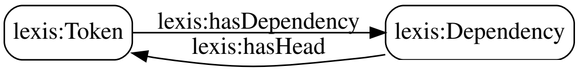

Figure 2: Lexis Token and Dependency Relation

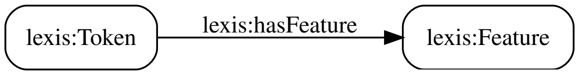

Figure 3: Lexis Token and Feature

Figure 1 shows that a second language text fragment, namely a `LanguageItem`, may have one or more entities of type `Token` related to it by the `hasToken` property. The `hasNext` property points to the next ordered `Token` for the `LanguageItem`. A number of OWL datatype property relations are further defined for `Token`, e.g. the text value of the token.

The property `hasDependency` (Figure 2) connects a `Token` and a `Dependency` according to the dependency relations specified by the Universal Dependencies project (Nivre, 2016). The head of a dependency relation is another token, indicated by the `hasHead` object property. Morphological features of tokens, including part of speech and grammatical gender, are assigned to the `Feature` class, related to a token via the object property `hasFeature` (Figure 3).

The Lexis ontology imports the Lemon ontology (Section 2.2), which is used to connect word senses of tokens to the corresponding WordNet entries (Figures 4 and 5). The lemma of a token is saved as a datatype property of the token's sense.

For the Lexis ontology, in addition to Lemon, it would have been possible to reuse other existing ontologies designed for representing linguistic annotation, like POWLA, GOLD or OLiA (Section 2.2). For this initial research prototype, however, the design decision was made to create a new, minimal ontology and the mapping of Lexis to other ontologies is left for future research.

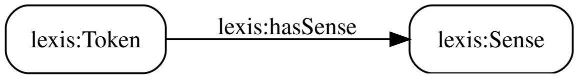

Figure 4: Lexis Token and Sense

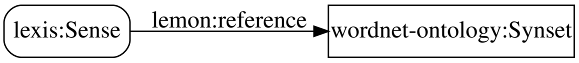

Figure 5: Lexis Sense and Lemon Reference

#### 3.3.2 Linking LLOD Resources

As mentioned above, the RDF version (McCrae et al., 2014) of WordNet (Miller, 1995) was used, connecting synsets to tokens via lexical sense (Figure 5). As an expedient initial assignment, the part of speech and lemma of a token were used to search for the corresponding WordNet synset with the highest frequency (tag count). Links to DBnary (Sérasset, 2015) were created in a similar way.

#### 3.3.3 Linking Other Linguistic Resources

The majority of open-source linguistic resources are currently not available as five-star linked open data according to Berners-Lee's (2006) definition. However, as long as the data is three star, then it can generally be meaningfully converted into linked data, usually with some manual work involved to create a mapping. Three-star data is available to use with an open licence; available as structured, machine-readable data; and available in a non-proprietary format (Berners-Lee, 2006). Indeed this is the source of many of the LLOD resources, like DBpedia, whose data were originally available in some other format. For the current research prototype, two main resources were converted to RDF, the Specialist lexicon[7] and the FreeLing Spanish dictionary[8]. These were then linked to the learning content in a similar way to the LLOD resources (Section 3.3.2).

The Specialist lexicon (Browne et al., 2000) is a large English lexicon developed within the Unified Medical Language System by the US National Library of Medicine (Bodenreider, 2004). The XML version of the lexicon was imported using the provided (but slightly adapted) XML format specification. A custom ontology was created in OWL that paralleled the underlying structure of the dictionary entries. A Java program was then written to convert the XML to RDF according to the ontology. The ontology and Java program have been made available as an open-source project[9].

The FreeLing Spanish dictionary entry files were converted into RDF triples according to the

[7] http://specialist.nlm.nih.gov/lexicon

[8] https://github.com/TALP-UPC/FreeLing

[9] https://github.com/babbel/specialist_rdf

Lemon ontology (McCrae et al., 2010).

### 3.4 Storing Linguistic Linked Data

With the recent rise in popularity of NoSQL databases, there are now a number of databases specifically designed for storing linked data as RDF triples, such as Ontotext's GraphDB (based on RDF4J, formerly Sesame) and Apache Jena Fuseki. The created and collected linguistic linked data described in Section 3.3 was stored in GraphDB.

## 4 Evaluation

A suite of example use cases were built on top of the database, serving as experimental evaluation. These use cases included a Spanish conjugation exercise (Section 4.1) and an English syntax display (Section 4.2). Apart from unit testing to assure the graph is produced as expected, the quality of the data produced was not evaluated. The quality of the linguistic annotation depends on the tools used to generate it, e.g. Stanford CoreNLP. The evaluation of the quality of the sense linking with WordNet and DBnary is left for further research.

### 4.1 Spanish Conjugation

A learning exercise for verb conjugation in Spanish was built on top of the existing learning content in the database[10]. Learning content for Spanish was searched for sentences in the present tense of the form subject–verb–direct object. Spanish verbs in the present tense have a different form depending on politeness (Helmbrecht, 2013) and the person and number of the subject. The verb was then replaced with its infinitive form and a drop-down menu showing all present tense verb forms for the same verb. The user is then asked to choose the correct form of the verb. For example, "Este piso tiene un jardín privado" becomes "Este piso tener un jardín privado", with a drop-down menu for "tener" displaying all the present tense forms of the verb. If the user selects the incorrect verb form from the drop-down menu, a message is displayed and they may try again. If the user selects the correct verb form from the drop-down menu, the exercise is complete.

### 4.2 English Syntax

A further use case was built on top of the database for selecting English language items containing auxiliary verbs. The SPARQL request shown in Listing 1 selects English language items that have a dependency relation where one verb acts as an auxiliary to another verb. This query returns URIs for languages items such as "Which pants should I buy?", where 'should' is the auxiliary verb and 'buy' is the main verb. A further SPARQL query retrieves the tokenization for this language item, enabling the auxiliary verb and main verb to be identified and highlighted for the user in the GUI. Such a use case could be extended to any other syntactic construction, so that the user could revise the construction in question, e.g. by highlighting the correct verb types.

Listing 1: SPARQL Query

```
PREFIX lexis: <http://www.babbel.com/lexis#>
PREFIX lesson: <http://www.babbel.com/lesson#>
SELECT DISTINCT ?subject
WHERE {
  ?subject a lexis:LanguageItem .
  ?subject lesson:alpha3 ‘eng’ .
  ?subject lexis:hasToken ?token .
  ?token lexis:hasDependency ?dependency .
  ?dependency lexis:dependencyFunction ‘aux’ .
  ?dependency lexis:hasHead ?head .
  ?head lexis:hasFeature ?feature .
  ?feature lexis:featureValue ?pos .
  ?feature lexis:featureName ‘pos’ .
  FILTER regex(?pos, ‘^V’)
} LIMIT 50
```

### 4.3 Performance

The technology for RDF triple stores is not as mature as for relational databases and this is reflected in their performance as witnessed by the so-called "RDF tax", although recent work has been done to improve this (Boncz et al., 2014). Performance for this prototype was also affected by the quality of the data contained in the database and the type of query performed. When the linguistic annotation saved in the database is clean and precise, the SPARQL query can be simpler and get the desired result faster.

The SPARQL query in Listing 1 sent via cURL took 0.035 seconds on average when run 100 times in a row on a MacBook Pro with 8GB RAM. The database stops searching and replies as soon as it has found 50 items that fulfill the request.

[10] The authors thank Raphaela Wrede, Pierpaolo Frasa, Katharina Schoppa and Simon Kreiser for their help in testing a prototype of this idea.

The SPARQL query in Section 4.1, however, took around seven seconds when executed in the GraphDB SPARQL GUI. This is not unexpected as the query searches through every single item in the database. A large number of complicated conditions were further required in the query, as the NLP tool did not distinguish between certain types of objects. For example, temporal phrases and direct objects were coded the same, so these had to be manually added as conditions to the SPARQL query, so as not to be included in the end result.

## 5 Conclusion and Further Work

The prototype database presented here combines RDF resources created from Babbel's learning content with linguistic annotation and existing resources from the LLOD cloud and elsewhere. The concept of the database was validated by experimental evaluation in the form of use cases built on top of it (Section 4).

In the first prototype, the minimal Lexis ontology was designed to test the concept. In future iterations, more work on this ontology could take place, including identification of areas where ontology design patterns (Blomqvist et al., 2016) could be used; and mapping to existing ontologies for linguistic annotation (Section 2.2). Likewise, work on conceptual (semantic) interoperability could take place, using ISOcat categories or similar, to enable use cases that incorporate linguistic annotation across more than one language, and to enable more use of external LLOD resources.

Future iterations could also incorporate improved word sense disambiguation techniques based on supervised machine learning (Navigli, 2009). Alternatively, the availability of translations of the learning content into multiple languages could be exploited to infer the correct mapping (Tufiş et al., 2004).

As seen in Section 4.1, query performance time suffers, when the query becomes too complex due to errors in the linguistic annotation or underspecification in annotation categories. Improving the quality of the linguistic annotation, either by swapping out a given NLP tool, or using a combination of multiple NLP tools and manual review, would further improve the efficiency and usefulness of the database. As the second-language text fragments generally do not have any context, manual review will likely always be necessary.

Future work could also be done on database performance in general, for example by exploring the use of the compact Header, Dictionary and Triples structure for storing RDF (Fernández et al., 2010).

## References

Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. 2016. Globally normalized transition-based neural networks. arXiv:1603.06042 .

Tim Berners-Lee. 2006. Linked data. http://www.w3.org/DesignIssues/LinkedData.html.

Tim Berners-Lee, James Hendler, and Ora Lassila. 2001. The Semantic Web. Scientific American .

Steven Bird and Mark Liberman. 2001. A formal framework for linguistic annotation. Speech Communication 33(1):23–60.

Eva Blomqvist, Pascal Hitzler, Krzysztof Janowicz, Adila Krisnadhi, Tom Narock, and Monika Solanki. 2016. Considerations regarding ontology design patterns. Semantic Web 7(1):1–7.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): Integrating biomedical terminology. Nucleic Acids Research 32(1):D267–D270.

Peter Boncz, Orri Erling, and Minh-Duc Pham. 2014. Advances in large-scale RDF data management. In Sören Auer, Volha Bryl, and Sebastian Tramp, editors, Linked Open Data – Creating Knowledge Out of Interlinked Data, Springer, pages 21–44.

Allen C Browne, Alexa T McCray, and Suresh Srinivasan. 2000. The Specialist lexicon. Technical report, National Library of Medicine.

Christian Chiarcos. 2012. Interoperability of corpora and annotations. In Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann, editors, Linked Data in Linguistics, Springer-Verlag, Berlin, pages 161–179.

Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff. 2012. Linking linguistic resources: Examples from the Open Linguistics Working Group. In Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann, editors, Linked Data in Linguistics, Springer-Verlag, Berlin, pages 201–216.

Christian Chiarcos and Maria Sukhareva. 2015. OLiA – Ontologies of Linguistic Annotation. Semantic Web 6(4):379–386.

Ismail El Maarouf, Hatem Mousselly Sergieh, Eugene Alferov, Haofen Wang, Zhijia Fang, and Doug

Cooper. 2015. The GuanXi network: A new multilingual LLOD for language learning applications. In Second Workshop on Natural Language Processing and Linked Open Data. Hissar, Bulgaria, pages 42–51.

Scott Farrar and D Terence Langendoen. 2003. A linguistic ontology for the Semantic Web. GLOT International 7(3):97–100.

Javier D. Fernández, Miguel A. Martínez-Prieto, and Claudio Gutierrez. 2010. Compact representation of large RDF data sets for publishing and exchange. In Peter F. Patel-Schneider, Yue Pan, Pascal Hitzler, Peter Mika, Lei Zhang, Jeff Z. Pan, Ian Horrocks, and Birte Glimm, editors, The Semantic Web – ISWC 2010, Springer-Verlag, Berlin, pages 193–208.

Fabien Gandon and Guus Schreiber. 2014. RDF 1.1 XML syntax. W3C recommendation, W3C.

Nicola Guarino, Daniel Oberle, and Steffen Staab. 2009. What is an ontology? In Handbook on Ontologies, Springer-Verlag, Berlin, pages 1–17.

Ulrich Heid, Helmut Schmid, Kerstin Eckart, and Erhard Hinrichs. 2010. A corpus representation format for linguistic web services: The D-SPIN Text Corpus Format and its relationship with ISO standards. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, LREC 2010, Seventh International Conference on Language Resources and Evaluation. Valletta, Malta.

Johannes Helmbrecht. 2013. Politeness distinctions in pronouns. In Matthew S. Dryer and Martin Haspelmath, editors, The World Atlas of Language Structures Online, Max Planck Institute for Evolutionary Anthropology, Leipzig. http://wals.info/chapter/45.

Jennifer Hill and Rahul Simha. 2016. Automatic generation of context-based fill-in-the-blank exercises using co-occurrence likelihoods and Google *n*-grams. In Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications. San Diego, California, USA, pages 23–30.

Girish Kumar, Rafael E. Banchs, and Luis F. D'Haro. 2015. RevUP: Automatic gap-fill question generation from educational texts. In The Tenth Workshop on Innovative Use of NLP for Building Educational Applications. Denver, Colorado, USA, pages 154–161.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pages 55–60.

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. 2012. Interchanging lexical resources on the Semantic Web. Language Resources and Evaluation 46(4):701–719.

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. 2010. The lemon cookbook. http://lemon-model.net/lemon-cookbook.pdf.

John Philip McCrae, Christian Chiarcos, Francis Bond, Philipp Cimiano, Thierry Declerck, Gerard de Melo, Jorge Gracia, Sebastian Hellmann, Bettina Klimek, Steven Moran, Petya Osenova, Antonio Pareja-Lora, and Jonathan Pool. 2016. The Open Linguistics Working Group: Developing the Linguistic Linked Open Data cloud. In LREC 2016, Tenth International Conference on Language Resources and Evaluation. Portorož, Slovenia, pages 2435–2441.

John Philip McCrae, Christiane Fellbaum, and Philipp Cimiano. 2014. Publishing and linking WordNet using lemon and RDF. In 3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing. Reykjavík, Iceland.

Pablo Mendes, Max Jakob, and Christian Bizer. 2012. DBpedia: A multilingual cross-domain knowledge base. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, LREC 2012, Eighth International Conference on Language Resources and Evaluation. Istanbul, Turkey.

George A Miller. 1995. WordNet: A lexical database for English. Communications of the ACM 38(11):39–41.

Roberto Navigli. 2009. Word sense disambiguation: A survey. ACM Computing Surveys 41(2):10.

Joakim Nivre. 2016. Universal Dependencies: A cross-linguistic perspective on grammar and lexicon. In Eva Hajičová and Igor Boguslavsky, editors, Grammar and Lexicon: Interactions and Interfaces. Osaka, Japan, pages 38–40.

Llus Padró and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards wider multilinguality. In LREC 2012, Eighth International Conference on Language Resources and Evaluation. Istanbul, Turkey, pages 2473–2479.

Gilles Sérasset. 2015. DBnary: Wiktionary as a lemon-based multilingual lexical resource in RDF. Semantic Web 6(4):355–361.

Dan Tufiş, Radu Ion, and Nancy Ide. 2004. Fine-grained word sense disambiguation based on parallel corpora, word alignment, word clustering and aligned wordnets. In COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics. Geneva, Switzerland, 1192, pages 1312–1318.

Roumen Vesselinov and John Grego. 2016. The Babbel efficacy study. http://press.babbel.com/en/releases/downloads/Babbel-Efficacy-Study.pdf.

Menzo Windhouwer and Sue Ellen Wright. 2012. Linking to linguistic data categories in ISOcat. In Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann, editors, Linked Data in Linguistics, Springer-Verlag, Berlin, pages 99–107.

Amir Zeldes, Florian Zipser, and Arne Neumann. 2013. PAULA XML documentation: Format version 1.1. Technical report, University of Potsdam.

Florian Zipser. 2009. Entwicklung eines Konverterframeworks für linguistisch annotierte Daten auf Basis eines gemeinsamen (Meta-)modells. Diplomarbeit, Humboldt-Universität zu Berlin.

Florian Zipser and Laurent Romary. 2010. A model oriented approach to the mapping of annotation formats using standards. In Language Resource and Language Technology Standards  State of the Art, Emerging Needs, and Future Developments, LREC 2010. Valletta, Malta.