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 number of fields. However for Linguistics, much data is still published in 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 describe the application of this paradigm to the modeling of two language resources, WordNet and the MASC corpus, that 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.

1 Motivation and Overview

Language is arguably one of the most complex forms of human behavior, 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 the language resources, a problem that is at best partially solved (Ide and Pustejovsky, 2010). 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 principle 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 Linked Data Principles represent rules of best practice that should be followed when publishing data on the Web (Bizer et al., 2009):

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 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, Klyne et al., 2004) and querying linked data (SPARQL, Prud'Hommeaux and Seaborne, 2008). 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 labeled directed (multi-)graphs, which can be serialized in different formats, where 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 labeled edge) that connects a *subject* (a resource, in graph-theoretical terms a labeled 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 Ta-

[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.

Table 1 Selected relations from existing RDF vocabularies and possible fields of application

domain	example property	reference
meta data	`creator`	Dublin Core meta data categories
general relationships 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

ble 1, the links provided give a more detailed description of how they are to be used. 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 reused 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. For example, the Web Ontology Language (OWL) defines the datatypes 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 labeled directed graphs (feature structures, Ide et al., 1995), as can annotated corpora (Bird and Liberman, 2001). RDF is based on labeled directed graphs and thus particularly well-suited for modeling 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. Moreover, HTTP allows multiple formats for the same resource to be published at the same location.

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 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 the possibility of checking a lexicon or 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 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 paper. It is structured as follows: Section 2 describes the modeling of linguistic resources as linked data and identifies deficits and prospective advantages of using linked data for linguistic resources. Section 3 elaborates some of the benefits of this representation. Section 4 summarizes recent community activities promoting the publication of language resources as Linked Data.

2 Modeling Linguistic Resources as Linked Data

We consider two important classes of language resources, the first of which is **lexical-semantic resources** under which we group machine-readable dictionaries, semantic networks, semantic knowledge bases, ontologies and terminologies. The second class of language resources considered here are **annotated corpora**. For both types of resources, we describe state-of-the-art approaches, briefly motivate the application of linked data principles, and then describe modeling these resources using RDF and OWL.

Resource modeling 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.

Unlike other graph-based modeling formalisms applied to language resources, e.g., GraphML (Brandes et al., 2010), 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 modeling 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) of synonyms of 'land' in the sense of *'(the territory occupied by a nation)'* (in WordNet 3.1) 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 to represent different types of linguistic resources (corpora, treebanks, lexico-semantic resources such as WordNet etc.) in a uniform way thus supporting uniform querying across resources. The above query could be for example represented in SPARQL as follows:

```
PREFIX wn20: <http://www.w3.org/2006/03/wn/wn20/schema/> .
PREFIX rkbWN: <http://wordnet.rkbexplorer.com/id/> .
SELECT ?token {
  rkbWN:synset-land-noun-2
     wn20:containsWordSense ?sense .
  ?sense rdfs:label ?synonym .
  ?token powla:hasString ?synonym .
}
```

Assuming that there is a mechanism that can distribute this query to different RDF repositories or SPARQL endpoints that contain the relevant MASC and WordNet data, answering the query is indeed straightforward.

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

2.1 Modeling Lexical-Semantic Resources: WordNet

2.2 WordNet Data Structures

WordNet (Miller, 1995; Fellbaum, 1998) 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 indicated 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.[2]

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 arcs 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 Fellbaum, 1998).

2.2.1 Generic Data Structures: Lexical Markup Framework

To facilitate interoperability among different lexical-semantic resources, generalizations of their data structures have been developed on the basis of feature structures (i.e., directed acyclic graphs) as a flexible and general formalism (Vronis and Ide, 1992). On this basis, standards have been developed, in particular, the Lexical Markup Framework (LMF, Francopoulo et al., 2006). LMF represents a metamodel aiming to provide a standard to represent semantic information in NLP lexicons and machine-readable dictionaries. It has been successfully applied to develop resources such as Uby (Gurevych et al., to appear), an openly available large-scale lexico-semantic resource integrating a wide range of information from nine lexico-semantic 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 how resources can be uniquely identified on the web. Furthermore, according to the experience of Uby, application of the format requires making domain-specific modifications to the standard schema.

An RDF formalization of LMF allows us to tackle some of these problems, and this has been suggested by the LMF developers themselves (Francopoulo et al., 2009).[3] 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.

2.2.2 From LMF to RDF: Lemon

Independently from LMF, there has already been some work towards the integration of WordNet with the Semantic Web, notably the conversion of WordNet by Van As-

[2] `http://www.wordnet.princeton.edu`

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

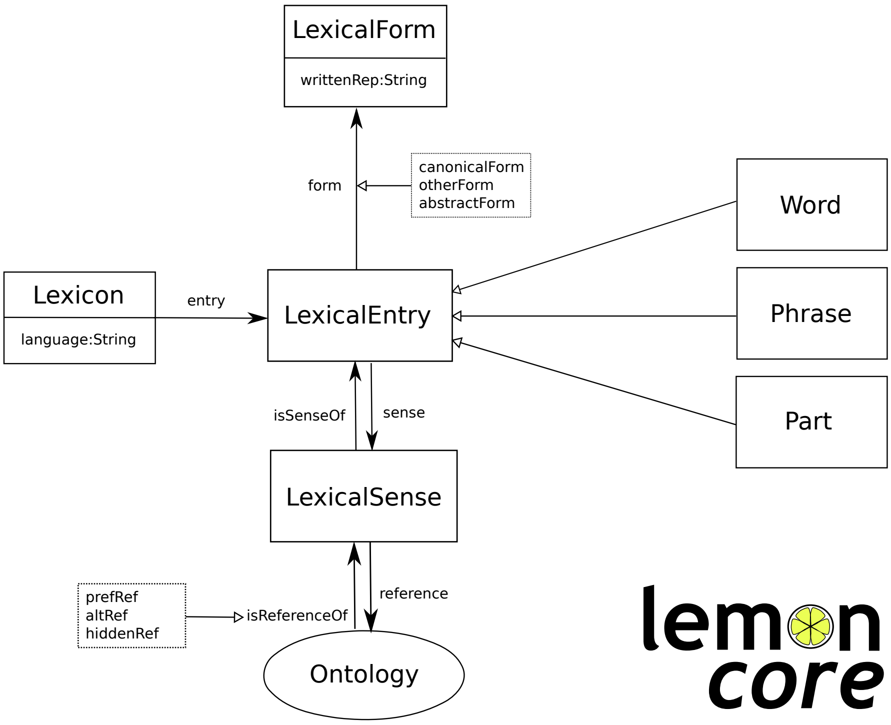

Fig. 1 The core of the *lemon* model

sem et al. (2006), 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, McCrae et al. have propose an interchange model called *lemon* supporting the publication of lexical-semantic resources as lexical linked data. *lemon* stands for (**Le**xicon **M**odel for **On**tologies) and builds on the following principles:

1. **based on LMF** (to allow easy conversion from non-linked data resources);
2. **RDF-native** (publishing as linked data, with RDFS and OWL used to describe the semantics of the model);
3. **modular** (separation of lexicon and ontology layers, so that *lemon* lexica can be linked to existing ontologies in the linked data cloud);
4. **externally defined data categories** (linking to data categories in repositories of annotation terminology rather than proposing a specific set of part-of-speech tags);
5. **the 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, Shadbolt et al., 2006).

This format is illustrated in Fig. 1. *lemon* has been used as a basis for integrating the data of the English Wiktionary [4], a (human-readable) dictionary created along 'wiki' principles, with the data from the RDF version of WordNet (see McCrae et al. (2012b)). 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 with only minor differences in the modelling of inflectional variants of lexical entries. However, the semantic modeling was more significantly different as *lemon* uses the OWL ontology language to represent semantics.

2.3 Modeling annotated corpora: MASC

2.3.1 The Manually Annotated Sub-Corpus

The Manually Annotated Sub-Corpus (MASC, Ide et al., 2010) 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.[5] 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 used different output formats. In order to establish interoperability between them, MASC distributions adopt a generic data model, the Graph Annotation Framework (GrAF, Ide and Suderman, 2007). 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.

2.3.2 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 labeled directed graphs (Bird and Liberman, 2001). To a certain extent, this echoes the application of feature structures to lexical-semantic resources as feature structures can be also be interpreted as directed acyclic graphs.

One representative example for graph-based generic formats is the Graph Annotation Framework (GrAF). Like other state-of-the-art approaches that implement graph-based data models for linguistic corpora (Carletta et al., 2005; Chiarcos et al., 2008), 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 differ-

[4] `http://en.wiktionary.org/`

[5] `www.anc.org/MASC`

ent layers of annotations. In Fig. 2, this is shown for an example sentence from the MASC corpus. As a multi-layer corpus, MASC is distributed in the GrAF format, with all annotations of a document 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 type of annotation. 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. Thus, it seems clear that as standoff formats become the norm, it is necessary to rely on models that are fundamentally linked as opposed to hierarchical models that are easy to serialize in formats such as XML.

Figure 2 shows the graph-based modeling 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 (Baker et al., 1998) and to the right PennTreebank-style syntax annotations (Marcus et al., 1994) are shown. Both annotations are synchronized with each other and the primary data through a shared base segmentation file.

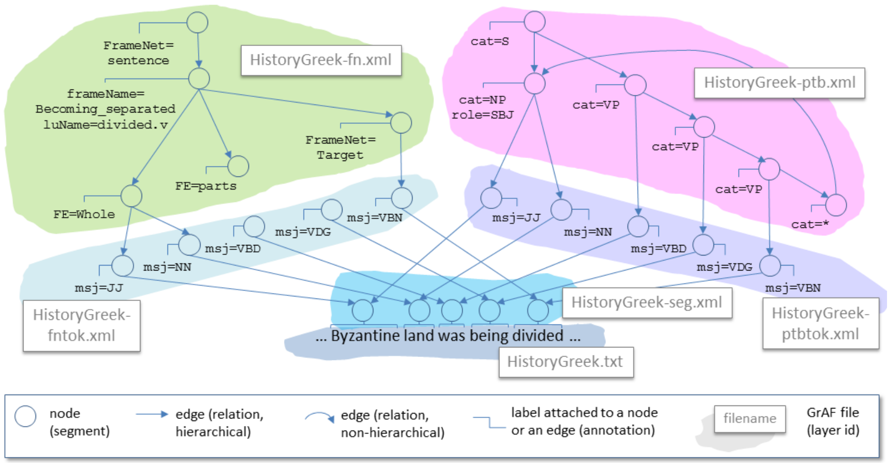

Fig. 2 Representing and integrating annotations for syntax and frame-semantics in a directed graph

2.3.3 From Standoff XML to RDF: POWLA

As mentioned above, standoff formats can be relatively hard to process, and corresponding infrastructures and standards are still under development, whereas RDF already provides a rich technological ecosystem for labeled directed graphs. Accordingly, 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 (2010) and Chiarcos (2012).

Chiarcos (2012) described POWLA, an RDF/OWL linearization of PAULA, a generic data model for the representation of annotated corpora (Dipper, 2005; Chiarcos et al., 2011). PAULA is similar in scope and design to GrAF and also builds on traditional standoff annotations. 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 to RDF converter is provided under `http://purl.org/powla`, it allows to replicate 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.

3 Benefits of Linked Data for Linguistics

Aside from representation, we have identified five specific advantages of modeling 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).

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'

(Ide and Pustejovsky, 2010). 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**.

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**, i.e., a user or agent that accesses a particular resource can specify the format they want to obtain 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* 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.

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. As such it is 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, i.e., synoynms from the WordNet synset `land%1:15:02::`.

Using RDF representations of WordNet and MASC, however, accessing separate APIs for MASC, GrAF and WordNet is not necessary. Instead, the task to integrate information from different resources can be easily achieved by applying standard RDF query languages like SPARQL (Prud'Hommeaux and Seaborne, 2008) to a repository in which both resources are contained. The sense keys are thus URIs in a

RDF version of WordNet such as `http://wordnet.rkbexplorer.com/id/synset-land-noun-2`. Hence a query as below can be formulated:

```
PREFIX wn20: <http://www.w3.org/2006/03/wn/wn20/schema/> .
PREFIX rkbWN: <http://wordnet.rkbexplorer.com/id/> .
SELECT ?token {
  rkbWN:synset-land-noun-2
     wn20:containsWordSense ?sense .
  ?sense rdfs:label ?synonym .
  ?token powla:hasString ?synonym .
}
```

3.2 *Linking and Federation*

Linked Data is built on URIs as globally unique and unambiguous identifiers. They have the key advantage that resources can be uniquely 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 and unlabelled hyperlinks used in HTML. Linked Data thus does not exist as a set of files on a hard disk but instead as a network of related resources on the web. Thus, linguistic data need not necessarily to exist in a single repository or database, but instead queries can be **federated** over multiple different repositories, physically located at potentially different servers across the world (Quilitz and Leser, 2008; Hartig et al., 2009; Guéret et al., 2011; Buil-Aranda et al., 2011).

As such, instead of querying for WordNet senses and linguistic annotations stored in a single RDF repository, we can directly address the public SPARQL endpoint of RKB Explorer[6] in a subquery:

```
PREFIX wn20: <http://www.w3.org/2006/03/wn/wn20/schema/> .
PREFIX rkbWN: <http://wordnet.rkbexplorer.com/id/> .
SELECT ?token {
  service <http://wordnet.rkbexplorer.com/sparql> {
    rkbWN:synset-land-noun-2
       wn20:containsWordSense ?sense .
    ?sense 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 (Schenk and Petrk, 2008). 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 in different repositories.

[6] `http://wordnet.rkbexplorer.com/sparql`

3.3 Conceptual Interoperability

RDF does not only establish structural interoperability among and between LSRs 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 potential 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 (2010) 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 annoations, 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, Farrar and Langendoen, 2003, 2010) and the ISO TC37/SC4 Data Category Registry (ISOcat, Ide and Romary, 2004; Windhouwer and Wright, 2012). 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 (Chiarcos, 2008, 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 in the respective Annotation Model and the Reference Model. In the same way, the OLiA 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, attribute-value pairs like `msj=NN` attached to a particular `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 (McCrae et al., 2011), 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 defines, inter alia, the corresponding part of speech tags. This information can then be exploited, for example, 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 can then applied, for example, even to corpora that are not provided as Linked Data.

3.4 Ecosystem

RDF comes with a relatively rich repository of tools and formalisms for the processing of graph-based data structures. Using it as a representation formalism for multi-layer annotations provides us with convenient means for modeling, manipulating, storing and querying directed labeled 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, several search engines index all the linked data available and allow the discovery of new services.[10]

3.5 Dynamic Import

In the traditional approach on modeling 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

[7] Other domains where the Linked Data paradigm has been applied, include, e.g., geography (Goodwin et al., 2008), biomedicine (Ashburner et al., 2000), cultural history (`http://www.europeana.eu`) or government data (e.g., `data.gov` in the US and `data.gov.uk` in the UK).

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

[9] Berlin SPARQL Benchmark: `http://www4.wiwiss.fu-berlin.de/bizer/berlinsparqlbenchmark/`

[10] Examples include Swoogle `swoogle.umbc.edu`, Sindice `www.sindice.net`, SWSE `swse.deri.ie`, and Watson `watson.kmi.open.ac.uk`

resolvable URI, and is thus accessible in its complete and up-to-date form. If the resource that is referred to is augmented by additional information, then a system can access this information even though it was not available at the time when the word sense annotation 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 practice in the publication of linguistic resources (both corpora and lexical-semantic resources), it is thus advisable to focus stable release editions and to indicate these differences in the corresponding URIs.

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 resources to be connected, thereby creating a web of (linguistic) data. 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 neighboring fields. Its genericity allows researchers from all these different subcommunities to share data and experiences; thereby, RDF encourages interdisciplinary cooperations.

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.

4.1 The Open Linguistics Working Group

The Open Linguistics Working Group (OWLG)[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, cor-

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

pus linguistics), applied linguistics (e.g. computational linguistics, lexicography and language documentation), and information technology (e.g. Natural Language Processing, Semantic Web). The primary goals of the working group are to promote the idea of open linguistic resources, to develop 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 neighboring fields are currently prepared. Related efforts, e.g. those assembled in Chiarcos et al. (2012), include fields so diverse as language acquisition, the study of folk motifs, phonological typology, translation studies, pragmatics and comparative lexicography. The OWLG represents a platform for the exchange of ideas, data and information across all these different disciplines.

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 byproduct 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 lexico-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.

Multilinguality 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, in particular lexicon models such as LMF.

5 Summary

In this paper, we suggested that modeling linguistic resources as linked data provides a number of crucial advantages as compared to existing formalisms. In particular, modeling 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 the use of 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 paper 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 and OWL. 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

M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25, 2000.

C. F. Baker and C. Fellbaum. WordNet and FrameNet as Complementary Resources for Annotation. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 125–129, August 2009.

C.F. Baker, C.J. Fillmore, and J.B. Lowe. The Berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, pages 86–90, 1998.

T. Berners-Lee. Tim Berners-Lee on the next Web, February 2009. URL `http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html`.

S. Bird and M. Liberman. A formal framework for linguistic annotation. *Speech Communication*, 33(1):23–60, 2001.

C. Bizer, T. Heath, and T. Berners-Lee. Linked data-the story so far. *International Journal on Semantic Web and Information Systems*, 14:9, 2009.

U. Brandes, M. Eiglsperger, J. Lerner, and C. Pich. Graph Markup Language (GraphML). *Handbook of Graph Drawing and Visualization*, 2010.

V. Bryl, C. Giuliano, L. Serafini, and K. Tymoshenko. Using background knowledge to support coreference resolution. In *Proceedings of the 19th European Conference on Artificial Intelligence (ECAI 2010), August*, 2010.

C. Buil-Aranda, M. Arenas, and O. Corcho. Semantics and optimization of the SPARQL 1.1 federation extension. *The Semanic Web: Research and Applications*, pages 1–15, 2011.

J. Carletta, S. Evert, U. Heid, and J. Kilgour. The NITE XML Toolkit: data model and query. *Language Resources and Evaluation Journal*, 39(4):313–334, 2005.

S. Cassidy. An rdf realisation of laf in the dada annotation server. *Proceedings of ISA-5, Hong Kong*, 2010.

C. Chiarcos. An ontology of linguistic annotations. *LDV Forum*, 23(1):1–16, 2008.

C. Chiarcos. Interoperability of Corpora and Annotations. In C. Chiarcos, S. Nordhoff, and S. Hellmann, editors, *Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata*, pages 161–179, Heidelberg, 2012. Springer.

C. Chiarcos, S. Dipper, M. Götze, U. Leser, A. Lüdeling, J. Ritz, and M. Stede. A Flexible Framework for Integrating Annotations from Different Tools and Tagsets. *TAL (Traitement automatique des langues)*, 49(2), 2008.

C. Chiarcos, J. Ritz, and M. Stede. By all these lovely tokens ... Merging conflicting tokenizations. *Journal of Language Resources and Evaluation*, 4(45), 2011. to appear.

C. Chiarcos, S. Nordhoff, and S. Hellmann, editors. *Linked Data in Linguistics. Representing Language Data and Metadata*. Springer, Heidelberg, 2012. companion volume of the Workshop on Linked Data in Linguistics 2012 (LDL-2012), held in conjunction with the 34th Annual Meeting of the German Linguistic Society (DGfS), March 2012, Frankfurt/M., Germany.

S. Dipper. XML-based stand-off representation and exploitation of multi-level linguistic annotation. In *Proc. Berliner XML Tage 2005 (BXML 2005)*, pages 39–50, 2005.

S. Farrar and D. T. Langendoen. A Linguistic Ontology for the Semantic Web. *GLOT International*, 7:97–100, 2003.

S. Farrar and D. T. Langendoen. An OWL-DL implementation of GOLD: An ontology for the Semantic Web. In A. W. Witt and D. Metzing, editors, *Linguistic Modeling of Information and Markup Languages: Contributions to Language Technology*. Springer, Dordrecht, 2010.

C. Fellbaum. *WordNet*. MIT Press, Cambridge, MA, 1998.

G. Francopoulo, M. George, N. Calzolari, M. Monachini, N. Bel, M. Pet, C. Soria, et al. Lexical Markup Framework (LMF). In *International Conference on Language Resources and Evaluation (LREC 2006)*, 2006.

G. Francopoulo, N. Bel, M. George, N. Calzolari, M. Monachini, M. Pet, and C. Soria. Multilingual resources for NLP in the Lexical Markup Framework (LMF). *Language Resources and Evaluation*, 43(1):57–70, 2009.

A. Gangemi, N. Guarino, C. Masolo, and A. Oltramari. Sweetening wordnet with dolce. *AI magazine*, 24(3):13, 2003.

J. Goodwin, C. Dolbear, and G. Hart. Geographical linked data: The administrative geography of great britain on the semantic web. *Transactions in GIS*, 12:19–30, 2008.

C. Guéret, S. Kotoulas, and P. Groth. Triplecloud: An infrastructure for exploratory querying over web-scale rdf data. In *Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pages 245–248, 2011.

I. Gurevych, J. Eckle-Kohler, S. Hartmann, M. Matuschek, C. M. Meyer, and C. Wirth. Uby – A large-scale unified lexical semantic resource based on LMF. In *Proceedings of EACL 2012*, Avignon, France, April to appear.

O. Hartig, C. Bizer, and J.C. Freytag. Executing SPARQL queries over the web of linked data. *The Semantic Web-ISWC 2009*, pages 293–309, 2009.

N. Ide and J. Pustejovsky. 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)*, 2010.

N. Ide and L. Romary. A registry of standard data categories for linguistic annotation. In *Proceedings of the Fourth Language Resources and Evaluation Conference (LREC 2004)*, pages 135–139, 2004.

N. Ide and K. Suderman. GrAF: A graph-based format for linguistic annotations. In *Proceedings of Linguistic Annotation Workshop (LAW 2007)*, pages 1–8, 2007.

N. Ide, J. Le Maitre, and J. Vronis. Outline of a model for lexical databases. In A. Zampolli, N. Calzolari, and M.S. Palmer, editors, *Current Issues in Computational Linguistics: In Honour of Don Walker*, pages 283–320. Giardini, Pisa, 1995.

N. Ide, C. Fellbaum, C. Baker, and R. Passonneau. The manually annotated subcorpus: A community resource for and by the people. In *Proceedings of the ACL-2010*, pages 68–73, 2010.

G. Klyne, J.J Carroll, and B. McBride. Resource Description Framework (RDF): Concepts and Abstract Syntax. Technical report, W3C Recommendation, 2004. URL `http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/`.

M.P. Marcus, B. Santorini, and M.A. Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1994.

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

J. McCrae, G. Aguado-de Cea, P. Buitelaar, P. Cimiano, T. Declerck, A. Gomez-Perez, J. Gracia, L. Hollink, E. Montiel-Ponsoda, D. Spohr, and T. Wunner. Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 2012a.

J. McCrae, E. Montiel-Ponsoda, and P. Cimiano. Integrating WordNet and Wiktionary with lemon. In C. Chiarcos, S. Nordhoff, and S. Hellmann, editors, *Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata*, pages 25–34, Heidelberg, 2012b. Springer.

D.L. McGuinness, F. Van Harmelen, et al. OWL web ontology language overview. Technical report, W3C recommendation, 2004.

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

E. Prud'Hommeaux and A. Seaborne. SPARQL query language for RDF. *W3C working draft*, 4(January), 2008.

B. Quilitz and U. Leser. Querying distributed rdf data sources with sparql. *The Semantic Web: Research and Applications*, pages 524–538, 2008.

S. Schenk and J. Petrk. Sesame RDF repository extensions for remote querying. In *Proceedings of Znalosti 2008*, 2008.

N. Shadbolt, W. Hall, and T. Berners-Lee. The semantic web revisited. *IEEE intelligent systems*, 21(3):96–101, 2006.

M. Van Assem, A. Gangemi, and G. Schreiber. Conversion of WordNet to a standard RDF/OWL representation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 237–242, 2006.

J. Vronis and N. Ide. A feature-based model for lexical databases. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING'92)*, pages 588–594, Nantes (France), 1992.

M. Windhouwer and S. E. Wright. Linking to linguistic data categories in ISOcat. In C. Chiarcos, S. Nordhoff, and S. Hellmann, editors, *Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata*, pages 99–107. Springer, Heidelberg, 2012.