Chapter 2
Towards Open Data for Linguistics: Linguistic Linked Data

Christian Chiarcos, John McCrae, Philipp Cimiano, and Christiane Fellbaum

Abstract 'Open Data' has become very important in a wide range of fields. However for linguistics, much data is still published in proprietary, closed formats and is not made available on the web. We propose the use of linked data principles to enable language resources to be published and interlinked openly on the web, and we describe the application of this paradigm to the modeling of two resources, WordNet and the MASC corpus. Here, WordNet and the MASC corpus serve as representative examples for two major classes of linguistic resources, lexical-semantic resources and annotated corpora, respectively.

Furthermore, we argue that modeling and publishing language resources as linked data offers crucial advantages as compared to existing formalisms. In particular, it is explained how this can enhance the interoperability and the integration of linguistic resources. Further benefits of this approach include unambiguous identifiability of elements of linguistic description, the creation of dynamic, but unambiguous links between different resources, the possibility to query across distributed resources, and the availability of a mature technological infrastructure. Finally, recent community activities are described.

C. Chiarcos (✉)
Information Sciences Institute, University of Southern California, Marina del Rey, CA, USA
e-mail: chiarcos@isi.edu

J. McCrae · P. Cimiano
Semantic Computing Group, Cognitive Interaction Technology Center of Excellence (CITEC), University of Bielefeld, Bielefeld, Germany
e-mail: jmccrae@cit-ec.uni-bielefeld.de; cimiano@cit-ec.uni-bielefeld.de

C. Fellbaum
Computer Science Department, Princeton University, Princeton, NJ, USA
e-mail: fellbaum@princeton.edu

2.1 Motivation and Overview

Language is arguably one of the most complex forms of human behaviour, and accordingly, its investigation involves a broad width of formalisms and resources used to analyze, to process and to generate natural language. An important challenge is to store, to connect and to exploit the wealth of language data assembled in half a century of computational linguistics research. The key issue is the **interoperability** of language resources, a problem that is at best partially solved [25]. Closely related to this is the challenge of **information integration**, i.e., how information from different sources can be retrieved and combined in an efficient way.

As a principal solution, Tim Berners-Lee – the founder of the World Wide Web – proposed the so called *linked data principles* to publish open data on the Web. These principles represent rules of best practice that should be followed when publishing data on the Web [4]:

1. Use URIs as (unique) names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using Web standards such as RDF, and SPARQL.
4. Include links to other URIs, so that they can discover more things.

We argue that applying the linked data principles to lexical and other linguistic resources has a number of advantages and represents an effective approach to publishing language resources as open data. The first principle means that we assign a unique identifier (URI) to every element of a resource, i.e., each entry in a lexicon, each document in a corpus, every token in a corpus as well as to each data category that we use for annotation purposes. The benefit is that this makes the above mentioned resources uniquely and globally identifiable in an unambiguous fashion. The second principle entails that any agent wishing to obtain information about the resource can contact the corresponding web server and retrieve this information using a well-established protocol (HTTP) that also supports different 'views' on the same resource. That is, computer agents might request a machine readable format, while web browsers might request a human-readable and browseable view of this information as HTML. The third principle requires the use of standardized, and thus, inter-operable data models for representing (RDF, [29]) and querying linked data (SPARQL, [35]). The fourth principle fosters the creation of a network of language resources where equivalent senses are linked across different lexical-semantic resources, annotations are linked to their corresponding data categories in data category repositories, etc.

In the definition of linked data, the **Resource Description Framework (RDF)** receives special attention. RDF was originally designed as a language to provide metadata about resources that are available both offline (e.g., books in a library) and online (e.g., eBooks in a store). RDF provides a data model that is based on labelled directed (multi-)graphs, which can be serialized in different formats, where

Table 2.1 Selected relations from existing RDF vocabularies and possible fields of application

Domain	Example	Reference
Meta data	creator	Dublin core meta data categories
General relations between resources	sameAs	Web ontology language (OWL)
Concept hierarchies	subClassOf	RDF schema
Relations between vocabularies	broader	Simple knowledge organization scheme
Linguistic annotation	lemma	NLP interchange format

the nodes identified by URIs are referred to as 'resources'.[1] On this basis, RDF represents information in terms of *triples* – a *property* (relation, in graph-theoretical terms a labelled edge) that connects a *subject* (a resource, in graph-theoretical terms a labelled node) with its *object* (another resource, or a literal, e.g., a string). Every RDF resource and every property is uniquely identified by a URI. They are thus globally unambiguous in the web of data. This allows resources hosted at different locations to refer to each other, and thereby to create a network of data collections.

A number of RDF-based vocabularies are already available, and many of them can be directly applied to linguistic resources. A few examples are given in Table 2.1. In this way, the RDF specification provides only elementary data structures, whereas the actual *vocabularies* and domain-specific *semantics* need to be defined independently. For reasons of interoperability, existing vocabularies should be re-used whenever possible, but if a novel type of resource requires a new set of properties, RDF also provides the means to introduce new relations, etc.

RDF has been applied for various purposes beyond its original field of application. In particular, it evolved into a generic format for data exchange on the Web. It was readily adapted by disciplines as diverse as biomedicine and bibliography, and eventually it became one of the building stones of the Semantic Web. Due to its application across discipline boundaries, RDF is maintained by a large and active community of users and developers, and it comes with a rich infrastructure of APIs, tools, databases, and query languages. Further, RDF vocabularies do not only define the labels that should be used to represent RDF data, but they also can introduce additional constraints to formalize specialized RDF sub-languages. For example, the **Web Ontology Language (OWL)** defines the data types necessary for the representation of ontologies as an extension of RDF, i.e., *classes* (concepts), *instances* (individuals) and *properties* (relations).

In the remainder of this chapter, we explore the benefits of linked data, considering in particular the following advantages:

Representation and modelling Lexical-semantic resources can be described as labelled directed graphs (feature structures, [27]), as can annotated corpora [3].

---

[1]The term 'resource' is ambiguous here. As understood in this chapter, resources are structured collections of data which can be represented, for example, in RDF. Hence, we prefer the terms 'node' or 'concept' whenever *RDF resources* are meant.

RDF is based on labelled directed graphs and thus particularly well-suited for modelling both types of language resources.

Structural interoperability Using a common data model eases the integration of different resources. In particular, merging multiple RDF documents yields another valid RDF document, while this is not necessarily the case for other formats.

Federation In contrast to traditional methods, where it may be difficult to query across even multiple parts of the same resource, linked data allows for federated querying across multiple, distributed databases maintained by different data providers.

Ecosystem Linked data is supported by a community of developers in other fields beyond linguistics, and the ability to build on a broad range of existing tools and systems is clearly an advantage.

Expressivity Semantic Web languages (OWL in particular) support the definition of axioms that allow to constrain the usage of the vocabulary, thus introducing formal data types and the possibility of checking a lexicon or an annotated corpus for consistency.

Conceptual interoperability The linked data principles have the potential to make the interoperability problem less severe in that globally unique identifiers for concepts or categories can be used to define the vocabulary that we use and these URIs can be used by many parties who have the same interpretation of the concept. Furthermore, linking by OWL axioms allows us to define the exact relation between two different concepts beyond simple equivalence statements.

Dynamic import URIs can be used to refer to external resources such that one can thus import other linguistic resources "dynamically". By using URIs to point to external content, the URIs can be resolved when needed in order to integrate the most recent version of the dynamically imported resources.

We elaborate further on these aspects in this chapter. It is structured as follows: Sect. 2.2 describes the modelling of linguistic resources as linked data and identifies deficits and prospective advantages of using linked data for linguistic resources. Section 2.3 elaborates some of the benefits of this representation. Section 2.4 summarizes recent community activities promoting the publication of language resources as linked data.

2.2 Modelling Linguistic Resources as Linked Data

We consider two important classes of language resources, the first of which is **lexical-semantic resources**, i.e., resources that provide information about lexemes and their relation to other lexemes (e.g., machine-readable dictionaries, semantic networks, semantic knowledge bases, ontologies and terminologies). The second class of language resources considered here are **annotated corpora**, i.e., collections of textual (spoken, written or gestural) data annotated with linguistic characteristics.

For both types of resources, we describe state-of-the-art approaches, briefly motivate the application of linked data principles, and then describe modelling these resources using RDF and OWL.

Resource modelling involves two aspects: (1) the specification of data structures and consistency constraints over these, and (2) the conversion of data into these representations. RDF encodes labelled directed graphs and is thus capable to represent both lexical-semantic resources and linguistic corpora, as both can be described with directed graphs. For reasons of symmetry, also different types of annotated corpora are enumerated.

Unlike other graph-based modelling formalisms applied to language resources, e.g., GraphML [5], RDF provides additional means to formalize specific data types, and thereby to establish a **reserved vocabulary** and to introduce **structural constraints** for nodes, edges or labels. Such constraints are necessary, e.g., for corpora, to avoid confusion between RDF representations of corpus infrastructure (corpus, subcorpus, document, annotation layer) and meta data (information about the resource as a whole).

As an illustration of the benefits of modelling linguistic data as linked data, let us consider the following example. Imagine we would like to get all occurrences in a corpus (e.g. MASC, Sect. 2.2.2) of synonyms of 'land' in the sense of *(the territory occupied by a nation)'* (in WordNet 3.1, Sect. 2.2.1) with synonyms 'country' and 'state'. In order to get such occurrences, one would first use the WordNet data model – suitably abstracted by some API – and query for elements in the synset corresponding to 'land' as *(the territory occupied by a nation)'*. This 'query' would yield: 'land', 'country' and 'nation'. Then, using another data model and appropriate APIs or query interfaces, we would then search for occurrences of 'land', 'country' or 'nation' in the MASC corpus annotated with the corresponding sense ID key from WordNet. This shows that it is cumbersome and difficult to answer such queries which span multiple resources as one is forced to use different data models, APIs etc.

The benefit of using RDF and linked data principles to model linguistic resources is that it provides a graph-based model that allows representing different types of linguistic resources (corpora, treebanks, lexical-semantic resources) in a uniform way, thus supporting uniform querying across resources. The query sketched above, for example, can be represented in a single, and simple SPARQL expression as shown in Sect. 2.3.1.2.[2] And as RDF and SPARQL employ URIs to designate elements, it is even possible to query data not stored in a single repository, but that are accessible through different SPARQL endpoints. With a mechanism that can distribute the relevant parts of a query to the repositories that contain the relevant MASC and WordNet data (Sect. 2.3.2), answering such a query is indeed straightforward.

---

[2]We provide a SPARQL endpoint under http://monnetproject.deri.ie/lemonsource_query, which provides access to the examples discussed in this chapter.

In the following we discuss in more detail how both corpora (such as MASC) and lexical-semantic resources (such as WordNet) can be modelled using RDF and what the particular advantages are.

2.2.1 Modelling Lexical-Semantic Resources: WordNet

2.2.1.1 WordNet Data Structures

WordNet [17, 34] is a particularly influential lexical-semantic resource, and very prototypical in many aspects. It is a manually constructed electronic lexical resource, organized around concepts and the words expressing them. WordNet draws its motivation from theories of human lexical memory, which indicate that people store knowledge about concepts in a well-structured, economic fashion and attempts to implement this model. The current version 3.1 includes over 117,000 concepts expressed by nouns, verbs, adjectives, and adverbs.[3]

A concept in WordNet is represented as a set of (roughly) synonymous words that all refer to the same entity, event, or property. Synset members can be interchanged without altering the truth value of a context. Formally, WordNet is a directed acyclic graph, where synsets are interlinked by edges standing for means of conceptual-semantic relations. The most important is the super-/subordinate (hyponymy) relation. It links generic to increasingly specific synsets like *land* to *kingdom* and *sultanate*. Synset pairs referring to part-whole concepts (*land-midland*, *wheel-car*, etc.) are also connected, as are synsets expressing semantic opposition (*hot-cold*, *arrive-leave*, etc.) and a range of temporal relations (see [17]).

2.2.1.2 Generic Data Structures: Lexical Markup Framework

To facilitate interoperability among lexical-semantic resources, feature structures (i.e., directed acyclic graphs) have been suggested as a generalization over resource-specific data structures [40]. Feature structures are a flexible and general formalism, which became the basis for subsequent standardization, in particular, in the Lexical Markup Framework (LMF, [19]). LMF represents a metamodel to represent semantic information in NLP lexicons and machine-readable dictionaries. It has been successfully applied to develop resources such as Uby [22], an openly available, large-scale lexical-semantic resource. Uby integrates nine independent resources for English and German, including WordNet, Wiktionary, Wikipedia, FrameNet, VerbNet, and OmegaWiki, which are linked with each other on sense level. However, the LMF format is not an open format (in the sense that its specification is not freely available), and in its standard serialization as XML, it does not consider

---

[3]http://www.wordnet.princeton.edu

how resources can be uniquely identified on the web. Furthermore, according to the experience of Uby, application of the format requires domain-specific modifications to the standard schema.

An RDF formalization tackles some of these problems, and this has been suggested by the LMF developers themselves.[4] Providing lexical-semantic resources as linked data actually allows us to integrate LMF resources with other resources previously converted to RDF, e.g., in the context of the developing Semantic Web.

2.2.1.3   From LMF to RDF: *lemon*

Independently from LMF, there has already been some work towards the integration of WordNet with the Semantic Web, notably [39], who provided a simple mapping from WordNet to RDF, and augmented it with OWL semantics so that reasoning could be applied to the structure of the resource. However the format chosen for this resource was specific to the underlying data model of WordNet. For this reason, [33] propose the interchange model *lemon* (Lexicon Model for Ontologies) that supports publishing lexical-semantic resources as linked data on the basis of the following principles:

LMF-based (to allow easy conversion from non-linked data resources);
RDF-native (publishing as linked data, with RDFS and OWL used to describe the semantics of the model);
Modular (separation of lexicon and ontology layers, so that *lemon* lexica can be linked to existing ontologies in the linked data cloud);
Externally defined data categories (linking to data categories in annotation terminology repositories, rather than being limited to a specific part-of-speech tag set);
Principle of least power (the smaller the model and the less expressive the language, the wider its adoption and the higher the reusability of the data, [38]).

This model is illustrated in Fig. 2.1. *lemon* has been used as a basis for integrating the data of the English Wiktionary,[5] a (human-readable) dictionary created along 'wiki' principles, with the RDF version of WordNet [33]. As *lemon* derives from LMF but integrates with the existing Semantic Web formalisms, there was some need to adapt the data model. It was found that WordNet's model was fairly close to *lemon* and LMF, with only minor differences in the modelling of inflectional variants of lexical entries. However, the semantic modelling was more significantly different as *lemon* uses OWL to represent semantics.

---

[4]http://www.tagmatica.fr/lmf/LMF_revision_14_In_OWL29october2007.xml

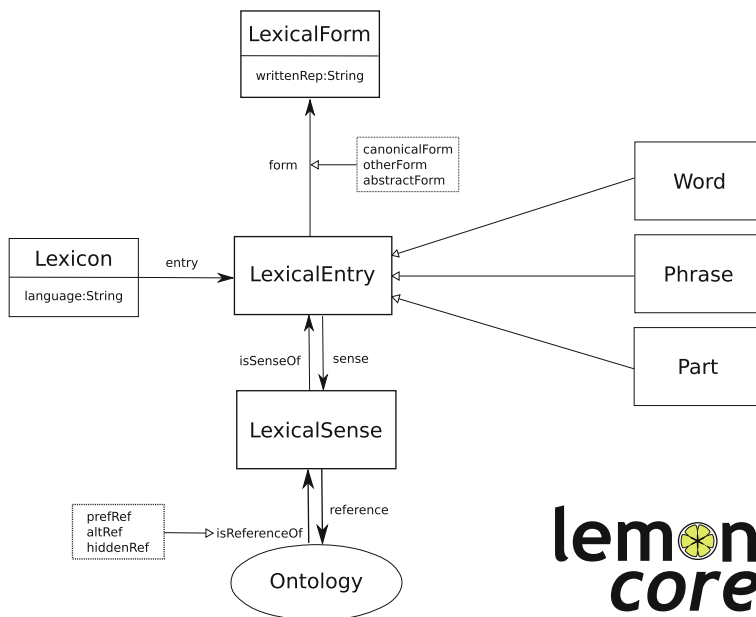[5]http://en.wiktionary.org/

Fig. 2.1 The core of the *lemon* model

2.2.2 Modelling Annotated Corpora: MASC

2.2.2.1 The Manually Annotated Sub-Corpus

The Manually Annotated Sub-Corpus (MASC, [28]) is a corpus of 500,000 tokens of contemporary American English text drawn from the Open American National Corpus, written and spoken, and chosen from a variety of genres.[6] MASC comprises various layers of annotations, including parts-of-speech, nominal and verbal chunks, constituent syntax, annotations of WordNet senses, frame-semantic annotations, coreference, document structure and illocutionary structure. The tools that generated the annotations of the MASC corpus use different output formats. In order to establish interoperability between them, MASC distributions adopt a generic data model, the Graph Annotation Format (GrAF, [26]). By use of multi-layer annotations, MASC allows all annotations of a particular piece of text to be integrated into a common representation that provides lossless and comfortable access to their linguistic information.
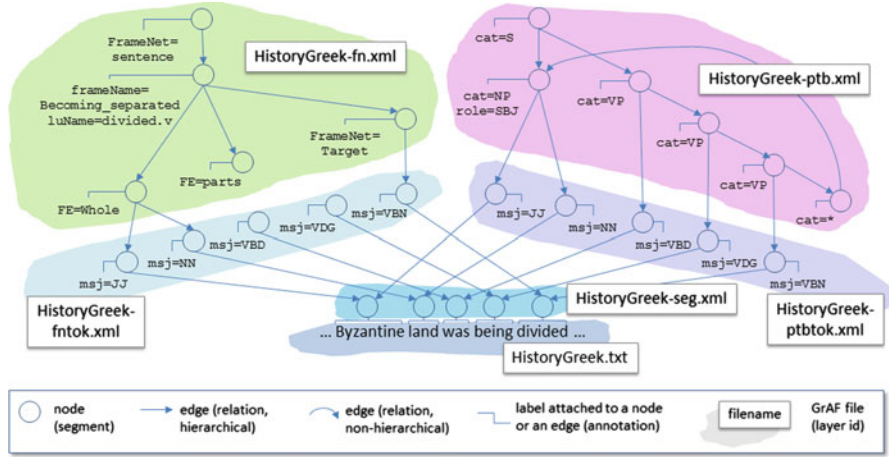
---

[6]www.anc.org/MASC

Fig. 2.2 Representing and integrating annotations for syntax and frame-semantics in a directed graph

2.2.2.2 Generic Data Structures for Annotated Corpora: GrAF

State-of-the-art approaches on interoperable formats for annotated corpora are based on the assumption that all linguistic annotations can be represented by means of labelled directed acyclic graphs [3]. To a certain extent, this echoes the application of feature structures to lexical-semantic resources (feature structures are labelled directed acyclic graphs).

One representative example for graph-based generic formats is the GrAF format. Like other state-of-the-art approaches that implement graph-based data models for linguistic corpora [7, 11], GrAF is a special-purpose XML standoff format. Standoff formats are based on a physical separation between primary data (e.g., text, audio or video) and different layers of annotations. In Fig. 2.2, this is shown for an example sentence from the MASC corpus: All annotations of a document are grouped together in a set of XML files pointing to the same piece of primary data. Different file names in the figure represent the respective annotation layer. Distributing annotations across different files, however, results in a highly complex structure with multiple dependencies between individual files. Consequently, standoff formats introduce a relatively large technical overhead that makes it difficult to work with large data in practice. While standoff formats have become widely accepted, the efficient processing, storing and retrieval of standoff data requires formalisms that support the free linking of elements, and that are thus fundamentally different from hierarchical data models such as XML that are optimized for tree structures, rather than general graphs.

Figure 2.2 shows the graph-based modelling and its XML standoff serialization for two selected layers of annotations for the clause *'Byzantine land was being divided'*. To the left, the figure shows FrameNet annotations [2] and to the right

PennTreebank-style syntax annotations [30]. Both annotations are synchronized with each other and the primary data through a shared base segmentation file.

2.2.2.3 From Standoff XML to RDF: POWLA

Standoff XML can be hard to process, and the corresponding infrastructures and standards are still under development. RDF, however, already provides a rich technological ecosystem for labelled directed graphs, and GrAF data structures can be easily converted to RDF. Rendering generic data models for annotated corpora in RDF has been suggested before, e.g., by Cassidy [8] and Chiarcos [10].

Chiarcos [10] described POWLA, an RDF/OWL linearization of PAULA, a generic data model for the representation of annotated corpora [14, 15]. PAULA is similar in scope and design to GrAF and also builds on traditional standoff XML. POWLA consists of two basic components: (1) an OWL/DL ontology that defines the valid data types, relations and constraints as classes, properties and axioms; (2) an RDF document that represents a corpus as a knowledge base consisting of individuals, instantiated object properties and data values assigned to individuals through datatype properties. POWLA formalizes the structure of annotated corpora and linguistic annotations of textual data. With respect to the latter, it provides data types such as `Node` and `Relation` (as well as more specialized data types) that directly reflect the underlying graph-based data model. With OWL/DL axioms, the relationship between these data types can be formalized and automatically verified, e.g., that `Relation` and `Node` are disjoint, and that every `Relation` is connected by one `hasSource` and one `hasTarget` property with a particular `Node`.

A GrAF converter is provided under http://purl.org/powla, it replicates the structure of the GrAF file exactly in RDF/OWL. As with the original GrAF representation, annotated corpora represented in this way are structurally interoperable (different annotations use the same representation formalism), but in this form, they can be queried using RDF query languages like SPARQL, they can be stored in RDF databases, and OWL/DL reasoners can be applied to validate the consistency of the data.

2.3 Benefits of Linked Data for Linguistics

Aside from representation, Sect. 2.1 identified five specific advantages of modelling linguistic resources as linked data. These include structural interoperability (same format for different types of resources), the querying of physically distributed resources (federation), enhanced conceptual interoperability (same vocabulary for different resources), a rich ecosystem of formalisms and technologies, and the possibility to create resolvable links between resources that are maintained by different data providers (dynamic import).

2.3.1 Structural Interoperability

Structural ('syntactic') interoperability of a language resource in NLP corresponds to the 'ability [of an NLP tool] to process it immediately without modification to its physical format', i.e., structural interoperability 'relies on specified data formats, communication protocols, and the like to ensure communication and data exchange' [25]. This involves two aspects: The capability to **provide access to the data** depending on the needs of the data consumer (a human user or some software tools), and the use of the **same format** for different resources such that they can be processed in a uniform way. To this definition of structural interoperability we should add another desideratum that partially follows from both aspects, namely that different resources are accessible with uniform query languages, and that information from different sources can be easily **merged**.

2.3.1.1 Structural Interoperability by Content Negotiation

Servers that publish data on the web can (and should) provide multiple versions of the data. This is possible as the HTTP protocol supports **content negotiation** [18, p. 67–70], i.e., a user or agent that accesses a particular resource can specify the format they want by means of the HTTP `Accept` header. This allows a lexical resource to be identified by a single URI, but display human-readable HTML to users accessing the page through a web browser and the original RDF data to web agents. Upon accessing a resource URI, the server responds with the first specified data format given by the user or an error if no acceptable format can be rendered. In this way, language resources can be published on the web using Semantic Web standards, human readable forms and other serializations.

A similar method called *transparent content negotiation* [24] allows the RDF and HTML versions of the page to be identified by a separate URI to the resource itself. Here instead of responding with the correct data type, the server redirects the client to a new URL for the appropriate data format. or example, the server may direct the client to add the suffix `.rdf` for the linked data and `.html` for the human-readable version.

2.3.1.2 RDF as a Structurally Interoperable Format

We have seen that RDF is suitable for representing two major types of linguistic resources, and thus we can achieve structural interoperability in the sense that information from these two RDF documents (and actually, the documents themselves) can be merged without the need to create a new schema. It is thus easy to formulate uniform queries that work over heterogeneous language resources. As an example, we can combine information from the linked data version of WordNet and the POWLA formalization of the MASC corpus, e.g., the task to find all tokens in a corpus that refer to *land* as a political unit (synonyms from the WordNet synset `land%1:15:02::`).

Using RDF representations of WordNet and MASC, however, it is no longer necessary to access separate APIs for MASC, GrAF and WordNet. Instead, the task to integrate information from different resources can be easily achieved by applying standard RDF query languages like SPARQL [35] to a repository in which both resources are contained. The sense keys are thus URIs in a RDF version of WordNet such as lwn:synset-land-noun-2. Hence a query as below can be formulated:

```
PREFIX lemon: <http://www.monnet-project.eu/lemon#>
PREFIX lwn: <http://monnetproject.deri.ie/lemonsource/wordnet/> .
SELECT ?token {
  lwn:synset-land-noun-2 lemon:isReferenceOf ?sense
  ?sense lemon:isSenseOf ?entry .
  ?entry rdfs:label ?synonym .
  ?token powla:hasString ?synonym .}
```

2.3.2 Linking and Federation

Linked data is built on URIs as globally unique identifiers. They have the key advantage that resources can be unambiguously identified, thus supporting the creation of a linked web in analogy to the current web of documents (but using properties to link resources instead of the document-oriented, unlabelled HTML hyperlinks). Linked data thus does typically not exist as a set of files on a hard disk or as data in a single data base, but instead as a network of related resources on the web. In other words, techniques must be (and have been) provided that allow queries over linked data to be **federated** over multiple different repositories, physically located at different servers across the world [6, 21, 23, 36].

Rather than querying for WordNet senses and linguistic annotations stored in a single RDF repository, we thus can directly address the public SPARQL endpoint of *lemon source* [32] to access WordNet senses in a subquery:

```
PREFIX lemon: <http://www.monnet-project.eu/lemon#> .
PREFIX lwn: <http://monnetproject.deri.ie/lemonsource/wordnet/> .
SELECT ?token {
  service <http://monnetproject.deri.ie/lemonsource_query/> {
    lwn:synset-land-noun-2 lemon:isReferenceOf ?sense .
    ?sense lemon:isSenseOf ?entry .
    ?entry rdfs:label ?synonym .
  }
  ?token powla:hasString ?synonym .
}
```

If the query engine was configured to do so, it may be able to infer which endpoints to query for certain data based on the URIs used in the query [37]. By building on a standard method for federation of queries on the web, we ensure that the systems take advantage of effective algorithms for federating queries. In this way, information from corpora and lexical-semantic resources can be successfully integrated with each other even if these resources are physically distributed over different repositories.

2.3.3   *Conceptual Interoperability*

RDF does not only establish structural interoperability among and between lexical-semantic resources and corpora, but also between these and resources like terminology repositories or meta-data repositories. In combination with the possibility to query distributed resources, this can also be exploited to enhance the **conceptual interoperability** between language resources, i.e., the use of shared vocabularies for linguistic analyses and metadata.

Ide and Pustejovsky [25] define conceptual ('semantic') interoperability of NLP tools as 'the ability to automatically interpret exchanged information meaningfully and accurately in order to produce useful results'. Further, they suggest that this can be achieved 'via deference to a common information exchange reference model' for language resources and NLP tools.

Different communities create their own grammatical annotations, and although they follow the common goal to establish conceptual interoperability, they have been developed for different use cases, and – even worse – they represent different terminological traditions. Two representative repositories are the General Ontology of Linguistic Description (GOLD, [16]) and the ISO TC37/SC4 Data Category Registry (ISOcat, [41]). Adopting a linked data approach, however, it is possible to link these repositories with each other, i.e., either to link from one resource to the other, or to create mediator ontologies that provide a linking between these repositories. The Ontologies of Linguistic Annotation [9, OLiA] are a modular set of ontologies that establish such a linking. OLiA consists of a *Reference Model*, which specifies the common terminology that different annotation schemes can refer to, as well as *Annotation Models* that formalize annotation schemes and tagsets for about 70 different languages. For every Annotation Model, a *Linking Model* defines relationships between concepts/properties with the Reference Model. In the same way, the Reference Model is linked with several terminology repositories, including GOLD and ISOcat.

Considering annotations in a corpus, say, the syntax annotations of the word *land* from Fig. 2.2, attribute-value pairs like `msj=NN` attached to a particular POWLA `Node` can be exploited to assign this `Node` the superclass `penn:CommonNoun` from the Annotation Model that formalises the corresponding annotation scheme. Through the linking, it can be inferred that this `Node` is also an `olia:CommonNoun` in the Reference Model and that it is an instance of both `isocat:DC-1256` and `gold:CommonNoun`. It would thus become compatible and aligned with any annotation scheme that is linked to either GOLD or ISOcat.

By this kind of linking we can create chains of resources leading to links that would not have been trivial to discover otherwise. As an example, assume that we are interested in studying a particular lexeme in a lexical-semantic resource and that we would like to inspect its usage in a particular corpus. Many lexicons, e.g., those developed on the basis of LexInfo [31], include references to ISOcat data categories. The link between these and the OLiA Reference Model can be discovered – for example – by querying a Semantic Web Search Engine for references to the

ISOcat data category. Dereferencing the OLiA Reference Model, we can find the corresponding Annotation Model concepts that define, inter alia, the corresponding part-of-speech tags. This information can then be exploited to generate corpus queries to retrieve example sentences for the lexeme which combine lemma and spelling information with the appropriate part-of-speech tags. Such queries could then be applied even to corpora that are not provided as linked data.

2.3.4   Ecosystem

RDF comes with a rich repository of tools and formalisms for the processing of graph-based data structures. Using it as representation formalism for multi-layer annotations provides us with convenient means for modelling, manipulating, storing and querying directed labelled graphs. Linked data has achieved success in a wide variety of fields and in fact the linked data paradigm is being applied to a number of domains[7] and is thus supported by a comparably large and active user community.

One consequence is the existence of multiple standards and recommendations maintained by the W3C (e.g., RDFS, OWL, SPARQL) for which new extensions are being developed at a rapid pace.[8] Moreover, there exist a large number of commercial and open-source tools to process linked data, in particular repositories for storing and querying. There are frequent benchmarks of the performance of these tools.[9] In addition, search engines index all the linked data available and allow the discovery of new services.[10]

2.3.5   Dynamic Import

In the traditional approach on modelling language resources, cross-links between different resources are typically represented by attribute-value pairs whose value contains the string representation of IDs as defined within another language resource. Within the linked data approach, however, such information can be represented by a resolvable URI, and is thus accessible in its complete and up-to-date form. When the resource that is referred to is augmented by additional

---

[7]Other domains where the linked data principles have been applied, include, e.g., geography [20], biomedicine [1], cultural history (http://www.europeana.eu) or government data (e.g., http://data.gov and http://data.gov.uk).

[8]For example, the W3C Semantic Web Activity reported on developments for Media Resources, Data Provenance and Microdata in the first 2 weeks of February 2012

[9]http://www4.wiwiss.fu-berlin.de/bizer/berlinsparqlbenchmark

[10]Examples include http://swoogle.umbc.edu, http://www.sindice.net, http://swse.deri.ie, and http://watson.kmi.open.ac.uk.

information, then a system can access this information even though it was not available at the time when the annotation (say, a WordNet sense) was created. Maintenance efforts nowadays necessary to maintain the proper linking of corpora with the most recent WordNet edition available can thus be reduced to a minimum. Furthermore, the use of URIs instead of system-defined IDs solves another problem, namely that such informal ID references are usually not unambiguous. For example, the version of the WordNet referred to a resource can be indicated by its full URI avoiding the need to explicitly state the version number.

However, dynamism can be a "double-edged sword". Although continuous corrections may improve the quality of a resource, this entails the risk that references from external resources are no longer valid, e.g., because a sense has been redefined, split or merged with another. Following an established publication practice for linguistic resources, it is thus advisable to provide stable release editions and to indicate these differences in the corresponding URIs.

2.4 Community Efforts Towards Lexical Linked Data

Publishing language resources using such interoperable representations, formally defined data types and resolvable URI to designate elements of linguistic analysis/ annotation allows existing linguistic resources to be connected. Aside from the benefits enumerated in the last section, this facilitates the distributed, but highly synchronized development of linguistic resources. The technological infrastructure developed around RDF makes it an attractive candidate for the creation, exchange and processing of language resources in different sub-disciplines of linguistics, NLP and neighbouring fields. Its genericity allows researchers from all these different subcommunities to share data and experiences; thereby, RDF encourages interdisciplinary cooperation.

Consequently, linked data is at the core of recent community activities. We describe two initiatives heading towards the creation of a linked (open) data cloud of linguistic data.

2.4.1 The Open Linguistics Working Group

The Open Linguistics Working Group (OWLG, [12])[11] of the Open Knowledge Foundation was founded in late 2010 as an initiative of experts from different fields concerned with linguistic data, including academic linguists (e.g. typology, corpus linguistics), applied linguistics (e.g. computational linguistics, lexicography and language documentation), and information technology (e.g. Natural Language

---

[11]http://linguistics.okfn.org

Processing, Semantic Web). The primary goals of the working group are to promote the idea of **open linguistic resources**, to explore **means for their representation**, and to encourage the **exchange of ideas** across different disciplines.

A number of concrete community projects have been initialized,[12] including the documentation of workflows, documenting best practice guidelines and collecting use cases with respect to legal issues of linguistic resources. Of particular importance in this context is the collection of representative resources available under open licenses, the identification of possible links between these resources and, consequently, the creation of a **Linguistic Linked Open Data cloud**.[13]

For resources published under open licenses, an RDF representation yields the additional advantage that resources can be interlinked, and it is to be expected that an additional gain of information arises from the resulting network of resources. So, although the OWLG is dedicated to open resources in linguistics in general, and not a priori restricted to linked data, a general consensus has been established within the OWLG that Semantic Web formalisms provide crucial advantages for the publication of linguistic resources, some of which have been illustrated here as well.

The idea of linked data is gaining ground: data sets from different subdisciplines of linguistics and neighbouring fields are currently prepared. Recent activities include subject areas as diverse as language acquisition, the study of folk motifs, phonological typology, translation studies, pragmatics and comparative lexicography [13]. The OWLG represents a platform for the exchange of ideas, data and information across all these different fields.

2.4.2   W3C Ontology-Lexica Community Group

The Ontology-Lexica Community (OntoLex) Group,[14] was founded as a W3C Community and Business Group in September 2011. It aims to produce specifications for a **lexicon-ontology model** that can be used to provide rich linguistic grounding for domain ontologies. Rich linguistic grounding includes the representation of morphological, syntactic properties of lexical entries as well as the syntax-semantics interface, i.e. the meaning of these lexical entries with respect to the ontology in question. An important issue herein will be to clarify how extant lexical and language resources can be leveraged and reused for this purpose. As a by-product of this work on specifying a lexicon-ontology model, it is hoped that such a model can become the basis for a web of lexical linked data: a network of lexical and terminological resources that are linked according to the linked data principles forming a large network of lexical-syntactic knowledge.

---

[12]http://wiki.okfn.org/Wg/linguistics

[13]http://linguistics.okfn.org/llod

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

Five general requirements for the lexicon-ontology model were identified:

RDF/OWL The actual model is an OWL ontology, a specific lexicon instantiating the model is a plain RDF document.

Multilingualism The model supports the specification of the linguistic grounding with respect to any language.

Semantics by reference The meaning of a lexical entry is specified by referencing the URI of the concept or property in question.

Flexible infrastructure The lexicon-ontology model is extensible by new constructs as needed, e.g. by a certain application, and it makes no unnecessary choices with respect to which linguistic data categories to use, i.e., leaving open the possibilities to have very different instantiations of the model.

Interoperability Reuse of relevant standards (e.g. LMF).

2.5 Summary

In this chapter, we suggested that modelling linguistic resources as linked data provides a number of crucial advantages as compared to existing formalisms. In particular, modelling linguistic resources in RDF can lead to enhanced **interoperability** (and thus, scalability) for applications, **knowledge integration**, and access to **distributed resources**, and last but not least the rich **infrastructure** provided by the Semantic Web community can be applied to develop infrastructures for NLP, computational lexicography or corpus linguistics. In this way, linked data might facilitate the work of application developers, users of language resources and the natural language processing community as a whole.

A specific characteristic of RDF and linked data in general is that resources and their components (e.g., entries in a dictionary) are represented by URIs, thus enabling the **globally unambiguous referencing** of data. By using resolvable URIs to refer to other resources, resources can be **interlinked** and thereby integrated. For example, a corpus can be directly connected to a lexical-semantic resource, different lexical-semantic resources can be queried simultaneously and information from various sources can be combined. Further, we described recent **community efforts** in the NLP and Semantic Web communities heading towards the provision of a larger set of linguistic resources as linked data.

Overall, in this chapter we have discussed the benefits of publishing linguistic data as linked data and outlined a vision, sketching the potential, implications and applications thereof. The vision we have outlined is not a far-fetched one. From a technological point of view, the main ingredients are already in place, in particular RDF, OWL and SPARQL. Furthermore, as linked data grows in popularity across multiple disciplines, tools that can be applied to linguistic linked data will only increase in number and power.

References

1. Ashburner, M., Ball, C.A., et al.: Gene ontology: tool for the unification of biology. Nat. Genet. **25**(1), 25–29 (2000)
2. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The Berkeley FrameNet project. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL-1998), Montréal, pp. 86–90 (1998)
3. Bird, S., Liberman, M.: A formal framework for linguistic annotation. Speech Commun. **33**(1), 23–60 (2001)
4. Bizer, C., Heath, T., Berners-Lee, T.: Linked data – the story so far. Int. J. Semant. Web Inf. Syst. (IJSWIS) **5**(3), 1–22 (2009)
5. Brandes, U., Eiglsperger, M., et al.: Graph markup language (GraphML). In: Tamassia, R. (ed.) Handbook of Graph Drawing and Visualization. Chapman & Hall/CRC, London (2010)
6. Buil-Aranda, C., Arenas, M., Corcho, O.: Semantics and optimization of the SPARQL 1.1 federation extension. In: The Semantic Web: Research and Applications, pp. 1–15. Springer, Heraklion (2011)
7. Carletta, J., Evert, S., et al.: The NITE XML Toolkit: data model and query. Lang. Resour. Eval. J. (LREJ) **39**(4), 313–334 (2005)
8. Cassidy, S.: An RDF realisation of LAF in the DADA annotation server. In: Proceedings of the 5th Joint ISO-ACL/SIGSEM Workshop on Interoperable Semantic Annotation (ISO-5), Hong Kong (2010)
9. Chiarcos, C.: An ontology of linguistic annotations. LDV Forum **23**(1), 1–16 (2008)
10. Chiarcos, C.: Interoperability of corpora and annotations. In Chiarcos, C., Nordhoff, S., Hellmann, S. (eds.) Linked Data in Linguistics, pp. 161–179. Springer, Heidelberg (2012)
11. Chiarcos, C., Dipper, S., et al.: A flexible framework for integrating annotations from different tools and tagsets. TAL (Traitement automatique des langues) **49**(2), 217–246 (2008)
12. Chiarcos, C., Hellmann, S., et al.: The open linguistics working group. In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2012), Istanbul (2012a)
13. Chiarcos, C., Nordhoff, S., Hellmann, S. (eds.): Linked Data in Linguistics. Representing Language Data and Metadata. Springer, Heidelberg (2012b)
14. Chiarcos, C., Ritz, J., Stede, M.: By all these lovely tokens . . . Merging conflicting tokenizations. J. Lang. Resour. Eval. (LREJ) **4**(45), 53–74 (2012c)
15. Dipper, S.: XML-based stand-off representation and exploitation of multi-level linguistic annotation. In: Eckstein, R., Tolksdorf, R. (eds.) Proceedings of Berliner XML Tage 2005 (BXML-2005), Berlin, pp. 39–50 (2005)
16. Farrar, S., Langendoen, D.T.: An OWL-DL implementation of GOLD: an ontology for the Semantic Web. In: Witt, A., Metzing, D. (eds.) Linguistic Modeling of Information and Markup Languages. Springer, Dordrecht (2010)
17. Fellbaum, C.: WordNet. MIT, Cambridge (1998)
18. Fielding, R., Gettys, J., et al.: Hypertext transfer protocol – HTTP/1.1. Internet RFC 2068 (1997)
19. Francopoulo, G., George, M., et al.: Lexical markup framework (LMF). In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006), Genoa (2006)

20. Goodwin, J., Dolbear, C., Hart, G.: Geographical linked data: the administrative geography of Great Britain on the Semantic Web. Trans. GIS **12**, 19–30 (2008)
21. Guéret, C., Kotoulas, S., Groth, P.: TripleCloud: an infrastructure for exploratory querying over web-scale RDF data. In: Proceedings of the 2011 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT 2011), Lyon, pp. 245–248 (2011)
22. Gurevych, I., Eckle-Kohler, J., et al.: Uby – a large-scale unified lexical semantic resource based on LMF. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2012), Avignon, pp. 580–590 (2012)
23. Hartig, O., Bizer, C., Freytag, J.C.: Executing SPARQL queries over the web of linked data. In: The Semantic Web – ISWC 2009, Heraklion, pp. 293–309 (2009)
24. Holtman, K., Mutz, A.: Transparent content negotiation in HTTP. Internet RFC 2295 (1998)
25. Ide, N., Pustejovsky, J.: What does interoperability mean, anyway? Toward an operational definition of interoperability. In: Proceedings of the 2nd International Conference on Global Interoperability for Language Resources (ICGL 2010), Hong Kong (2010)
26. Ide, N., Suderman, K.: GrAF: A graph-based format for linguistic annotations. In: Proceedings of the First Linguistic Annotation Workshop (LAW 2007), Prague, pp. 1–8 (2007)
27. Ide, N., Le Maitre, J., Véronis, J.: Outline of a model for lexical databases. In: Zampolli, A., Calzolari, N., Palmer, M.S. (eds.) Current Issues in Computational Linguistics: In Honour of Don Walker, Giardini, pp. 283–320 (1995)
28. Ide, N., Fellbaum, C., et al.: The manually annotated sub-corpus: a community resource for and by the people. In: Proceedings of the ACL 2010 Conference Short Papers, Uppsala, pp. 68–73 (2010)
29. Klyne, G., Carroll, J.J, McBride, B.: Resource description framework (RDF): concepts and abstract syntax. Technical report, W3C Recommendation (2004)
30. Marcus, M.P., Santorini, B., Marcinkiewicz, M.A.: Building a large annotated corpus of english: the penn treebank. Comput. Linguist. **19**(2), 313–330 (1994)
31. McCrae, J., Spohr, D., Cimiano, P.: Linking lexical resources and ontologies on the Semantic Web with Lemon. In: The Semantic Web: Research and Applications, Heraklion, pp. 245–259 (2011)
32. McCrae, J., Montiel-Ponsoda, E., Cimiano, P.: Collaborative semantic editing of linked data lexica. In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2012), Istanbul (2012a)
33. McCrae, J., Montiel-Ponsoda, E., Cimiano, P.: Integrating WordNet and wiktionary with lemon. In: Chiarcos, C., Nordhoff, S., Hellmann, S. (eds.) Linked Data in Linguistics, pp. 25–34, Springer, Heidelberg (2012b)
34. Miller, G.A.: WordNet: a lexical database for English. Commun. ACM **38**(11), 39–41 (1995)
35. Prud'Hommeaux, E., Seaborne, A.: SPARQL query language for RDF. W3C working draft (2008)
36. Quilitz, B., Leser, U.: Querying distributed RDF data sources with SPARQL. In: The Semantic Web: Research and Applications, pp. 524–538. Springer, Berlin/Heidelberg (2008)
37. Schenk, S., Petrák, J.: Sesame RDF repository extensions for remote querying. In: Proceedings of the 7th Znalosti Conference (Znalosti-2008), Bratislava (2008)
38. Shadbolt, N., Hall, W., Berners-Lee, T.: The semantic web revisited. IEEE Intell. Syst. **21**(3), 96–101 (2006)
39. Van Assem, M., Gangemi, A., Schreiber, G.: Conversion of WordNet to a standard RDF/OWL representation. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006), Genoa, pp. 237–242 (2006)
40. Véronis, J., Ide, N.: A feature-based model for lexical databases. In: Proceedings of the 14th International Conference on Computational Linguistics (COLING-1992), Nantes, pp. 588–594 (1992)
41. Windhouwer, M., Wright, S.E.: Linking to linguistic data categories in ISOcat. In: Chiarcos, C., Nordhoff, S., Hellmann, S. (eds.) Linked Data in Linguistics, pp. 99–107. Springer, Heidelberg (2012)