Natural Language Engineering: page 1 of 49. © Cambridge University Press 2018
doi:10.1017/S1351324918000347

*Models to represent linguistic linked data**

J. BOSQUE-GIL[1], J. GRACIA[1,2],
E. MONTIEL-PONSODA[1] and A. GÓMEZ-PÉREZ[1]

[1]*Ontology Engineering Group, Escuela Técnica Superior de Ingenieros Informáticos, Universidad Politécnica de Madrid, Spain*
e-mails: jbosque@fi.upm.es, emontiel@fi.upm.es, asun@fi.upm.es
[2]*Aragon Institute of Engineering Research (I3A), University of Zaragoza, Spain*
e-mail: jogracia@unizar.es

(*Received 13 June 2017; revised 6 August 2018; accepted 10 August 2018*)

Abstract

As the interest of the Semantic Web and computational linguistics communities in linguistic linked data (LLD) keeps increasing and the number of contributions that dwell on LLD rapidly grows, scholars (and linguists in particular) interested in the development of LLD resources sometimes find it difficult to determine which mechanism is suitable for their needs and which challenges have already been addressed. This review seeks to present the state of the art on the models, ontologies and their extensions to represent language resources as LLD by focusing on the nature of the linguistic content they aim to encode. Four basic groups of models are distinguished in this work: models to represent the main elements of lexical resources (group 1), vocabularies developed as extensions to models in group 1 and ontologies that provide more granularity on specific levels of linguistic analysis (group 2), catalogues of linguistic data categories (group 3) and other models such as corpora models or service-oriented ones (group 4). Contributions encompassed in these four groups are described, highlighting their reuse by the community and the modelling challenges that are still to be faced.

1 Introduction

1.1 Background and motivation

Language resources (dictionaries, terminologies, corpora, etc.) developed in the fields of corpus linguistics, computational linguistics and natural language processing (NLP) are often encoded in heterogeneous formats and developed in isolation from one another. This makes their discovery, reuse and integration for both the development of NLP tools and daily linguistic research a difficult and cumbersome task. In

*We are very grateful to the anonymous reviewers for their meticulous reading of the survey and for providing us with numerous insightful and constructive suggestions to improve it. We would also like to thank Dr Guadalupe Aguado-de-Cea for her help in proofreading this manuscript. This work is supported by the Spanish Ministry of Education, Culture and Sports through the Formación del Profesorado Universitario (FPU) program, and by the Spanish Ministry of Economy and Competitiveness through the project 4V (TIN2013-46238-C4-2-R) within the FEDER funding scheme, the Juan de la Cierva program, and the Excellence Network ReTeLe (TIN2015-68955-REDT).

order to alleviate such an issue and to enhance interoperability of language resources on the Web, a community of language technologies experts and practitioners has started adopting techniques coming from the area of study of linked data (LD). The LD paradigm emerges as a series of best practices and principles for 'exposing, sharing and connecting data on the Web' (Bizer, Heath and Berners-Lee 2011), independently of the domain. These principles state that unique resource identifiers should be used to name things in a way that allows people to look them up, find useful information represented with standard formalisms and discover more things that are linked to those resources. LD emerged in the context of the Semantic Web, an extension of the Web 'in which information is given well-defined meaning, better enabling computers and people to work in cooperation' (Berners-Lee *et al.* 2001). General and domain-specific on-line ontologies (such as the ones we revisit in this survey) provide this well-defined meaning on the Semantic Web and are used to represent the data that will be linked to other data. Following this line, an ontology is properly defined as a 'formal, explicit specification of a shared conceptualisation' (Studer, Benjamins and Fensel 1998). This definition refers to an ontology formalised as a machine-readable, explicit description of a domain common to a group of domain experts. Thanks to the semantics encoded in ontologies, LD is transforming the Web into a vast cloud of interlinked information (commonly referred to as the 'Web of Data') in which resources are linked across datasets and sites, and where facts and related knowledge are available for consumption by advanced, knowledge-based software agents as well as by humans through suitable interfaces.

The Semantic Web (and hence, LD too) has as formal backbone the Resource Description Framework (RDF),[1] which describes the data through statements in the form of triples *subject – predicate – object*. This simple model allows structured and semi-structured data to be mixed, exposed and shared across different applications. These triples can be serialised in different formats such as Turtle[2] or JSON-LD.[3] The semantic layer that ontologies provide is defined using the Ontology Web Language (OWL)[4] and other formalisms and schemes such as RDF Schema.[5]

The following example (in Turtle syntax), taken from DBpedia,[6] is an RDF statement about Francisco de Goya (a famous Spanish painter) representing that 'Francisco de Goya was born in Fuendetodos':[7]

```
dbr:Francisco_Goya dbo:birthPlace dbr:Fuendetodos.
```

Notice that since another triple in DBpedia states that

[1] `http://www.w3.org/RDF/`
[2] `http://www.w3.org/TR/turtle/`
[3] `http://www.w3.org/TR/json-ld/`
[4] `https://www.w3.org/OWL/`
[5] `https://www.w3.org/TR/rdf-schema/`
[6] `http://dbpedia.org`
[7] In this example, *dbr* and *dbo* are the prefixes or shorthands for the namespaces http://dbpedia.org/resource/ and http://dbpedia.org/ontology/, respectively. These namespaces are the domains where the elements `Francisco_de_Goya` and `Fuendetodos` on the one hand, and the property `birthPlace` on the other, are uniquely identified.

```
dbr:Fuendetodos dbo:country dbr:Spain.
```

a semantic reasoner might infer that Goya was born in Spain even though this information is not explicitly declared in the data.

In the previous example, knowledge of some 'real live' entities (Goya, country, birth place, ...) has been described in RDF. However, the resources that RDF can describe and link can be literally anything, which means also linguistic information. For instance, the grammatical properties of the word 'country' can be represented in RDF as well. In the following example, we state that a certain resource (*apertiumEN:country-n-en*) is a lexical entry and has 'country' as written form and *noun* as part of speech:[8]

```
apertiumEN:country-n-en a lemon:LexicalEntry ;
        lemon:form [lemon:writtenRep "country"@en] ;
        lexinfo:partOfSpeech lexinfo:noun .
```

This example (encoded in compact Turtle syntax) illustrates also how information coming from different datasets (i.e., Apertium RDF[9] and LexInfo[10]), available in different Web spaces and built by different people for different purposes, have been interlinked in a rather straightforward way.

Linking linguistic information from heterogeneous sources that contain complementary information allows to create a graph of language resources that can be consumed by applications to perform NLP tasks or assist linguists in their research. Indeed, LD is increasingly being adopted by the computational linguistics and the digital humanities communities (Declerck *et al.* 2012; Chiarcos, Nordhoff, and Hellmann 2012c; Chiarcos *et al.* 2013; Hellmann *et al.* 2013; Flati and Navigli 2014; McCrae, Fellbaum and Cimiano 2014; El Maarouf *et al.* 2015; González-Blanco, del Rio and Martínez Cantón 2016; Villegas *et al.* 2016) and an extensive number of efforts are now devoted towards the conversion of language resources to RDF. These works apply LLD for language documentation efforts (Nordhoff 2012) and to facilitate dialectal and etymological studies (Chiarcos and Sukhareva 2014; Abromeit *et al.* 2016), lexical analysis (Khan, Boschetti and Frontini 2014), morphological studies (Klimek *et al.* 2016) as well as querying and reuse of lexicographical data (Declerck, Wandl-Vogt and Mörth 2015; Bosque-Gil *et al.* 2016a), among numerous examples. Specifically in relation to NLP tasks, the possibility of querying graphs of distributed sources, integrating and exchanging annotations between different NLP tools within a pipeline, as well as automatically enriching resources with complementary information are especially highlighted advantages that LD offers (Chiarcos, Hellman and Nordhoff 2012a, 2012b; Chiarcos *et al.* 2013; Hellmann *et al.* 2013; Chiarcos and Sukhareva 2015, just to name a few). These applications are possible thanks to the use of shared conceptualisations

[8] Where *apertiumEN*, *lemon*, and *lexinfo* represent the namespaces `http://linguistic.linkeddata.es/id/apertium/lexiconEN/`, `http://www.lemon-model.net/lemon#` and `http://www.lexinfo.net/ontology/2.0/lexinfo#` respectively.

[9] `http://linguistic.linkeddata.es/apertium/`

[10] `http://www.lexinfo.net`

and terminologies, such as the ones we revisit throughout this paper, and common formats.

Given these advantages and the potential of LD and, in particular, LLD, these last years have seen a wide range of contributions to the literature in LLD which aim at the publication of linguistic information on the Web following LD best practices. This has led to the emergence of the so-called Linguistic Linked Open Data (LLOD) cloud,[11] which constitutes a growing ecosystem of semantically connected linguistic data on the Web. The LLOD cloud is a community effort launched by The Open Knowledge Foundation's Working Group on Open Data in Linguistics (OWLG)[12] (Chiarcos, Hellmann and Nordhoff 2012a; McCrae *et al.* 2016) as a first step to bridge the gap between the advances in language technologies, and linguistics in general, and those taking place in the Semantic Web and artificial intelligence communities. Its main goal is to promote and track the use of LD in linguistics and facilitate the access to available language resources.

In addition, some recent advancements in LLD have also been driven by the activities developed within the framework of international projects such as LIDER[13] or FREME[14], among others. Workshops, datathons and conferences such as the Multilingual Semantic Web Workshop,[15] the Linked Data in Linguistics Workshop,[16] the Workshop on Knowledge Extraction and Knowledge Integration,[17] the Summer Datathon on Linguistic Linked Open Data,[18] the Conference on Language, Data and Knowledge,[19] the NLP&DBpedia Workshop Series,[20] among other initiatives, have encouraged interdisciplinary contributions and community gathering, and provide a perfect scenario to establish new collaborations along these lines of work.

In order to support the representation of the range of linguistic information contained in language resources, experts in the field have been developing numerous *models*, extensions and category registries to account for general and fine-grained linguistic description. A linguistic model provides the ontology entities (classes, properties and individuals) needed to represent the linguistic information of a language resource and to account for its structure and internal/external connections (e.g., lexical entry, syntactic frame, morph, root and antonym). Such models will be used to represent the linguistic facts contained in language resources. For example, defining what a synonym relation is belongs to the model, while stating that *car* and *auto* are synonyms belongs to the lexicon and is the result of a particular instantiation of such a model.

In general, the models presented in the literature are complementary to one another, but sometimes they show some overlap among them. Indeed, authors have

[11] http://linguistic-lod.org/llod-cloud
[12] http://linguistics.okfn.org/
[13] http://lider-project.eu/
[14] http://www.freme-project.eu/
[15] http://msw4.insight-centre.org/
[16] http://ldl2018.linguistic-lod.org/
[17] http://keki2016.linguistic-lod.org/
[18] http://datathon2017.retele.linkeddata.es/
[19] http://ldk2017.org/
[20] http://nlpdbpedia2015.wordpress.com/,http://nlpdbpedia2016.wordpress.com/

developed their own *ad-hoc* extensions due to the actual lack of existing models that account for the specific features of the resource they aim to convert, due to the lack of awareness of a partially similar resource, or even due to the difficulty of finding the appropriate documentation. Such a proliferation of models for representing LLD, along with their potential redundancies and their differences in scope, *makes it difficult to a newcomer to LLD to decide to which model to adhere when converting resources to LLD*. The review we are sketching would help, together with other initiatives such as the Linked Open Vocabularies website,[21] in preventing them from missing previous work that could be applicable to their resource.

The process of generating LLD consists of several steps. For a detailed walk-through, we refer the reader to guidelines for multilingual data generation and publication (Gómez-Pérez *et al.* 2013; Vila-Suero *et al.* 2014), as well as to the best practices suggested in W3C Best Practices for Multilingual Linked Open Data community group,[22] and the LIDER project guidelines and reference cards,[23] which cover a range of language resources. These steps for LLD generation can be broadly summarised as follows: (1) source data and source data model analysis, (2) unique resource identifier naming strategy definition, (3) modelling, (4) RDF generation (or conversion to RDF), (5) linking, and (6) publication. The first step, source data and data model analysis, gives already some valuable information on the kind of linguistic description provided in the resource and also guides in the selection of an appropriate LLD model to be reused. Indeed, in the third step, modelling, reusing existent ontologies and vocabularies is recommended (Gómez-Pérez *et al.* 2013). This involves a careful analysis of the models available out there, and it is at this point where this survey aims to offer support.

Even though summaries and overviews of the work towards the transformation of language resources to LLD are available (Hellmann 2013; McCrae *et al.* 2016), to the best of our knowledge, there still lacks a comprehensive review of the *state of the art on the models or ontologies that support the representation of linguistic data as LD in terms of the nature of the linguistic content they encode*, the ways in which they do so, and what representation challenges still remain ahead.

Our classification of contributions seeks to go over the current literature in LLD with an emphasis on how various efforts approach linguistic knowledge, how they capture it in their proposals and how they relate to one another. Each effort has its own background and tries to satisfy different representation needs, varying from one to another in terms of representation content and granularity. Many of them are complementary to each other and can be used simultaneously to represent different linguistic features. Owing to that, a critical comparative analysis of the considered models would need a preliminary descriptive analysis and classification of the whole landscape of LLD models, which is precisely the main target of this work. Therefore,

[21] Linked Open Vocabularies (LOV) is a catalogue of reusable vocabularies used in the Web of Data. Among other search functions, it allows users to search for ontology terms and filter their search by term type (class, property), vocabulary, etc. `http://lov.okfn.org/`

[22] `http://www.w3.org/community/bpmlod/`

[23] `http://www.lider-project.eu/guidelines`

this review is descriptive rather than critical[24] and focuses on the observed resource-specific problems and general tendencies that we explain and discuss throughout the article as answers to the research questions defined in Section 2. For instance, given two models sharing (1) a common representation goal, (2) an equivalent level of granularity of linguistic description, (3) the type of resources to which they are applied and (4) linking capabilities to other resources, the decision to which to adhere still remains on the LD expert's side. In practice, this may be done on the basis the widespread use of each model, design preferences, linguistic standpoint, etc.

The classification that we provide thus aims to serve as a broad framework for thinking about these models allowing experts to make informed decisions about the models that best cover their needs.

1.2 Classification scheme

The LLOD cloud offers a classification of language resources in LD format in terms of resource type and license. The resource type classification organises resources according to the following categories: corpora, terminologies, thesauri and knowledge bases; lexica and dictionaries; linguistic resource metadata; linguistic data categories, and, last, typological databases. These classes can be roughly grouped into corpora, lexico-conceptual resources and metadata resources (McCrae *et al.* 2016), a grouping which comes close to the classification provided in Meta-Share.[25]

In contrast to the LLOD cloud classification, which focuses on datasets, ours is a classification of the models that are used to represent such datasets semantically. Our proposed classification hinges on three gradually emerging groups into which contributions to the LLD field can be organised, namely (1) *general models* that aim at capturing the main elements of lexical resources, (2) vocabularies developed as *extensions* to models in Group 1 along with ontologies that can be used with them to provide more granularity on *specific layers and dimensions* and (3) *catalogues* of linguistic data categories. A fourth group is included consisting of *other* models, namely, those geared towards the representation of corpora and service-oriented models. These last models are not dealt with in detail here, since these contributions include many structural (e.g., properties to anchor annotations to a given text span) or NLP-based entities and therefore do not strictly focus on the representation of content from the different levels of linguistic description.

Group 1 (general models) encompasses ontology lexica models, such as *lemon* (McCrae *et al.* 2012) and OntoLex,[26] and initiatives such as TELIX (Rubiera *et al.* 2012) or the recent Oxford Global Languages Ontology (OGL) (Parvizi *et al.* 2016).

[24] We leave a critical analysis/comparison of the models as future work, on the basis of the common classification framework that this article will propose.

[25] Meta-Share is a network of repositories that provides a multi-layer infrastructure to share, manage and document language resources, tools and services. `http://www.meta-share.org/` (Desipri *et al.* 2012).

[26] `http://www.w3.org/2016/05/OntoLex/`

Group 2 (model extensions) encompasses two subgroups: *Group 2a* revisits models that address a particular level of linguistic description: phonology, morphology, syntax, semantics, discourse, etc. Some of the models that are covered in our study are the PHOIBLE model for phonological typology (Moran and Wright 2009; Moran 2012), the Multilingual Morpheme Ontology (MMoOn) ontology for morphology (Klimek 2017) and the Predicate Model for Ontologies (PreMOn) model for verbal predicates and each of its extensions, among many others. This group also includes the small extensions developed to account for specificities of a resource due to its grounding on a specific linguistic theory, for instance, *lemonGL* (Khan *et al.* 2013) in the Generative Lexicon Theory (GL) (Pustejovsky *et al.* 1991). *Group 2b* focuses on specific branches of linguistics or related areas, and presents models such as *lemonDIA* (Khan *et al.* 2014) for diachronic linguistics and *lemonet* (Chiarcos and Sukhareva 2014; Abromeit *et al.* 2016) for etymological data, along with models used in contrastive linguistics and lexicography.

Group 3 (catalogues) contains catalogues of linguistic data categories that cover the whole linguistic domain and are widely used in conjunction with models of Group 1 and Group 2a and 2b (concrete examples provided in Section 3.3). In this group, we go over ontologies of linguistic description, category and language identifier registries (General Ontology for Linguistic Description (GOLD) (Farrar and Langendoen 2003), LexInfo (Cimiano *et al.* 2011), ISOcat[27] (Kemps-Snijders *et al.* 2008; Windhouwer and Wright 2012) and Lexvo (de Melo 2015)) and the Ontologies of Linguistic Annotation (OLiA)[28] (Chiarcos and Sukhareva 2015)), among other resources.

The remainder of this paper is organised as follows. Section 2 is devoted to explain the methodology we have followed in order to conduct this review, including the research questions we aim to answer and the literature selection criteria. Section 3 focuses on each group in detail, followed by a discussion in Section 4. The conclusions and the main challenges that we have identified as a result of this study are outlined in Section 5.

2 Methodology

2.1 Research questions and goals

The reviewing method used in this paper (described in detail in the following subsections) is inspired from Dyba, Dingsoyr and Hanssen (2007) and aims to answer the following research questions:

(1) What are the *available models* that allow to represent linguistic information as LD in its different description levels?
(2) What are the main *modelling difficulties* that arise when representing linguistic information as LD and how do different models tackle them?
(3) How have such LD-based linguistic models been *(re)used, adapted and extended*?

[27] http://www.isocat.org
[28] http://acoli.cs.uni-frankfurt.de/resources/olia/

(4) What are the major remaining *challenges* to describe linguistic content in the current LD-based models?

The goal of this review therefore is to provide answers to questions (1)–(4). This review focuses on the current literature in LLD. Other models that do not provide linguistic descriptions as LD are outside of the scope of this paper. Further, techniques, tools and best practices for building ontologies, as the ones used in LLD, are not examined in this survey. On this topic, we refer the interested reader to the literature in ontology engineering (e.g., reviews of methodologies for building ontologies by Corcho, Fernández-López and Gómez-Pérez (2003) and Iqbal *et al.* (2013).

In the remainder of this section, we describe the different steps that we have followed in this process.

2.2 Selection of bibliographic sources

As a starting point, the datasets available through the LLOD cloud page at the time of writing were listed, and their associated papers searched on the Web. Further, the whole proceedings of the following related conferences and workshops were examined: all Multilingual Semantic Web Workshop proceedings available to the date of writing; all Linked Data in Linguistics Workshop issues to the date of writing; OntoLex Workshop 2008, LREC 2014, 2016; ISWC 2014, 2015, 2016; ESWC 2014, 2015; ASIALEX 2013, 2015, 2016; EURALEX 2014, 2016; GLOBALEX 2016, eLex 2013, 2015, and the ISA-LSA Chicago 2015 Conference. In addition, a series of Google Scholar searches involving the combination of the following terms, plus the specific level of linguistic analysis, linguistic theory or branch, were carried out: {linked data, lemon, RDF, linguistic linked data, RDF conversion, migration, (level of analysis), (linguistic theory), (branch of linguistics)}, e.g., '*linguistic linked data GL lemon RDF*'.

2.3 Selection of papers

From the numerous references obtained in the previous step, a first title-based filtering was conducted guided by the following questions, with one positive answer being enough to pass the filter:

(1) Does the title of the paper suggest that it deals with the RDF conversion of a linguistic resource or a modelling of a linguistic resource for the Web of Data?[29]
(2) Does the title of the paper mention lexical resources and refers to their linking, integration or merge?
(3) Does the title of the paper suggest the authors propose a new ontology or model, or an extension to an existing one to encode linguistic information?

[29] In order to know whether a title suggested the topics mentioned in this question, we looked for the following keywords in it: *linking, RDF, Linguistic Linked (Open) Data, mode(l)ling, representing, lemon, Semantic Web, NIF, OWL.*

(4) Does the title of the paper suggest the authors are introducing a new project or initiative related to LLD?

A second filtering process was performed on the remaining papers, abstract-based, to make sure the answers to the previous four questions were indeed correct and the title had not been misleading. An additional crucial question was added to this filter: Is the paper related to LLD in any way? This ruled out works which integrate or link resources but do not rely on the LD paradigm. The third filter consisted in detecting which of the remaining papers reported on the same work or project, but were presented from different perspectives in multiple publications.

Furthermore, if an abstract included references to previous work on which that paper was building upon, those references were searched in Google Scholar and underwent the same filtering process. Likewise, if the abstract included a potential useful keyword that had not been used in the Google Scholar searches (e.g., searches including 'discourse' but not 'information structure'), another round of searches was performed and the results were subject to the filtering processes. This step was repeated throughout the whole review as the papers were read and more references and relevant keywords included in them were taken into consideration. Exceptions were only made in those cases in which (1) a paper served as basis for a model or inspired its creation, but it itself was not directly related to LD and (2) a paper was a (classic) reference paper in linguistics, philosophy or cognitive science.

This review seeks to have a broad scope in order to provide an overview of available models, their coverage and their use by the community, without carrying out a detailed evaluation of the quality of such models. Although the report tries to be exhaustive, we are aware that some contributions may not have been included in the initial selection due to different reasons. For instance, recent publications take some time to appear in Web indexes like Google Scholar. Also the visibility of journal papers may be delayed by the publication process.

2.4 Methodology of analysis

Language can be studied with different units of analysis in focus. The levels of linguistic description are defined in terms of these units of analysis, which, if arranged from smallest to largest, lead to a structure similar to the one shown in Figure 1. This arrangement of the linguistic description levels is widely found in reference books in linguistics and introductory courses and literature in the field (Hayes *et al.* 2013) (traditionally, morphology and syntax would constitute the grammatical level).

Inspired by the thematic analysis description provided by Dixon-Woods *et al.* (2005) and referred to by Dyba *et al.* (2007), this review identifies the set of themes into which current work in LLD may be systematically classified. Specifically, we follow a thematic approach based on the main levels of linguistic description showed in the figure above and commonly addressed in theoretical linguistics, and some of the sub-fields in applied linguistics and areas related to it. By relying on these levels as backbone for our thematic approach, we help the reader to better locate each

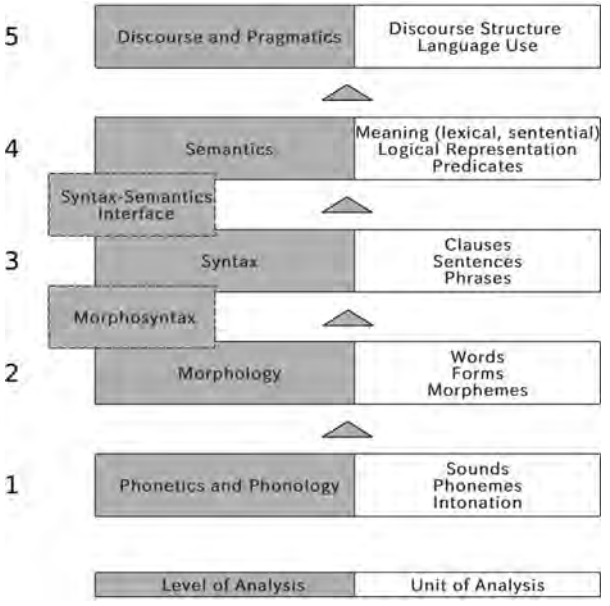

Fig. 1. Levels of linguistic description and their units of analysis.

studied model in a broader context and thus facilitate the decision which one to reuse in a particular project. The main categories we propose follow the same aim.

The papers remaining from the selection and filtering processes outlined above were classified and sub-classified into the groups in which this review is structured. Group 1 models mainly represent information pertaining to the 2nd–4th levels in Figure 1 (counted bottom-up), whereas Group 3 models include elements to capture information from all levels and their interfaces, specially morphosyntax. Models belonging to Group 2a are rather focused on one level (e.g. morphology, (lexical) semantics), as opposed to models in Group 2b, which address aspects that either touch upon different levels or are not easily covered by them. We analyse and compare the contributions in each of these groups in the following sections.

Notice that this survey devotes more space to some models with respect to others, in particular, those concerned with the representation of (lexico-)semantic information. The main reason is that such a kind of models gather more interest in the LLD community, therefore needing a more thorough description. Furthermore, these efforts (aimed at the representation of the lexicon-ontology interface, the lexico-grammar interface and word meaning in general) are more numerous than those addressing phonology, morphology or, in applied linguistics, terminology or typology. This makes the semantics section of the survey longer than other sections. On the other hand, each work either addresses different, concrete aspects in the representation of lexico-semantic information (e.g., from a specific theory) not covered by the others, proposes new ontologies for this kind of information, or highlights gaps in other models. Since we do not find this scenario in other levels or areas of Group 2 or Group 3 yet (e.g., discourse and diachronic linguistics), consisting of works with different goals in mind (exceptions to this would be lexicography and

etymology), one of our aims was to highlight this heterogeneity. This allows us to detect potential shared aspects, even though some of these models have not been reused in later contributions and others are still on-going work.

3 Models for the representation of language resources as LD

This section provides an overview of the main current models to represent linguistic content as LD. Four main groups are distinguished: models to represent the main elements of lexical resources (Section 3.1), ontologies and extensions to such models and *ad-hoc* vocabularies to encode content from different levels of linguistic analysis, linguistic theories, or other disciplines related to linguistics (Section 3.2), linguistic terminology registries and OLiA and description (Section 3.3); and other models, particularly those concerned with the representation of corpora and service-oriented data. Section 3.1 is not intended to go through schemes and frameworks to represent linguistic annotations in XML that fall outside the scope of LD (e.g., the Linguistic Annotation Framework (Ide and Romary 2004)). We limit ourselves to mention those initiatives that were the most influential in the development of current models for LLD and the most used formats in resources that are nowadays being converted to RDF.

3.1 Group 1: Models to represent the main elements of lexical resources

3.1.1 Previous work

The idea of structuring a lexical resource as a graph is not exclusive to LD-based representations, but goes back to initiatives such as WordNet (Fellbaum 1998, 2005) and Polguère's lexical systems (Polguère 2006, 2014) among others (e.g., ConceptNet (Liu and Singh 2004)). *The Lexicon Graph Model* (Trippel 2006), for instance, addresses the combination of diverse lexica from different sources with disparate structures (micro, macro and mesostructures), which, however, contain information that makes them complementary to one another. The author proposes a graph of lexical items as nodes and typed edges, where the original lexica are sub-graphs, and the information can be accessed through various entry points corresponding to the elements of the entry such as the orthography or the spelling. **Spohr** (Spohr 2012), inspired also by the Lexicon Graph Model, turns to OWL and RDF to represent a *multifunctional lexical resource*. Here, *multifunctional* is understood as both serving a range of NLP applications and, from the perspective of functional lexicography, modelling lexical entries as dynamic entries generated and presented to the user depending on their profile and situation needs.

In the development of resources for NLP and outside the context of LD, the *Lexical Markup Framework* (LMF) (Francopoulo *et al.* 2006) was defined to serve as standard for the creation of lexical resources, such as machine readable dictionaries, to allow for their exchange. LMF comprises a core package and several extensions to represent morphological, syntactic and multilingual information. Its core describes the basic structure of a lexical entry, and, together with its extensions, it inspired part of the core of the Lexicon Model for Ontologies (*lemon*) (McCrae *et al.*

2012), designed in RDF and nowadays a *de facto* standard for converting language resources to LLD. *lemon* inherits some classes and properties from LMF, adapts others to the concept of semantics by reference (Cimiano *et al.* 2011) and to the use of external ontologies for linguistic description, and proposes new classes as well (McCrae *et al.* 2010). We will describe *lemon* in detail in Section 3.1. In the context of the ISO's Technical Committee TC37 SC4,[30] we briefly refer also to the *Linguistic Annotation Framework* (Ide and Romary 2004) and its XML serialisation, the *Graph Annotation Format* (Ide and Suderman 2007), the *Morpho-syntactic Annotation Framework* (de la Clergerie and Clément 2005) and the *Syntactic Annotation Framework* (Declerck 2006).

In addition to LMF, other formats suggested to deal with the syntactic heterogeneity in linguistic resources for NLP are *TEI* (Sperberg-McQueen and Burnard 1994) and *Term Base eXchange* (ISO DIS 30042, 2007) for terminologies.

3.1.2 Linked Data-based models

The following models, in contrast to the ones mentioned in the previous section, were developed in the context of the Semantic Web and the LD initiative.

Simple Knowledge Organisation System. The Simple Knowledge Organisation System or SKOS (Miles and Bechhofer 2009) is a W3C Recommendation[31] that is widely used to represent thesauri and taxonomies (Food and Agriculture Thesaurus AGROVOC,[32] the GEMET thesaurus of the European Environment Agency,[33] the Multilingual Thesaurus of the EU (EUROVOC),[34] to name a few). SKOS comprises a series of specifications that provide guidelines to model concepts, concept schemes, hierarchical (e.g., `skos:broader`) and associative (e.g., `skos:related`) semantic relations between concepts, as well as preferred and alternative (lexical) labels for concepts (e.g., `skos:prefLabel`), represented as strings. SKOS-XL is the extension of this model for labels (SKOS eXtension for Labels).[35] This extension allows to reify lexical labels of resources and provides properties to relate them (e.g., an acronym relation can be a subproperty of a label relation) or to indicate what type of label it is (preferred, alternative, etc.). However, and in contrast to the models below, SKOS-XL does not support the linguistic description of such labels: their phonetic, grammatical and semantic features, or how the different labels relate to one another. Importantly also, SKOS-XL labels are not described as 'linguistic objects', in contrast to some of the models presented below.[36]

[30] `http://www.iso.org/committee/297592.html`
[31] `https://www.w3.org/TR/skos-reference/`
[32] `http://aims.fao.org/standards/agrovoc/linked-open-data`
[33] `http://www.eionet.europa.eu/gemet`
[34] `http://eurovoc.europa.eu/drupal/`
[35] `http://www.w3.org/TR/skos-reference/skos-xl.html`
[36] See `http://www.w3.org/2016/05/ontolex/#skos-xl`

lemon and OntoLex. *lemon* emerges from a combination, review and extension of prior models such as LingInfo (Buitelaar *et al.* 2006), LexOnto (Cimiano *et al.* 2007), LIR (Montiel-Ponsoda *et al.* 2011) and SKOS. Its successor, OntoLex, is the result of opening *lemon* to the community under the umbrella of the W3C Ontology-Lexica community group, in order to extend and formally modularise it. Both *lemon* and OntoLex belong to the group of ontology lexica models, i.e., models for the lexicalisation of an ontology. The criteria guiding the development of *lemon* and OntoLex are conciseness, descriptiveness (not prescriptiveness), modularity and RDF-native design. These models were actually not devised, in their origin, to represent lexica as LLD but to ground a given ontology linguistically (McCrae *et al.* 2010). Despite this, they are increasingly being used to convert lexica and other linguistic resources to LD and have become a *de facto* standard to represent and interchange lexical data in the Semantic Web.

Since these models aim to lexicalise an ontology, the conceptual layer (the ontology) is kept separate from the lexical (and linguistic information-bearing, in general) layer, being the class `LexicalSense` the bridge between a lexical entry (`LexicalEntry`) and its meaning (the concept in the ontology). This separation between the lexical and ontological layer is also implemented in previous models such as Senso Comune (Oltramari *et al.* 2008), which follows the Descriptive Ontology for Linguistic and Cognitive Engineering Upper Ontology (Gangemi *et al.* 2002) for its reference ontology. *lemon* comprises five modules: (1) the core, for the representation of grammatical, (basic) morphological and semantic information, and the (2) lexical and terminological variation, (3) phrase structure, (4) syntactic frames and (5) morphological variation modules. OntoLex also consists of a core, which describes lexical entries, their forms and their mapping to the denoted meaning in an ontology, and a series of modules to represent the syntax-semantics interface (synsem), variations and translation (vartrans), the internal structure of an entry (decomp) and linguistic metadata (lime). Section 3.2 presents the extensions to *lemon* and OntoLex that have been proposed in the literature.

The Oxford Global Languages Ontology (OGL). With the conversion of lexicographically rich resources to RDF new modelling problems arise, and the lack of mechanisms to represent certain lexicographic annotations moves authors to design extensions to OntoLex (Section 3.1) or, recently, to the creation of a new ontology, partially from scratch, targeted at the conversion of lexicographic resources. The *OGL* (Parvizi *et al.* 2016) has been developed to model and integrate multilingual linguistic data from Oxford Dictionaries. It includes elements to account for a range of information found in dictionaries, from inflected forms to semantic relations, pragmatic features and etymological data.

The approach followed in OGL is nonetheless not focused on the reuse of existing vocabularies through their extension (in contrast to models presented in the next subsection), but on the creation of a new vocabulary designed with this concrete aim to ensure that the dictionary representation requisites are met. The ontology allows linkage with *lemon* content and ontologies of linguistic description. The emphasis of the approach is set on representing grammatical information with cross-linguistic

validity and on maintaining grammar traditions in different languages as key points. Although some of the modelling decisions differ from those followed in *lemon*-based approaches and the proposed classes are differently defined (e.g., Lexical Entry in *lemon* vs. Lexical Entry in OGL, the use of forms in OGL to represent dictionary headwords, etc.), other aspects are common with *lemon* and are also modelling practices that are spreading throughout the community. Examples of these are the treatment of the given translations of a dictionary entry as lexical entries (or headwords) of a dictionary in the target language (cf. Bosque-Gil *et al.* 2016a; Gracia *et al.* 2018) or the exploitation of translation relations among senses to obtain indirect translations through a pivot language (Villegas *et al.* 2016; Gracia *et al.* 2018).

TELIX. Last, the TELIX model for the representation of linguistic annotations (Rubiera *et al.* 2012) takes a slightly different approach from the one followed in *lemon*, OntoLex and the OGL ontology. Just as corpus oriented models (Section 3.4), it focuses on representing linguistic annotations in corpora, but it also proposes means to cover the lexical information about word tokens, for which some corpus-oriented models rely on *lemon*, used along with linguistic data category registries (see Section 3.3). TELIX is presented as an annotation-oriented lightweight ontology that refines SKOS-XL, and, in fact, models lexical entries as `skosxl:Labels` that stand in opposition to their occurrences, `telix:LabelOccurrence` (word forms in a text). In contrast to OntoLex and the OGL ontology, the lexical sense is not one of the core elements of the model, i.e., `skosxl:Labels` and occurrences directly point to ontology entities, and word forms can be linked to a range of linguistic information through custom properties, classes and values that cover morphosyntax, syntax and discourse concepts, among other aspects.

Figures 2–4 show the different representations of the lexical entry *book* according to OntoLex, the OGL ontology and TELIX, respectively.[37] Example 1 provides the RDF Turtle serialisation of the entry shown in Figure 2 and also illustrates the subject–predicate–object model that was introduced earlier. The *book* example is inspired by examples provided in the literature (Rubiera *et al.* 2012; Parvizi *et al.* 2016) and the OntoLex Specification, which in turn have been adapted for comparative purposes. As shown in Figure 3, in OGL Ontology, the word *book* is treated as a form of both the verb *book* and the noun *book*, i.e., forms can be shared by different lexical entries, which are part-of-speech dependent. The inflection of *book* in plural form, *books*, is therefore modelled as a lexical entry as well. Interestingly, the element `Entry` refers to the dictionary entry from which the information concerning the lexical entry is extracted, acting then as provenance. In OntoLex (Figure 2), in contrast, forms are not shared by lexical entries, and inflectional word forms are viewed as forms of the lexical entry. In OntoLex, there is no *entry* element to track the provenance, but it relies on properties available in other vocabularies, such as

[37] Throughout the survey, we adopt the following graphical representation conventions: boxes indicate instances (actual data), with the shaded section being the type of that instance according to a vocabulary. Entities without namespace in a figure have the namespace of the illustrated model.

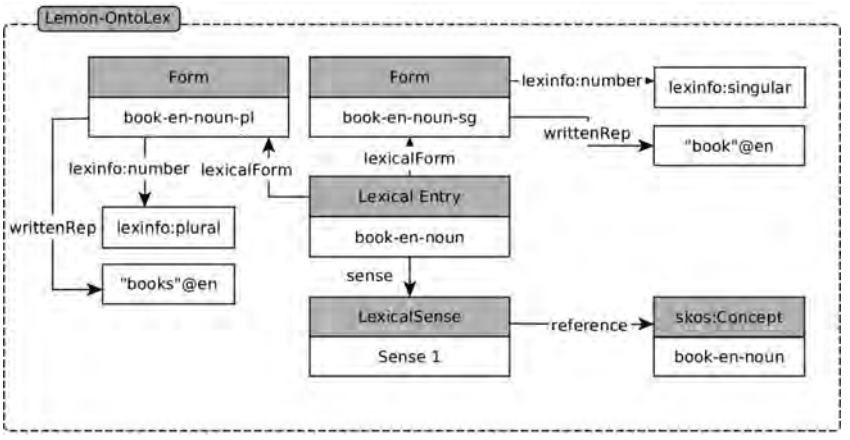

Fig. 2. The lexical entry *book* modelled with OntoLex.

`dc:source`, to that aim. Last, since TELIX is oriented towards the representation of annotations, Figure 4 shows the modelling of an occurrence of the word *books* in a sentence (see Rubiera *et al.* (2012) for more details). The `LabelOccurrence` class is used to model tokens and their morphosyntactic properties, whereas SKOS labels act as the types of those tokens and link to their meaning as defined in an ontology.

Example 1. Turtle RDF serialisation of the example in Figure 2

```
@prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#> .
@prefix lexinfo: <http://www.lexinfo.net/ontology/2.0/lexinfo#>.
@prefix : <http://www.example/ns#>.

:book-en-noun a ontolex:LexicalEntry .
       lexinfo:partOfSpeech lexinfo:noun ;
       ontolex:lexicalForm :book-en-noun-sg, :book-en-noun-pl .
:book-en-noun-sg a ontolex:Form ;
       ontolex:writtenRep "book"@en ;
       lexinfo:number lexinfo:singular .

:book-en-noun-pl a ontolex:Form;
       ontolex:writtenRep "books"@en ;
       lexinfo:number lexinfo:plural .
```

Table 1 summarises the models mentioned in this group, along with the domain in which they have been developed or are commonly used, and the type of resources to which they can be applied.

3.2 Group 2: Level and theory-specific models and extensions

In addition to the models that describe the structure of a lexical entry, its relations to other elements in the lexicon or to concepts in an ontology, we find a series of contributions which present extensions to the models mentioned above and new ontologies to capture, with a high level of granularity, content from a specific level

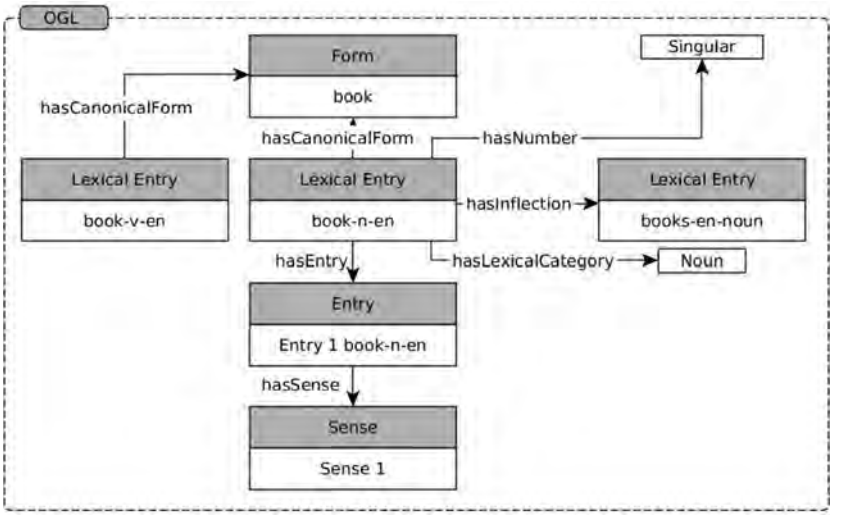

Fig. 3. The lexical entry *book* modelled with the OGL ontology.

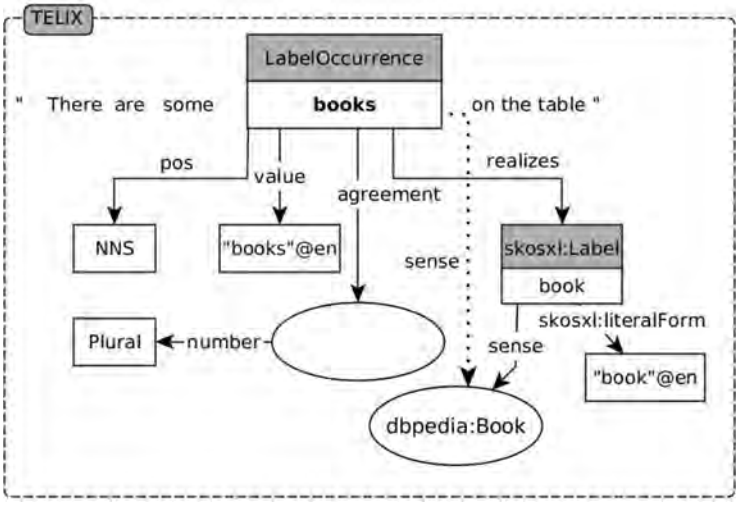

Fig. 4. The lexical entry *book* modelled with TELIX.

of linguistic analysis and notions coming from branches or sub-branches related to linguistics (lexicography, comparative linguistics, historical linguistics, terminology, etc.) and linguistic theories.

3.2.1 Models for different levels of linguistic analysis

In the following, we go over the ontologies and extensions to the models presented in Section 3.1 developed to address linguistic content from each of these levels and their interfaces following the order presented in Figure 1.

Table 1. *Group 1: Models to represent the main elements of lexical resources*

Model	Related area	Applicable to
MLR model (Spohr 2012)	Lexicography	Dictionaries and lexica
SKOS (Miles and Bechhofer 2009)	Terminology	Terminologies, thesauri
lemon (McCrae *et al.* 2012)	Ontology engineering, NLP, lexicography	Ontologies and lexica, corpora (with a corpus model)
OntoLex[a]	Ontology engineering, NLP, lexicography	Ontologies and lexica, corpora (with a corpus model)
OGL (Parvizi *et al.* 2016)	Lexicography	Dictionaries
TELIX (Rubiera *et al.* 2012)	NLP	Corpora

`http://www.w3.org/2016/05/ontolex/`

Phonetics and phonology. Phonetics and phonology remain two areas with relatively low coverage in the LLOD cloud. *PHOIBLE* (Moran and Wright 2009; Moran 2012) was proposed for the creation of a knowledge base for phonological typology that integrates content from several segment inventory databases into one interoperable dataset. A model is developed to associate segments to the languages in which they occur (e.g., `hasSegment`), features to segments (`hasFeature`) and metadata and bibliographic information (sources) to each segment inventory. The GOLD (Farrar and Langendoen 2003), on the other hand, contains classes to encode phonetic, specifically, articulatory features (e.g., `SupraLaryngealProperty`), and phonological information, but to the best of the authors' knowledge, there is no instantiation of these concepts in phonetic datasets yet.

Morphology. In addition to the classes, properties and individuals available in linguistic data category registries (Section 3.3), MMoOn[38] (Klimek *et al.* 2016; Klimek 2017) was recently made available with the specific aim of representing lexemes, word forms, morphological patterns and any other morphological information at the sub-word level relevant in morphology. The ontology, with classes such as `mmoon:Morph`, `mmoon:Morpheme` and `mmoon:Meaning` and links to ontologies of linguistic description and annotation, allows for the creation of morpheme inventories for any inflectional language through language specific schema levels that extend the core of the ontology. Previously, the *lemon* model included a morphology module (McCrae *et al.* 2010) intended to capture morphological patterns and rules that would prevent the explosion of instantiated `lemon:Forms` in case of languages of rich synthetic and polysynthetic morphology. Classes such as `lemon:MorphPattern`, `lemon:MorphTransform` and properties as `lemon:rule` and `lemon:generates` allow to encode the morphological pattern the lexical entry shows and the set of transformations that would systematically generate a series

[38] `http://mmoon.org/`

of inflected forms of that lexical entry. This module is now available as the *Lemon Inflectional Agglutinative Morphology* ontology.[39]

Syntax. The syntax domain is mainly addressed by the annotations schemes present in linguistic data category registries, which are covered in detail in Section 3.3. At the time of writing, LLD resources that instantiate models covering this layer are still scarce except from annotated corpora such as the NEGRA Corpus (Skut *et al*. 1998; Chiarcos, Hellmann and Nordhoff 2012b).

Syntax-Semantics interface. The representation of the interface between syntax and semantics is receiving much attention by the community. The newly developed PreMOn (Corcoglioniti *et al*. 2016) and its extension to model NomBank,[40] PropBank,[41] VerbNet,[42] and FrameNet[43] are put forward as complements to *lemon* to represent predicate models and their mappings (SemLink[44]) as LD. The model, just as MMoOn for morphology, responds in an OntoLex-compliant fashion to the lack of a single ontological model in RDF/OWL to represent these resources in an homogeneous way. The proposed extensions for each of the resources are intended to encode the particular features of a predicate model, whereas the PreMOn core ontology, with classes such as `pmo:SemanticClass` and `pmo:SemanticRole` and properties that relate classes and roles, serves both as foundation to achieve semantic interoperability between the extensions and as link to OntoLex and SKOS. An extract of the entry *hesitate* modelled with PreMOn is illustrated in Figure 5. The entry *hesitate* evokes several concepts, represented in this figure as instantiations of semantic classes from VerbNet (`linger-53.1`) and PropBank (`hesitate.01`). Each of these instantiations is linked to the series of thematic roles that the verb selects (in the case of VerbNet) or its numbered arguments (PropBank), lumped together in the same box in this figure for space reasons. There is a conceptualisation mapping (which in the actual data takes two conceptualisations, `co-v-hesitate-vn32-linger-53` and `co-v-hesitate-vn32-linger-53.1`). This mapping allows to establish a semantic role mapping between VerbNet's *hesitate* `agent` role and PropBank's `arg0` for *hesitate*.

FrameNet, along with other lexico-syntactic and lexico-semantic resources (WordNet,[45] VerbNet, OmegaWiki[46] in German and English and Wiktionary[47] in English and German) had previously been converted to *lemon* in *lemonUby* (Eckle-Kohler, McCrae and Chiarcos 2015). The vocabulary *ubyCat* was proposed to extend *lemon* with data structures found in UBY resources (in LMF) and to link the data

[39] `http://lemon-model.net/liam`
[40] `http://nlp.cs.nyu.edu/meyers/NomBank.html`
[41] `http://verbs.colorado.edu/ mpalmer/projects/ace.html`
[42] `http://verbs.colorado.edu/ mpalmer/projects/verbnet.html`
[43] `http://framenet.icsi.berkeley.edu/`
[44] `http://verbs.colorado.edu/semlink/`
[45] `http://wordnet.princeton.edu/`
[46] `http://www.omegawiki.org/`
[47] `http://www.wiktionary.org/`

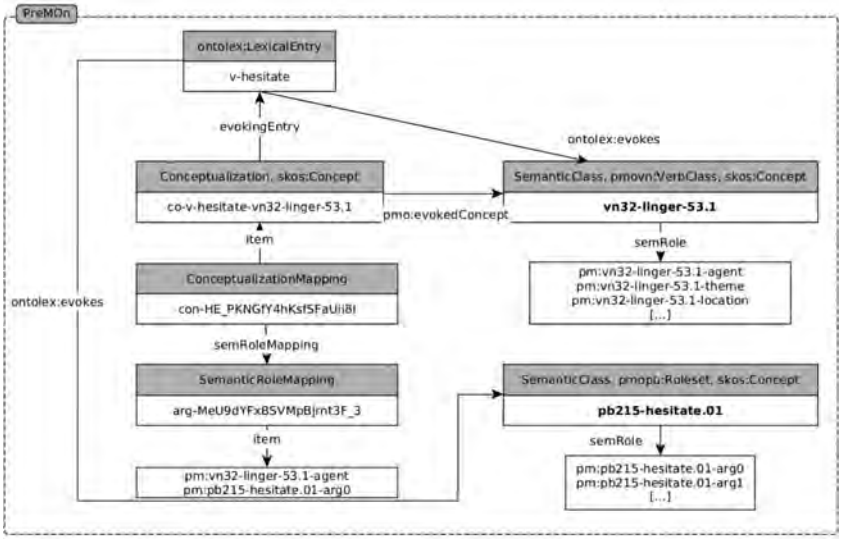

Fig. 5. The lexical entry *hesitate* modelled with PreMOn.

categories in UBY resources to those already defined in current linguistic data category registries. However, the challenge in this contribution was the mapping of LMF elements in UBY resources to *lemon*, rather than the development of a new ontology to cover a specific type of information (Eckle-Kohler *et al.* 2015). Nonetheless, the *ubyCat* extension is reused in the conversions of other works to RDF (e.g., Swedish FrameNet RDF (Borin *et al.* 2014)) and, since it includes equivalences from concepts present in the resources to their *lemon* counterpart, the ontology serves as guide to map concepts to *lemon* as well.

Prior to *lemonUby* and PreMOn, and independently from *lemon*, FrameNet 1.5[48] was converted to RDF through the use of the Semion[49] tool, and a method was proposed to extract knowledge patterns from FrameNet frames (Nuzzolese, Gangemi and Presutti 2011). The OWL metamodel developed in that work sticks strongly to the structure of FrameNet, as opposed to the extension suggested in PreMOn (see Corcoglioniti *et al.* (2016) for more details). An *integration of FrameNet in LOD with the NLP Interchange Format* (a model for the representation of corpora described in Section 3.4) building upon that previous work has also been put forward (Alexiev and Casamayor 2016).

A renown lexical resource for English is the Pattern Dictionary of English Verbs (PDEV),[50] a database of English verbs that gathers the common syntagmatic patterns of use of each verb, expresses them in terms of an ontology of semantic and syntactic categories, and associates them with a meaning, implicature or entailment in order to illustrate verbal behaviour (Hanks 2007). These patterns are extracted through Corpus Patterns Analysis (Hanks 2004), which is based on the Theory of Norms and

[48] `http://framenet.icsi.berkeley.edu/fndrupal/`
[49] `http://stlab-wiki.istc.cnr.it/stlab/Semion`
[50] `http://pdev.org.uk/`

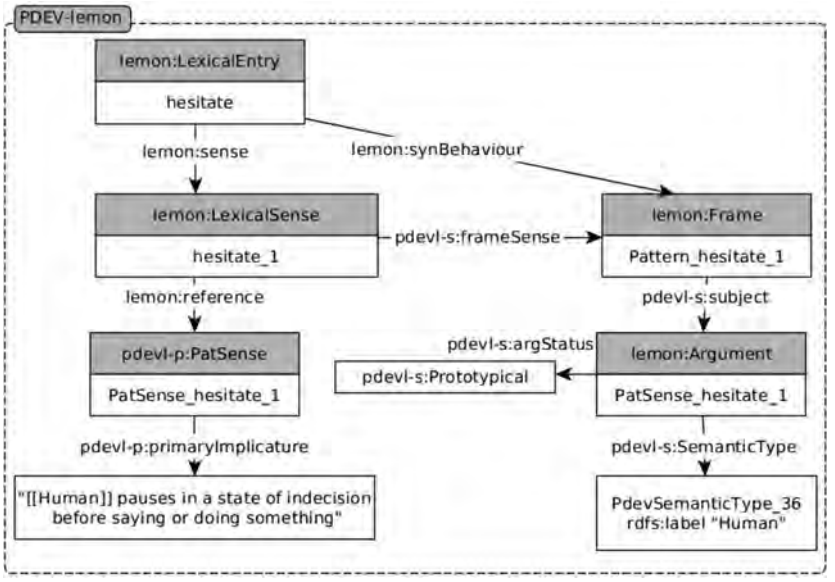

Fig. 6. The lexical entry *hesitate* modelled with PDEV-lemon.

Exploitations (Hanks and Pustejovsky 2005; Hanks 2013). One of the challenges of converting PDEV to *lemon* rests on the fact that the model lacks a way of connecting a syntactic frame to a lexical sense directly: This connection depends on the syntactic arguments, so that the meaning of the syntactic frame will then be inferred from the meaning of the concepts and properties corresponding to the arguments (El Maarouf, Bradbury and Hanks 2014). Cases in which the meaning is not compositionally built from the meaning of the arguments (such as idioms, one of the focus of the Theory of Norms and Exploitations) give rise to modelling problems.

The lack of direct connection from a syntactic frame to a lexical sense in *lemon* and OntoLex (link only through the class `OntoMap`) is also encountered in recent work in the conversion of other verbal inventories, as LVF (Falk and Stein 2016). In *PDEV-lemon*, the authors tackle the problem with the *ad-hoc* classes `:frameSense` and its inverse, included in the ontology suggested as extension to *lemon*, together with other four new ontologies covering pattern domains, pattern registries, semantic types and contextual roles, respectively. Figure 6 shows the modelling of the entry we mentioned earlier, *hesitate*, in PDEV-lemon. Note that, whereas PreMOn captures the semantics of *hesitate* in terms of verb classes and verbal propositions in VerbNet and PropBank, by representing semantic roles, their mappings, and selectional restrictions on the arguments, PDEV-lemon introduces semantic types and prototypicality of arguments in syntactic frames and gives the implicature of the verb in the specific patterns in which it occurs.

The ontologies put forward as part of the PDEV-lemon effort and the *ubyCat* vocabulary introduced above, in conjunction with *lemon*, are also suggested for the modelling of the 'Les Verbes Français' (LVF) (Falk and Stein 2016). LVF is a lexico-syntactic resource in French where each verb sense comprises a semi-formal

semantic description (an operator, which is a combination of primitive predicates and a semantic class), and a series of syntactic constructions which in turn contain information about the syntactic arguments, adjuncts and some features of their realisations. LVF-lemon turns to the *lemon* core to encode morphology, to the UBY ontology for syntactic constructions, to VerbNet for thematic roles, to PDEV-lemon ontologies for the conceptual layer, and provides some *ad-hoc* extensions for syntactic types and operators. The authors argue, however, that verb senses in LVF do not seem to fit into the `LexicalSense` class available in *lemon*: in LVF, they are induced from the semantic descriptions provided (through semantic class and semantic primitives information). The authors of LVF-lemon show that the mapping to *lemon* of the semantic primitives that form up the complex meaning of senses is not a trivial task in itself.

(Lexical) Semantics. In addition to those ontologies developed as part of PDEV-lemon in the context of the Theory of Norms and Exploitations and the efforts towards the migration of FrameNet to RDF (FrameNet as resource conceived from the Frame Semantics approach), we found a series of contributions grounded in other linguistic theories.

The SIMPLE-OWL ontology (Toral and Monachini 2007), for instance, is the OWL version of the SIMPLE ontology (Lenci *et al.* 2000), a language-independent ontology of semantic types, semantic units, and semantic and lexical relations grounded in the extended qualia structure of the GL (Pustejovsky *et al.* 1991). As a result of the PAROLE (Ruimy *et al.* 1998) and SIMPLE projects, which, respectively, aimed at creating a series of lexica and corpora with morphological and syntactic information in numerous European languages and at adding a semantic layer to them, the Parole-Simple datasets were developed. The Parole-Simple DTD has been mapped to *lemon* and to the linguistic data category catalogue LexInfo (Section 3.3) (Villegas and Bel 2015) in the Parole-Simple LexInfo Ontology for the conversion of the Spanish and Catalan portions of the data (Parole-Simple Spanish and Catalan lexica). Such conversions called for the definition of new classes such as `parole:TemplateTop` to encode semantic template following the SIMPLE ontology, e.g., `parole:SymbolicCreation, parole:Human`; and new syntactic and morphosyntactic properties subsumed by linguistic data category catalogues and *lemon* properties in order to represent, for instance, syntactic functions; and new semantic relations (e.g., causativity), among other aspects. Many elements in the DTD were in fact not longer needed once the model was mapped to RDF (Villegas and Bel 2015). The conversion of the Italian Parole-Simple lexica to *lemon* and the extensions to them developed in the context of the CLIPS Italian project (Ruimy *et al.* 2002) (hence, the Parole-Simple-Clips dataset) have also been addressed through the use of the SIMPLE-OWL ontology (Del Gratta *et al.* 2015). An example of this dataset for the entry *esitazione* 'hesitation', is provided in Figure 7. Note that one of the lexical senses of *esitazione* takes as conceptual reference an instantiation of the class `Psych_property` from the Simple Ontology (`USem5510esitazione`), and this instantiation is in turn a type of *comportamento* 'behaviour', an instantiation of the class `Relational_Act` in the model.

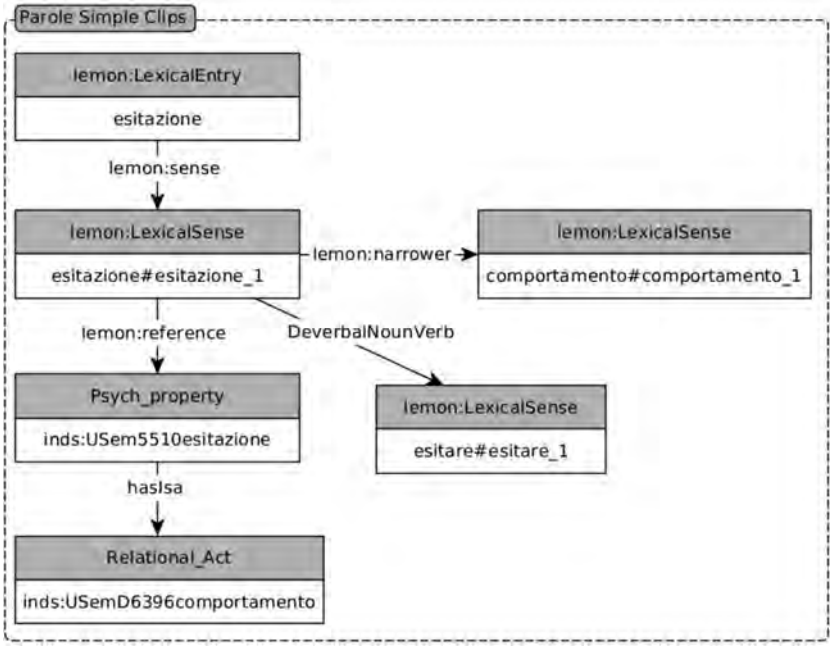

Fig. 7. The lexical entry *hesitate* modelled in Parole-Simple Clips.

Furthermore, *lemonGL* (Khan *et al.* 2013) was presented as an extension to *lemon* to better capture the structure of a lexical sense in terms of the GL qualia roles (formal, constitutive, agentive, and telic). The authors address the limitations of *lemon* to account for logical polysemy Pustejovsky and Bouillon (1995) by suggesting a lexical sense element in *lemon* that is linked to qualia and argument structures (`lemonGL:QualiaStructure`, `lemonGL:hasQualia`), which in turn are associated with roles (e.g., `lemonGL:hasTelic`) in order to represent, following the GL theory, the different aspects of a single meaning.

Models grounded on the Meaning Text Theory (Mel'cuk 1995, 1998) have also been proposed. The recent Lexical Functions Ontology Model (*lexfom*) (Fonseca, Sadat and Lareau 2016) addresses the representation of lexical functions in a *lemon*-compliant way and arises as a mechanism to migrate lexical networks (lexical units and their relations) to LLD. The model provides modules for the representation of lexical function families (e.g., paradigmatic versus syntagmatic), semantic classes of lexical functions; features, constituents and government patterns of lexical functions, and means to link *lemon* lexical senses as the keywords and values of these functions. The model is grounded on the idea that the nodes in a lexical network are already disambiguated, therefore, the connection of lexical items through lexical functions take place at the level of the lexical sense via `vartrans:SenseRelations`.

We also highlight at this point the work towards the migration of WordNet 3.0 to RDF using *lemon*. Prior to *lemon* and OntoLex, WordNet 2.0 had been previously converted to RDF, relying in a custom data model.[51] In the conversion

[51] `http://www.w3.org/TR/wordnet-rdf/`

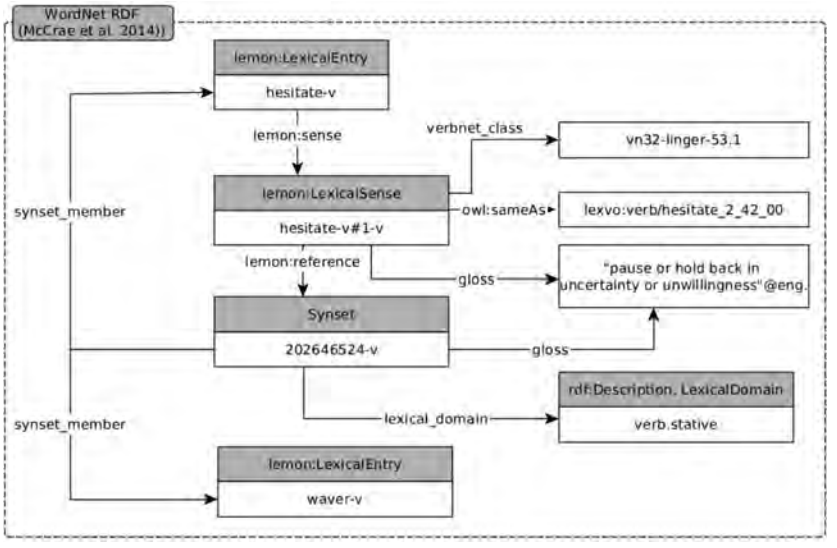

Fig. 8. The lexical entry *hesitate* modelled in WordNet RDF (2014).

from Princeton WordNet to LLD (McCrae *et al.* 2014), WordNet synsets are viewed as the ontological references of *lemon* lexical senses, and new classes (e.g., *Synset* as subclass of `skos:Concept`, *Lexical Domain*, etc.) properties to encode hyponymy, meronymy, entailment, etc., and individuals of lexical domains (e.g., `noun.process`) were defined to account for the lexical information present in WordNet. This is illustrated in Figure 8 with a sense of the entry *hesitate* linked to a synset as reference. Through this concept, we can access other members of that synset as well as information on its lexical domain, among other aspects. This modelling in *lemon* has been recently adapted to OntoLex.[52] Some of the main differences include the use of the class `ontolex:LexicalConcept` for synsets and the mapping of lexical concepts to the Collaborative Interlingual Index (Bond *et al.* 2016).

The Global WordNet Formats to publish WordNets[53] reuse the entities from the WordNet ontology as RDF and JSON-LD, along with the OntoLex vocabulary. Another ontology developed in order to convert a lexico-semantic resource is in fact the EuroWordNet Multilingual Central Repository Ontology,[54] generated from the LMF versions of WordNet Lexica included in the Multilingual Central Repository[55] (Gonzalez-Agirre, Laparra and Rigau 2012), which builds upon the EuroWordNet Top Ontology (Vossen *et al.* 1998). The EuroWordNet MCR Ontology serves as foundational backbone to generate the EuroWordNet *lemon* lexica available in the MCR.

[52] See `http://wordnet-rdf.princeton.edu/about`.
[53] `http://globalwordnet.github.io/schemas/`
[54] `http://github.com/martavillegas/EuroWordNetLemon`
[55] `http://adimen.si.ehu.es/web/MCR`

Semiotics. Before moving into models for the representation of discourse, and still closely related to semantics, we turn to the realm of semiotics, concerned with the study of the meaning of signs in communication systems. The Linguistic Meta-Model (Picca, Gliozzo and Gangemi 2008) enables the representation of knowledge according to different semiotic theories in terms of the semiotic triangle (*symbol – reference – thought of reference*) (Ogden *et al.* 1923; Peirce *et al.* 1958). Its core consists of the classes `Reference`, `Meaning` and `Expression`, each of them with a series of subclasses inspired or derived from Descriptive Ontology for Linguistic and Cognitive Engineering (Gangemi *et al.* 2002), and it is extended by a second module that accounts for specific linguistic references, e.g., `lmm2:ExtensionalReference`. This model serves as basis for the Semiotics Ontology,[56] imported by OntoLex, and includes also pragmatics-inspired elements such as the class `LinguisticAct`.

Discourse and information structure. The OLiA, explained in more detail in Section 3.3, aim to achieve conceptual interoperability among different annotation schemes and external terminology repositories through an intermediate conceptual layer, the OLiA Reference model. The OLiA Discourse extension (Chiarcos 2014) for the Reference Model addresses the information encoded in major discourse annotated corpora available nowadays and incorporates notions from other efforts that revolve around the representation of discourse (such as NERD[57] or the Grounded Annotation Framework[58]). The annotations this extension addresses pertain to discourse structure, coherence relations, coreference and bridging, information structure (topic-focus) and information status (given-new) with classes such as `DiscourseCategory`, `DiscourseRelation` and `DiscourseFeature` and a long list of subclasses and properties that capture the different phenomena. The extension focuses on the annotation schemes of the Rhetorical Structure Tree Bank,[59] grounded on the Rhetorical Structure Theory (Mann and Thompson 1988), OntoNotes (Pradhan *et al.* 2007) and the Penn Discourse Tree Bank[60] and Graph Bank (Wolf and Gibson 2005) (PDTB and PDGB, respectively). One of the author's main points, along with the NLP interoperability advantage and the potential use of the reference model in corpus query systems, is that this extended reference model of OLiA allows to generalise over the Rhetorical Structure Tree Bank and PDTB, which are conceived from different theoretical backgrounds, by providing a shared terminology that separates discourse relations (the central aspect of the PDTB approach) from discourse structural ones (the latter playing an important role in RST). This idea, in turn, is based on the benefits of a common conceptual layer among different schemes to compare, integrate, or revisit them (see Section 3.3).

Table 2 lists the models mentioned in this section, including the level of linguistic description they address, the theory on which they are grounded (if any) and the

[56] `http://www.ontologydesignpatterns.org/cp/owl/semiotics.owl#`
[57] `http://nerd.eurecom.fr/ontology`
[58] `http://groundedannotationframework.org/`
[59] `http://www.isi.edu/ marcu/discourse/Corpora.html`
[60] `http://www.seas.upenn.edu/ pdtb/`

type of resources to which they are applied. The structure of this table differs from Table 1 to reflect the fact that the systems described in this section deal with specific levels of linguistic analysis.

3.2.2 Group 2b: Models for specific branches of applied linguistics and related areas

There exist other kinds of extensions and efforts in the development of models for the representation of language resources as LD presented from the perspective of a sub-discipline of linguistics (not theoretical) or a related research area. These are targeted at addressing gaps in the representation of information common to a specific field of study (e.g., historical linguistics, contrastive linguistics/typology, terminology and lexicography), rather than to a particular level of linguistic analysis. We go through the ontologies and extensions designed for this purpose in the following paragraphs, focusing on historical linguistics, lexicography and cross-linguistic studies.

Historical linguistics. In historical linguistics, and, in particular, diachronic semantics, for instance, the semantic shift in word meanings is an object of study. *lemonDIA* (Khan *et al.* 2014) is developed as an extension to *lemon* to capture the meaning change in a *lemon* lexical entry throughout time as a perdurant entity in order to dodge the unfeasability of adding an extra time argument to RDF triples (Khan, Diaz-Vera and Monachini 2016a). The authors suggest the inclusion of a class `DiachronicShiftObject` with an associated `TimeInterval`, and one or more lexical senses can be linked to it. This approach is further updated with classes such as a perdurant LexicalSense (`LexicalpSense`) to which a Semantic Shift element can be linked, `LexicalDomain`, `NegatedShift` and links to the Time Ontology in OWL[61] for the conversion of historico-philological data (Khan *et al.* 2016a). An interface to create datasets based on *lemonDIA* is also provided (Khan, Bellandi and Monachini 2016b).

In relation to historical linguistics and moving into lexicography, the European Network of e-Lexicography is fostering work towards the conversion of a range of dictionaries, some of which contain historical data, to LLD based on the OntoLex model. To this aim, new classes and properties are defined to capture information which the *lemon* and OntoLex models fall short of covering. This is the case, for instance, of the class `Etymology` and the properties encoding different types of temporal information, in the recent conversion of thirteen dictionaries (dialectal, bilingual, monolingual, historical, etc.) carried out as part of European Network of e-Lexicography (Declerck *et al.* 2015). An extension to *lemon* to represent etymological information of lexical entries has also been proposed (*lemonet*) (Chiarcos and Sukhareva 2014). Recently, a new revisited version of *lemonet* builds upon the properties proposed in the modelling of the etymological WordNet[62] in order to undertake the conversion of the Tower of Babel (Starling)[63] with

[61] `http://www.w3.org/TR/owl-time/`
[62] `http://www1.icsi.berkeley.edu/~demelo/etymwn/`
[63] `http://starling.rinet.ru/`

properties such as `lemonet:cognate` and `lemonet:derivedFrom` as subproperties of `vartrans:lexRel` from the *vartrans* module of OntoLex (Abromeit *et al.* 2016). This last work has been launched in the context of the Linked Open Dictionaries project,[64] which, among other aspects, aims to apply the LD paradigm to develop new methodologies for the research in (historical) linguistics, cross-lingual lexicography and historical sciences.

Lexicography. Structures typically found in dictionaries, be they historical or not, such as the sense and sub-sense hierarchy in an entry, raise problems as well. *lemonLSJ* (Khan *et al.* 2016c) and *polyLemon* (Khan *et al.* 2016a) both emerge as extensions to *lemon* to capture this sense and sub-sense structure in a dictionary entry, the former specifically developed during the migration of the Liddell–Scott Greek–English Lexicon[65] to *lemon*. Both extensions suggest the inclusion of the properties `senseChild` and `senseSibling` to relate senses and their parent senses in the dictionary entry. *Ad-hoc* vocabularies for the conversion of monolingual and bilingual lexica to *lemon* and OntoLex have also been proposed, for instance, for the conversion of K Dictionaries data (Klimek and Brümmer 2015; Bosque-Gil *et al.* 2016a). Such works aim to account for the proprietary XML tags and values whose labels were not present or had an incompatible definition in external linguistic terminology repositories. The conversion from TEI to RDF of dialectal dictionaries of Arabic for their subsequent integration using OntoLex has been addressed as well (Declerck and Mörth 2016). The focus here has been on how the several lexical senses of an entry across different dialectal dictionaries, if gathered in a sense repository and mapped when equivalent, enable the navigation through the different entries and the enrichment of one another.

With regards to the representation of the multilingual aspect in dictionaries, and lexical resources in general, a translation module for *lemon* was proposed (Gracia *et al.* 2014), the *lemon translation module*. This module, reused in resources such as the Apertium RDF series (Gracia *et al.* 2018), inspired later the *vartrans* module of OntoLex. With classes such as `vartrans:Translation` and the properties `vartrans:source` and `vartrans:target`, for example, the *vartrans* module provides mechanisms to describe translation relations between senses of different lexical entries and their directionality.

Prior to the *lemon* translation module, Wiktionary[66] was converted to *lemon* in the DBnary resource (Sérasset 2015; Tchechmedjiev *et al.* 2015) which, at that time, needed the definition of a `db:Translation` class with a series of properties to refer to the target language and the information concerning the source sense and the target sense, etc. Other classes and properties included in DBnary do not attempt to fill a gap in *lemon* but rather to deal with the fact that Wiktionary has its own legacy structure: different lexical entries may occur in the same Wiktionary page, relations might be underspecified or link two senses, others may link a sense to

[64] `http://acoli.cs.uni-frankfurt.de/liodi/home.html`
[65] `http://www.tlg.uci.edu/lsj/`
[66] `http://www.wiktionary.org/`

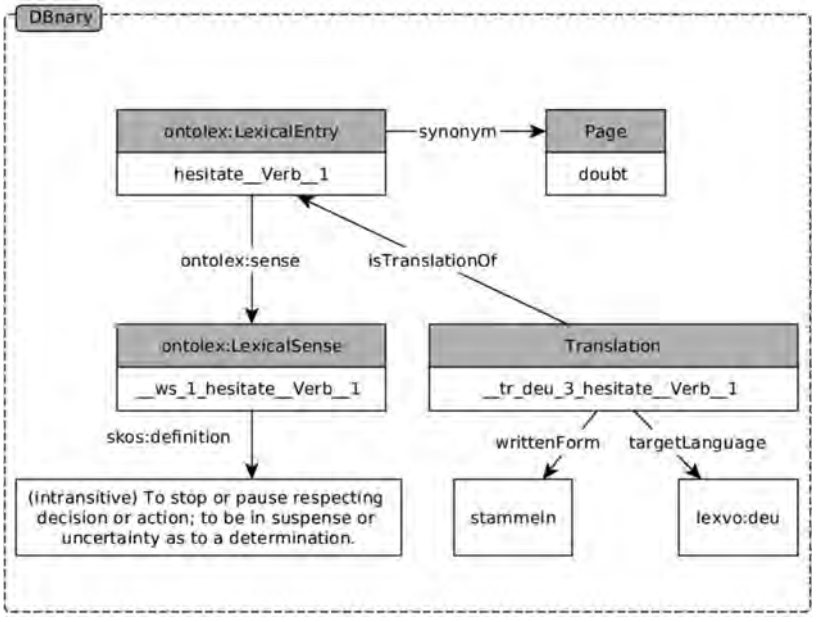

Fig. 9. The lexical entry *hesitate* modelled in DBnary.

a Wiktionary page, etc. Figure 9 illustrates the kind of information that DBnary provides for the entry *hesitate*: translations are represented at the lexical entry level with their own class and a target language. In contrast to WordNet RDF, lexical definitions or glosses are represented at the lexical sense level. Note that DBnary introduces classes such as `db:Page` reflecting the legacy structure of Wiktionary.

Also, in relation to translations, we mention the Panlex Vocabulary (Westphal, Stadler and Pool 2015), developed in the context of the PanLex project to gather translations among lexemes in different languages. PanLex includes both very general purpose classes such as `plx:Definition` or `plx:Expression` (the latter being used for lexemes as lexical entries), and classes to represent languages, their varieties and ISO language codes (`plxo: Language`, `plx:LanguageVariety`).

The problems encountered in the conversion of dictionaries, the majority of which have been highlighted in the definition of the requirements for the OGL (Section 3.1), prove that the mapping of proprietary formats to a standard framework like *lemon* is not always straightforward due to the type of information included in dictionary entries: lexicographic annotations, the structure of the dictionary entry itself and numerous elements that do not have an homologous one in other models and/or external ontologies for linguistic description. *lemon* and OntoLex, in fact, though devised from a lexicographic perspective, were originally intended to lexicalise ontologies, not to model lexicographically rich linguistic resources to RDF, which however is the use to which the vast majority of the community is currently turning. Even so, by using extensions and *ad-hoc* vocabularies, dictionary data are increasingly being converted to LLD, and the benefits of LD for lexicography are currently being explored (Declerck *et al.* 2015; Klimek and Brümmer 2015; Bosque-Gil *et al.* 2016a; Declerck and Mörth 2016; Parvizi *et al.* 2016) etc. At the time

Table 2. *Group 2a: Models for different levels of linguistic analysis*

Model	Level	Linguistic theory/ approach	Applicable to
PHOIBLE Model (Moran and Wright 2009; Moran 2012)	Phonology	Neutral	Phonological databases
MMoOn (Klimek 2017)	Morphology	Neutral	Morphological data
LIAM (McCrae *et al.* 2010)	Morphology	Neutral	Morphological data
PreMOn (Corcoglioniti *et al.* 2016)	Syntax-semantics interface, semantics	Frame Semantics	Predicate models
UbyCat (Eckle-Kohler *et al.* 2015)	Morphosyntax, syntax-semantics interface, semantics	Neutral	Lexico-semantic datasets, lexica
FrameNet Model pNuzzolese *et al.* (2011)	Syntax-semantics interface, Semantics	Frame Semantics	FrameNet
PDEV-lemon (El Maarouf *et al.* 2014)	Syntax-semantics interface, lexical semantics	Theory of Norms and Exploitations	Pattern dictionaries, lexica
LVF-lemon (Falk and Stein 2016)	Syntax-semantics interface, lexical semantics	Dubois and Dubois-Charlier (1997)'s classification, Maurice Gross' distributional grammar	Les Verbes Français Database
SIMPLE OWL Ontology (Toral and Monachini 2007)	Lexical semantics	Generative Lexicon	Lexica, lexico-semantic databases
Parole-Simple 'LexInfo' Ontology' (Villegas and Bel 2015)	Lexical semantics	Generative Lexicon	Lexica, lexico-semantic databases
lemonGL (Khan *et al.* 2013)	Lexical semantics	Generative Lexicon	Lexica, lexico-semantic databases
lexfom (Fonseca *et al.* 2016)	Lexical semantics	Meaning Text Theory	Lexical networks
WordNet RDF Vocabulary (McCrae *et al.* 2014)	Lexical semantics	WordNet	Wordnets, lexico-semantic databases

Table 2. *Continued*

Model	Level	Linguistic theory/ approach	Applicable to
Linguistic Meta-Model (LMM) (Picca *et al.* 2008)	Semiotics	Peirce (1958)	Any linguistic resource
OLiA Discourse (Chiarcos 2014)	Discourse, information structure	Neutral, but covers Rhetorical Structure Theory	Any discourse-annotated ling. resource

of writing, the OntoLex community is discussing a series of best practices for the representation of rich lexicographic resources as LLD as part of a new module.[67]

Typology and cross-linguistic studies. In the line of linguistic typology, language documentation and cross-linguistic studies in general, and in addition to some catalogues of linguistic categories (Lexvo.org and Glottolog/Langdoc, discussed later in Section 3.3) and PHOIBLE (introduced above), we point out the Typological Database System ontologies (Saulwick *et al.* 2005), which, through an architecture similar to OLiA's Reference Model mentioned above, aim to integrate the different theory-specific models of typological databases to allow for cross-database searches. To this aim, an ontology of linguistic typology is developed to serve as a shared vocabulary across the various local models. It encompasses the notions present in all of them (e.g., word order) through concepts that unify the different theory-dependent views, while keeping the theory-specific definitions at the local level. In addition, the Cross-Linguistic Linked Data Project (Forkel 2014) promotes the creation of a LD infrastructure to integrate and publish typological data sets (PHOIBLE, WALS,[68] WOLD,[69] Glottolog,[70] afbo,[71] among others) and suggests to that aim a typological data model that accommodates to the diverse resources.

Terminology. The conversion of resources to LLD has also been addressed in the domain of terminology. In addition to SKOS, the model to represent taxonomies and thesauri introduced at the beginning of Section 3, and its extension SKOS-XL, best practices have been defined for the generation of LLD from terminological resources in the TBX format.[72] This has been the case in the context of the conversion of IATE[73] and the European Migration Network[74] datasets to RDF using OntoLex

[67] http://www.w3.org/community/ontolex/wiki/Lexicography
[68] http://wals.info/
[69] http://wold.clld.org/
[70] http://glottolog.org/
[71] http://datahub.io/km/dataset/clld-afbo
[72] http://www.tbxinfo.net/
[73] http://iate.europa.eu/
[74] http://www.emn.ie/

Table 3. *Group 2b: Models for specific branches of applied linguistics and related areas*

Model	Related area	Applicable to
lemonDIA Khan *et al.* (2014)	Diachronic linguistics	Lexica, lexico-semantic databases
Dictionary extensions by Declerck *et al.* (2015)	Lexicography, historical linguistics, cross-lingual studies, dialectology	Lexica
lemonet Chiarcos and Sukhareva (2014)	Lexicography, historical linguistics, etymology	Etymological lexica, dictionaries and databases
lemonet Abromeit *et al.* (2016)	Lexicography, historical linguistics, etymology	Etymological Lexica, dictionaries and databases
lemonLSJ and polyLemon Khan *et al.* (2016c,a)	Lexicography	Dictionaries
K Dictionaries Vocabulary Klimek and Brümmer (2015)	Lexicography	Global Series K Dictionaries (monolingual)
K Dictionaries Vocabulary Bosque-Gil *et al.* (2016a)	Lexicography	Global Series K Dictionaries (multilingual)
DBnary model Sérasset (2015)	Lexicography, translation	Lexical, legacy data, Wiktionary
PanLex Westphal *et al.* (2015)	Language documentation, translation	PanLex Database
TDS Ontologies Saulwick *et al.* (2005)	Typology, cross-lingual studies	Typological databases
CLLD Data Model Forkel (2014)	Typology, language documentation	Typological databases
LIDER TBX Ontology Cimiano *et al.* (2015)	Terminology	Terminologies in TBX

and a specific vocabulary put forward for that purpose (Lider Term Base eXchange Ontology) (Cimiano *et al.* 2015). This vocabulary encodes header information, reliability codes or information related to transactions, for example. OntoLex alone (together with its modules), has also proved to be rich enough to cover basic terminological multilingual information (e.g., definitions and translations) as the one given in some resources such as the Terminoteca RDF (Bosque-Gil *et al.* 2016b).

Table 3 groups the models explained above, providing the area of linguistics to which they are related and the type of language resource to which they are being applied.

3.3 Group 3: Linguistic data category registries

This subsection dwells on ontologies of linguistic description and linguistic data category registries in general. Most of the models presented above often turn to these categories as a basis to create their custom vocabularies for three main

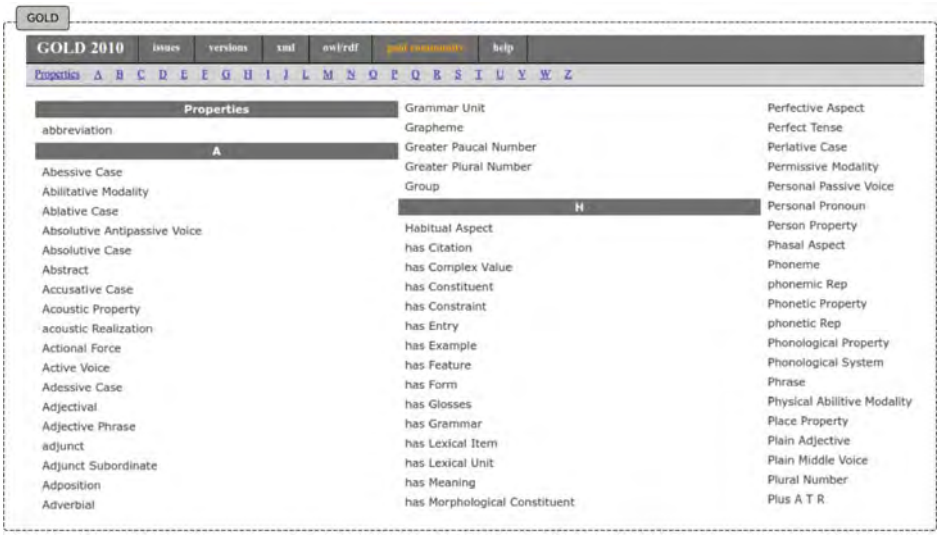

Fig. 10. (Colour online) A small section of the GOLD ontology in its Web interface.

reasons: To comply with the community's view of the domain or the linguistic tradition (Parvizi *et al.* 2016), to establish mappings between their defined elements and the ones in the linguistic terminology ontology (Klimek 2017), or to point to them as external resources (Hellmann *et al.* 2013) that can be used together with the developed model.

Developed from a typological viewpoint, the GOLD (Farrar and Langendoen 2003) constituted the first effort towards a repository of linguistic terminology in the Semantic Web and was developed as a comprehensive ontology that refines the Suggested Upper Merged Ontology[75] with linguistic constructs, basic concepts used in linguistic analysis, grammatical categories and types of linguistic expressions with the aim of serving as a central terminological resource to enable data comparison across languages. In its 2010 version, it addressed the morphosyntax, morphosemantic (e.g., mood, aspect, and tense) and phonetic layers (Figure 10). In contrast to ontologies such as Descriptive Ontology for Linguistic and Cognitive Engineering (Gangemi *et al.* 2002), which were devised for NLP purposes, GOLD is developed for cross-lingual studies and is inspired by typological initiatives such as WALS (Farrar and Langendoen 2003).

We also highlight the ISOcat ISO TC37 Data Category Registry in its adaptation from the XML data model to an RDF model (Windhouwer and Wright 2012) and the Relation Registry RELcat to relate the data categories (Windhouwer 2012). ISOcat is in turn based on ISO 12620:2009,[76] which establishes the specifications of data categories (the data model) and the management of such a registry (Kemps-Snijders *et al.* 2008). Importantly, it is not a formal ontology but a structured collection of terms (Eckle-Kohler *et al.* 2015), and ontological relationships are not

[75] `http://www.adampease.org/OP/`
[76] `http://www.iso.org/standard/37243.html`

contained in the registry, so that RELCat was developed to complement ISOcat in that respect. ISOcat was created initially to address data categories found in terminological databases, which in turn are common throughout annotation frameworks (Eckle-Kohler *et al.* 2015), and it was used along with LMF in the CLARIN infrastructure to encode terminological, morphosyntactic and metadata information. ISOcat was, however, designed in a data-oriented way and the needs of the ISO TC 37 and the CLARIN community turned out to differ after time. The latter advocated a simpler model, easily reusable, focused rather on concepts and without constraints on data types and data categories. This led to the creation of the CLARIN Concept Registry by building upon ISOcat (Eckle-Kohler *et al.* 2015; Shuurman *et al.* 2016).

Another linguistic data category registry is LexInfo (Cimiano *et al.* 2011), specifically developed to address the lexicon-ontology interface and widely used together with *lemon* and OntoLex (see El Maarouf *et al.* (2014), Ehrmann *et al.* (2014), Sérasset (2015), Gracia *et al.* (2018), among others). LexInfo, building on LingInfo (Buitelaar *et al.* 2006) and LexOnto (Cimiano *et al.* 2007), is a lexicon model that tackles the limitations of SKOS, RDF and RDFS in encoding linguistic information associated to the elements of an ontology, and it is grounded on the separation of the conceptual and linguistic layers. This separation lies behind the notion of 'semantics by reference' introduced above. The second version of LexInfo is an extensive ontology of types, values and properties derived partially from ISOcat,[77] and currently its elements capture information from the morphosyntactic (see Figure 11), syntactic, syntactic–semantic, semantic and pragmatic levels of linguistic description.

With an architecture different from LexInfo and ISOcat, the OLiA (Chiarcos and Sukhareva 2015) seek to achieve semantic and syntactic interoperability between current linguistic annotation schemes and external terminologies through the use of a reference model, annotation models (or external reference models) and linking models as mediators of reference and annotation or external ones, e.g., PennTreeBank annotation model and linking model. The idea behind this architecture is to abstract from the specific annotation schemes and to provide a common ontological ground among the different models to allow for the generalisation, comparison and potential revision of the models that capture a range of linguistic phenomena. As of today, they are being reused in multiple conversions of corpora and lexico-syntactic resources (e.g., *lemonUby*) to LLD, and the linguistic content they cover ranges from morphology, morphosyntax, syntax, semantics (partially) and, with the OLiA Discourse extension, to discourse phenomena, including discourse structure, coherence relations, coreference and information structure and status (Chiarcos 2014; Chiarcos, Fäth and Sukhareva 2016b) (Figure 12).

There has also been an effort to develop an ontology in OWL/DL from the Multext East Specifications for corpus morphosyntactic annotation (Chiarcos and Erjavec 2011). This effort provides numerous attributes and values for syntactic and morphosyntactic features of multiple languages along with notes and bibliography,

[77] `http://www.lexinfo.net/ontology/2.0/lexinfo.owl`

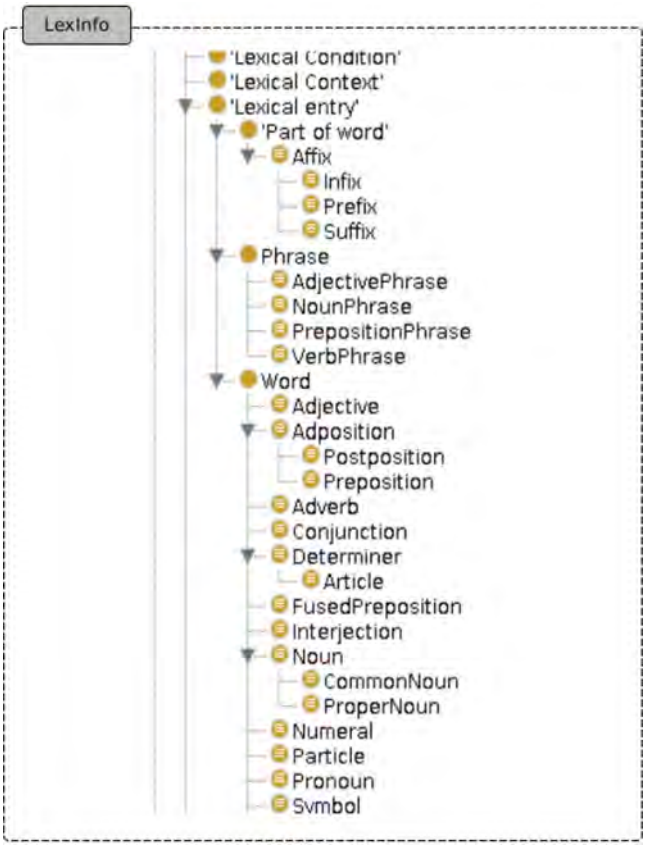

Fig. 11. (Colour online) A small section of the LexInfo ontology in the ontology editor *Protégé*.

etc. Through a rich feature set, the specifications account for a range of properties of languages with rich inflection and different typology.

In addition to the above-mentioned resources and also in relation with typology, there are typological databases such as Glottolog/Langdoc (Nordhoff and Hammarström 2011). This database provides a language family tree in terms of the notion of *languoid*, i.e., a language, dialect or language variety (Good and Hendryx-Parker (2006) as cited in Nordhoff (2012)) and links the information to the bibliographic records attesting it.

Table 4 summarises the information regarding the groups introduced in this subsection. For the sake of comparison of aim and scope, we include the level of linguistic description each catalogue covers and the branches of linguistics which may turn to it when representing their resources as LLD.

3.4 Other models

This section mentions some of the models that do not fit into the groups presented above and which do not address the representation of purely linguistic content, as opposed to the works mentioned so far, but are rather focused on describing

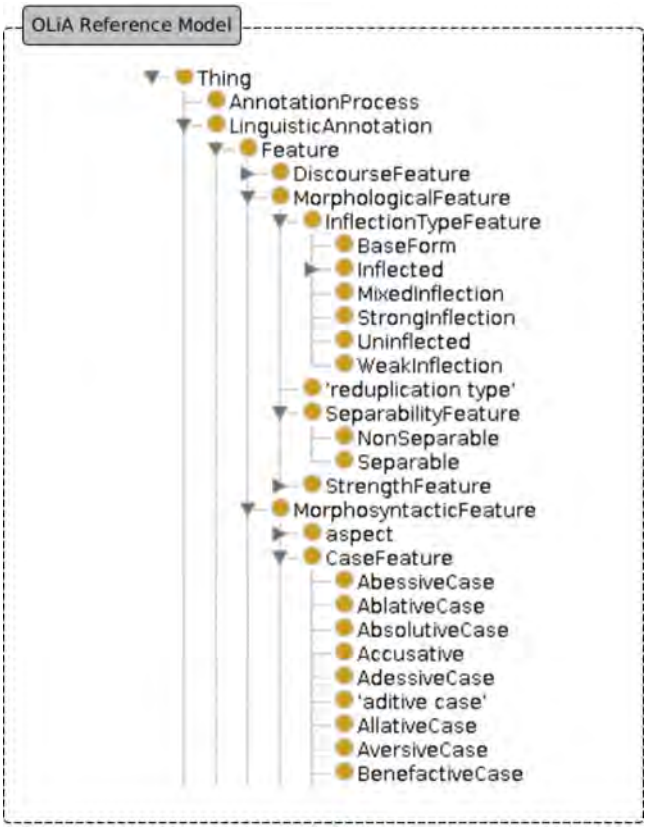

Fig. 12. (Colour online) A small section of the OLiA Refence Model in the ontology editor *Protégé*.

structural aspects of corpora or characteristics of NLP services. Although the analysis of these models falls outside the scope of this paper, it is worth mentioning them in order to enrich the context of this work. In fact, the models that we have reviewed in this survey are quite frequently used in conjunction with the ones that we report in the following paragraphs.

Corpora models. Models for corpora representation and their annotations as LLD focus on the structure of corpora, the documents themselves, the (multi-layer) annotations and the relation between the different annotations, and the text to which they are anchored. These models can often be combined with models to represent lexical resources (e.g., OntoLex) and linguistic data category catalogues (e.g., LexInfo, OLiA). Examples of efforts here are the NLP Interchange Format[78] (Hellmann *et al.* 2013), which can be used with the Open Annotation Data Model,[79] now superseded by the Web Annotation Data Model,[80] to represent corpora and annotations

[78] See ontologies and specifications at `http://persistence.uni-leipzig.org/nlp2rdf/`
[79] `http://www.openannotation.org/spec/core/`
[80] `http://www.w3.org/TR/annotation-model/`

Table 4. *Group 3 models*

Model	Level	Domain
ISOcat Kemps-Snijders *et al.* (2008)	Morphosyntax, syntax, (lexical) semantics	Terminology, lexicography, translation, sign language linguistics
CLARIN CR Shuurman *et al.* (2016)	Morphosyntax, syntax, (lexical) semantics	Terminology, lexicography, translation, sign language linguistics
LexInfo Cimiano *et al.* (2011)	Morphosyntax, syntax, (lexical) semantics, pragmatics	Terminology, lexicography, translation, NLP
OLiA Chiarcos and Sukhareva (2015)	Morphology, morphosyntax, syntax, semantics (partially), discourse	NLP, cross-lingual studies, historical linguistics
Multext East Specifications Chiarcos and Erjavec (2011)	Morphosyntax, syntax	NLP, cross-lingual studies
GOLD Farrar and Langendoen (2003)	Phonetics, phonology, morphology, morphosyntax, syntax	Descriptive linguistics, cross-lingual studies
Glottolog/LangDoc Nordhoff and Hammarström (2011)	N/A	Typology

as LD to allow for the exchange of content between different NLP tools, and POWLA (Chiarcos 2012) as a formalism based on RDF and OWL to represent (multi-layer) corpora annotations. The Open Annotation Data Model and other models developed in the context of Biomedical NLP have been applied to share and compare annotations from different corpora (Kim and Wang 2012) on top of which to develop multi-layer corpora search tools (Kim, Cohen and Kim 2015). Additional corpora models are the annotation layer of TELIX (Rubiera *et al.* 2012), the Fiesta/Mexico data model (Menke, McCrae and Cimiano 2013; Menke 2016) to account for annotations in multimodal corpora taking into account different timelines and the Semantic Quran vocabulary (Sherif and Ngonga Ngomo 2015), which was developed to model a multilingual dataset of translations of the Quranic Arabic Corpus and contains general purpose classes to represent chapters, verses, etc., and datasets with a hierarchical structure in general. As examples of corpora that reuse these models we mention The Universal Dependencies Treebank series in RDF, which use the NLP Interchange Format Format along with OLiA; the Manually Annotated Sub-Corpus (MASC) of the Open American National Corpus, which relies on POWLA and also reuses OLiA, and the Semantic Quran corpus, which uses its own vocabulary and links to GOLD to represent linguistic data categories.

Table 5. *Other models*

Model	Domain	Applicable to
NIF (Hellmann *et al.* 2013)	NLP, corpus linguistics	Corpora
POWLA (Chiarcos 2012)	NLP, corpus linguistics	Corpora, treebanks
Fiesta/Mexico (Menke *et al.* 2013; Menke 2016)	NLP, corpus linguistics	Multimodal corpora
Semantic Quran (Sherif and Ngonga Ngomo 2015)	Historical linguistics, digital humanities	Corpora, The Quran
Marl (Westerski and Sánchez-Rada 2013; Buitelaar *et al.* 2013)	NLP (Sentiment analysis, opinion mining)	Lexica, lexical databases
Onyx (Sánchez-Rada and Iglesias Fernández 2013)	NLP (Sentiment analysis, opinion mining)	Lexica, lexical databases
MLSA (Declerck and Krieger 2014)	NLP (Sentiment analysis, opinion mining)	Lexica, lexical databases

Service-oriented models. There have also been advances in the development of models oriented towards NLP services that leverage lexical resources in the LLOD cloud. Marl (Buitelaar *et al.* 2013; Westerski and Sánchez-Rada 2013), for instance, is an ontology that aims at capturing the properties of opinions as expressed online for opinion mining services, while its Onyx extension (Sánchez-Rada and Iglesias Fernández 2013) is aligned with the WordNet-Affect taxonomy[81] and the Emotion Markup Language (EmotionML)[82] in order to encode emotions and their relations to the lexical items lexicalising them in a *lemon*-compliant fashion. In addition, the multi-layered reference corpus for German Sentiment Analysis (MLSA) in its LLD version (Declerck and Krieger 2014) relies on a custom-vocabulary to encode polarity and factuality values.

Table 5 lists these models, the main areas of research in which they are or may be used, and the kind of resources that can instantiate them.

4 Discussion

The analysis of the different models described in the previous section reveals that representing linguistic content from language resources as LD is not a trivial task. Reusing already available representation mechanisms (e.g., *lemon*, SKOS) is a challenge when, at the same time, the modellers try to be respectful to the original resource annotations, be they inspired by tradition, linguistic theory or their use in applied and/or theoretical linguistics.

[81] `http://www.gsi.dit.upm.es/ontologies/wnaffect/`
[82] `http://www.w3.org/TR/emotionml/`

One of the main issues that our study illustrates is the *proliferation of independently developed vocabularies* that have a certain degree of overlap. In fact, vocabularies might have entities in common that may differ to smaller or greater extent from one another. Several resources may use the same type of annotations but define them differently in each case, and this leads in turn to different vocabularies created to transform them. Sometimes these differences between entities with the same label are motivated by practical or theoretical reasons. For example, the OGL `Lexical Entry` class used in the modelling of Oxford Dictionaries differs from `ontolex:LexicalEntry` because it was specifically devised for the conversion of those dictionaries, just as other *ad-hoc* elements defined in the K Dictionaries vocabularies which have a counterpart with the same name in data category registries not fully compatible with the original resource data model. Similarly, two ontologies with theoretical differences may include elements with the same label which may or may not refer to the exact same notion. This is the case of the class `Vehicle` in the PDEV-lemon ontology (derived from the Corpus Pattern Analysis Ontology), and another `Vehicle` as semantic class in the Parole Ontology (GL-based). Likewise, there is a `Time Period` in PDEV-lemon but no relation of this class to the time `Intervals` defined in the OWL Time Ontology, which, again, may be completely different or may share some basic definitions. A further example is the PanLex class `plx:Expression` and the homonym Semiotics Ontology class, and the list goes on. By pointing this out, we are not claiming that these classes should be related, that those models are lacking those relations or that one should reuse the other, but merely highlighting that there does not seem to be a reference model 'glueing' all these models to allow for their comparison. This proliferation of vocabulary entities with some overlap is also due to models being developed simultaneously, or to a model addressing gaps in the literature at that time (e.g., DBnary translation elements) that were further analysed and extensively covered by other models later on (e.g., *lemon* translation module, OntoLex vartrans module).

It is worth noting here that the *syntax-semantics interface and lexical semantics*, in general, seems to be one of the most challenging levels to represent as LD, and creating new classes in *ad-hoc* vocabularies does not always seem to achieve the required results. We have extensions for translations, discourse categories, etymology, morphology, etc., at other levels, but semantic classes, for instance, seem to be problematic still, specially if the resource includes content touching also upon semantic primitives and meaning decomposition, as is the case with LVF. A similar problem is addressed in PDEV-lemon when capturing the meaning of idioms and their relation (or lack of) to the meaning of their units, and the *lexfom* model showcases the difficulty in addressing collocations, pointing to future lines of work in that respect. *Other levels of analysis*, including pragmatics (e.g., speech acts), dialogue structure and phonetics and phonology remain as areas with still a very low coverage in the LLOD cloud.

Catalogues of linguistic data categories include numerous elements that thoroughly address the *morphosyntax* level. However, the number of corpora annotated with *syntax* is still low and the resources annotated (not as LD) with syntactic information beyond phrase or dependency structures, mainly for linguistic research, are hard

to discover: resources with syntactic phenomena, (e.g., syntactic ellipsis, control-constructions) or with the syntactic constructions those linguistic units give rise to (e.g., *seem* as a raising to-subject verb) are still difficult to find. Therefore, since the amount of resources that contain such information is still limited, their LLD versions are even more scarce.

There are *areas related to linguistics* studies in which the LD paradigm is being successfully adopted, such as historical linguistics, lexicography, terminology and cross-linguistic studies. However, its application in other areas remains largely unexplored. That is, the case of speech processing, patholinguistics, sociolinguistics, forensic linguistics, psycholinguistics and neurolinguistics, along with second-language acquisition. This might be caused by a less obvious relation between some of the latter areas and NLP-based applications, as the resources used in many of these disciplines, e.g., sociolinguistics, are not commonly involved in NLP tools for machine translation, named entity recognition and linking, text summarisation, and the like. Interestingly, speech resources and spoken language data are building blocks of NLP systems for speech recognition, but the representation of multimedia data as LLD has not been much addressed in the literature apart from the Fiesta/Mexico model (Menke *et al.* 2013). Other emerging disciplines at the interface between those mentioned above and computer science, e.g., computational neurolinguistics, computational forensic linguistics, computational sociolinguistics, etc. are likely to face the same problems of other branches in linguistics, namely, syntactic and semantic heterogeneity, which the LD paradigm and the Semantic Web would potentially help solve.

Multilingualism and the added value of linking multilingual resources in the context of NLP systems and cross-linguistic studies has emerged as one trending topic in the work with LLD, and the properties available in the OntoLex *vartrans* module (coming in turn from the *lemon* translation module), LexInfo, DBnary and SKOS account for that aspect in different ways, so that these mechanisms provide *a solid basis* to represent multilingual data in the LLD context. Several challenges arise when trying to automatically establish links among resources in different languages (Gracia *et al.* 2012a; Gracia, Montiel-Ponsoda and Gómez-Pérez 2012b), or when determining if two entities serve as cross-linguistic equivalents in any context or only in certain contexts (as some commonly cited examples reveal: *river* versus Fr. *rivière* or *flueve*; *president* or *prime minister* versus Sp. *presidente*, etc.), or even when addressing specific features of a multilingual resource (for instance, multilingual examples in a dictionary).

As for the reuse of available mechanisms, and in addition to *lemon* and OntoLex established already as *de facto* standards, LexInfo and OLiA also seem to be the *most reused external linguistic registries* that are used with *lemon* and OntoLex, over ISOcat and GOLD. *lemonUby* categories and the PDEV-lemon ontologies are also, to a smaller extent, reused by other works presenting their own extension. At the instantiation level, although not the focus of this review, we mention linkage to WordNet as one of the pivotal steps in the linking of multilingual information carried out in several of the efforts mentioned above, and BabelNet as an encyclopedic hub linking linguistic resources in the LLOD cloud.

Even though this scenario offers a promising perspective of the LLD landscape and points to directions for future developments, there are still some aspects to be addressed on the way to a wider community adoption of LLD:

- First, in comparison to the number of resources available and growing, there are still few tools and services that consume LLD and that leverage the links across datasets. The benefits of LLD are partially outshined by this lack, which makes them not fully evident to experts outside the Semantic Web. The LLD community, however, has started to tackle it (McCrae *et al.* 2015; Chiarcos *et al.* 2016a).
- Second, and related to what is described as the lack of expressivity in the LOD cloud (Jain *et al.* 2010), the datasets in the LLD cloud are usually not annotated using the full potential that semantic formalisms allow. This, together with the lack of schema level links (Jain *et al.* 2010; Millard *et al.* 2010; Anjomshoaa *et al.* 2014), highlighted in this survey in the case of lexical semantic models, prevents reasoners from inferring more knowledge, which in turn would help in showing the full potential of LLD. A way to mitigate this issue will come from the application of ontology matching techniques to establish links among the different linguistic models and from the development of specific reasoners ready to deal with linguistic knowledge.
- A third reason here would be that the gap between linguistic knowledge in the sub-disciplines of linguistics and the linguistic knowledge available in linguistic models in the Semantic Web is still very wide. The intended use of this knowledge varies, as mentioned before: the knowledge needed for an application to perform named entity linking is different from the one needed for conducting theoretical linguistic research. Thus, domain experts working on, for instance, the properties of verbs of movement in English and Spanish, may find the information about verbs in current ontologies of little use for their research. However, as the models in this work reveal, gradually more and more granularity is being added onto the different levels through extensions, new ontologies and new use cases.

5 Conclusions

In this survey, we have provided an overview of the current and available mechanisms to model linguistic content as LLD in order to give a sense of the main lines of work, trends and remaining challenges in the field. Throughout the paper, we have tried to provide answers to the research questions raised in Section 2, namely the following:

(1) What are the available models that allow to represent linguistic information as LD in its different description levels?
(2) What are the main modelling difficulties that arise when representing linguistic information as LD and how do different models tackle them?
(3) How have such LD-based linguistic models been (re)used, adapted and extended?

(4) What are the major remaining challenges to describe linguistic content in the current LD-based models?

Research question 1 has been addressed in Section 3, specifically in Section 3.2. Tables 1–5 provide summaries to the main points highlighted in each subsection of Section 3.

The modelling difficulties in the representation of linguistic content as LLD we identified in our analysis (question 2) and some remaining challenges that are still to be faced (question 4) were considered and discussed in Section 4. In addition, the reuse, adaptation and extensions of general models to represent the structure of lexica (question 3) were described throughout Section 3.2, and also referred to as part of the discussion.

In terms of future work, there are several directions to pursue. First, this survey will serve as a general framework for a critical, comparative analysis of the models presented here. This comparison could be performed in terms of the practices built for ontology development followed in each case, the structural and logical consistency of the ontologies (supported by ontology evaluation tools and reasoners), their coverage of the domain, their reuse in the community, their maintenance and documentation, etc. The results of such a comparison would provide some guidance on which model better suits a specific use case. Second, although this work mentioned some gaps in the presented models, it was not focused on specific modelling solutions for linguistic phenomena. For each group and sub-group introduced in this survey, it would be interesting to further investigate modelling gaps through the analysis of a set of representative examples extracted from various language resources of different kinds. In addition, and in order to complement our descriptive approach in this article, we plan to analyse the patterns of use of some of the models discussed in the previous pages so as to detect, compare and discuss common and emerging modelling practices and instantiation choices per resource type in the representation of language resources as LLD. This analysis will be accompanied by both an overview of the resources that instantiate these models as well as statistics of their term coverage. The obtained results are expected to reflect the current modelling heterogeneity in the conversion of language resources to LLD and will serve as input for the revision of these models and ontologies. Last, more research is needed regarding the issues on the way for a wider community adoption of LLD, some of them mentioned above.

References

Abromeit, F., and Fäth, C. 2016. Linking the tower of babel: modelling a massive set of etymological dictionaries as RDF. In *Proceedings of the 5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources (LDL 2016)*. Web. `https://pdfs.semanticscholar.org/62c1/53cd0ee24f40c7db9060dc064a489e54e480.pdf`

Alexiev, V., and Casamayor, G. 2016. FN goes NIF: integrating FrameNet in the NLP interchange format. In *Proceedings of the LDL 5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources*, pp. 1–10. `https://pdfs.semanticscholar.org/62c1/53cd0ee24f40c7db9060dc064a489e54e480.pdf`

Anjomshoaa, A., Kiesling, E., Tuan, D. T., Lam, D. B., Wetz, P., and Tjoa, A. M. 2014. Leveraging the web of data via linked widgets. *Journal of Service Science Research* **6**(1): 7–27. http://doi.org/10.1007/s12927-014-0001-9.

Berners-Lee, T., Hendler, J., and Lassila, O. 2001. The semantic web. *Scientific American* **284**(5): 34–43. Web. http://www.jstor.org/stable/26059207.

Bizer, C., Heath, T., and Berners-Lee, T. 2009. Linked data—the story so far. *Semantic services, interoperability and web applications: emerging concepts*. doi:10.4018/978-1-60960-593-3.ch008.

Bond, F., Vossen, P., McCrae, J. P., and Fellbaum, C. 2016. Cili: the Collaborative Interlingual Index. In *Proceedings of the Global WordNet Conference*, Bucharest, Romania, pp. 50–57. Amsterdam, The Netherlands: Global WordNet Association.

Borin, L., Dannells, D., Forsberg, M., and McCrae, J. P. 2014. Representing Swedish lexical resources in RDF with *lemon*. In *Proceedings of the International Conference on Posters and Demonstrations Track (ISWC-PD'14)*, Riva del Garda, Italy, vol. 1272, pp. 329–332. Aachen, Germany: CEUR-WS, pp. 205–227.

Bosque-Gil, J., Gracia, J., Montiel-Ponsoda, E., and Aguado-de-Cea, G. 2016. Modelling multilingual lexicographic resources for the web of data: the K dictionaries case. In *Proceedings of the GLOBALEX Lexicographic Resources for Human Language Technology Workshop Programme*. http://ailab.ijs.si/globalex/files/2016/06/LREC2016Workshop-GLOBALEX_Proceedings-v2.pdf

Bosque-Gil, J., Montiel-Ponsoda, E., Gracia, J., and Aguado-de-Cea, G. 2016. Terminoteca RDF: a gathering point for multilingual terminologies in Spain. In *Proceedings of TKE 2016 the 12th International conference on Terminology and Knowledge Engineering*, pp. 136–146. Copenhagen: Copenhagen Business School, pp. 65–72. http://hdl.handle.net/10398/9323.

Buitelaar P., *et al.* 2006. LingInfo: design and applications of a model for the integration of linguistic information in ontologies. In *Proceedings of the OntoLex Workshop at Language Resources and Evaluation Conference 2006*. Genoa, Italy: o. A, pp. 28–32. http://pub.uni-bielefeld.de/download/2497539/2525178.

Buitelaar, P., Arcan, M., Iglesias Fernández, C. A., Sánchez Rada, J. F., and Strapparava, C. 2013. Linguistic linked data for sentiment analysis. In *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013)*, pp. 1–8, Pisa, Italy. Stroudsburg, PA, USA: Association for Computational Linguistics. http://aclweb.org/anthology/W13-55.

Chiarcos, C. 2012. POWLA: modeling linguistic corpora in OWL/DL. In V. Presutti (ed.), *The Semantic Web: Research and Applications (ESWC 2012)*. Lecture Notes in Computer Science, vol. 7295, pp. 225–239. Berlin, Heidelberg, Germany: Springer.

Chiarcos, C. 2014. Towards interoperable discourse annotation. Discourse features in the ontologies of linguistic annotation. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. Paris, France: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/893_Paper.pdf

Chiarcos, C., and Erjavec, T. 2011. OWL/DL formalization of the MULTEXT-East morphosyntactic specifications. In *Proceedings of the 5th Linguistic Annotation Workshop*, pp. 11–20. Portland, Oregon. Stroudsburg, PA, USA: Association for Computational Linguistics.

Chiarcos, C., Hellmann, S., and Nordhoff, S. 2012a. The open linguistics working group of the open knowledge foundation. In C. Chiarcos, S. Nordhoff and S. Hellmann (ed.), *Linked Data in Linguistics*, pp. 153–160. Berlin, Heidelberg, Germany: Springer.

Chiarcos, C., Hellmann, S., and Nordhoff, S. 2012b. Linking linguistic resources: examples from the open linguistics working group. In C. Chiarcos, S. Nordhoff and S. Hellmann (ed.), *Linked Data in Linguistics*, pp. 201–216. Berlin, Heidelberg, Germany: Springer.

Chiarcos, C., McCrae, J., Cimiano, P., and Fellbaum, C. 2013. Towards open data for linguistics: linguistic linked data. In *New Trends of Research in Ontologies and Lexical Resources*, pp. 7–25. Berlin, Heidelberg, Germany: Springer.

Chiarcos, C., Nordhoff, S., and Hellmann, S. (eds.) 2012c. *Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata*. Berlin, Heidelberg, Germany: Springer.

Chiarcos, C., and Sukhareva, M. 2014. Linking etymological databases. A case study in Germanic. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing (LDL-2014)*, pp. 41–49. Stroudsburg, PA, US: Association for Computational Linguistics. http://toc.proceedings.com/28686webtoc.pdf

Chiarcos, C., and Sukhareva, M. 2015. OLiA – ontologies of linguistic annotation. *Semantic Web* **6**(4): 379–386. http://semantic-web-journal.net/content/olia-%E2%80%93-ontologies-linguistic-annotation.

Chiarcos, C., Fäth, C., Renner-Westermann, H., Abromeit, F., and Dimitrova, V. 2016a. Lin | gu | is | tik: building the Linguist's pathway to bibliographies, libraries, language resources and linked open data. In *Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC)*, pp. 4463–4471, Portoroz, Slovenia. Paris, France: European Language Resources Association (ELRA).

Chiarcos, C., Fäth, C., and Sukhareva, M. 2016b. Developing and using the ontologies of linguistic annotation (2006–2016). In *Proceedings of the LDL 5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources*. http://pdfs.semanticscholar.org/62c1/53cd0ee24f40c7db9060dc064a489e54e480.pdf

Cimiano, P., Buitelaar, P., McCrae, J., and Sintek, M. 2011. LexInfo: a declarative model for the lexicon-ontology interface. *Web Semantics: Science, Services and Agents on the World Wide Web* **9**(1): 29–51. http://www.sciencedirect.com/science/article/pii/S1570826810000892.

Cimiano, P., Haase, P., Herold, M., Mantel, M., and Buitelaar, P. 2007. LexOnto: a model for ontology lexicons for ontology-based NLP. In *Proceedings of the OntoLex07 Workshop*, vol. 7. Busan, South Korea. http://pub.uni-bielefeld.de/download/2497432/2524813

Cimiano, P., McCrae, J. P., Rodríguez-Doncel, V., Gornostay, T., Gómez-Pérez, A., Siemoneit, B., and Lagzdins, A. 2015. Linked terminologies: applying linked data principles to terminological resources. In *Proceedings of the eLex 2015 Conference*, pp. 504–517, 11–13 August 2015, Herstmonceux Castle, U.K. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd.

Corcho, O., Fernández-López, M., and Gómez-Pérez, A. 2003. Methodologies, tools and languages for building ontologies. Where is their meeting point?. *Data and Knowledge Engineering* **46**(1): 41–64. http://www.sciencedirect.com/science/article/pii/S0169023X02001957

Corcoglioniti, F., Rospocher, M., Aprosio, A. P., and Tonelli, S. 2016. PreMOn: a lemon extension for exposing predicate models as Linked Data. In *Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC)*, pp. 877–884. Portoroz, Slovenia. Paris, France: European Language Resources Association (ELRA).

Crystal, D. 1971. *Linguistics*. Baltimore: Penguin Books.

Declerck, T. 2006. SynAF: towards a standard for syntactic annotation. In *Proceedings of the 5th edition of the Language Resources and Evaluation Conference (LREC)*, pp. 229–232. Genoa, Italy: European Language Resources Association (ELRA).

Declerck, T., and Krieger, H. U. 2014. Harmonization of German lexical resources for opinion mining. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*, pp. 3872–3876. Reykjavik, Iceland. Paris, France: European Language Resources Association (ELRA).

Declerck, T., Lendvai, P., Mörth, K., Budin, G., and Váradi, T. 2012. Towards linked language data for digital humanities. In C. Chiarcos, S. Nordhoff and S. Hellmann (ed.), *Linked Data in Linguistics*, pp. 109–116. Berlin, Heidelberg, Germany: Springer.

Declerck, T., and Mörth, K. 2016. Towards a sense-based access to related online lexical resources. In *Proceedings of the XVII EURALEX International Congress*, pp. 660–667. Tbilisi, Georgia: Ivane Javakhishvili Tbilisi State University.

Declerck, T., Wandl-Vogt, E., and Mörth, K. 2015. Towards a pan European lexicography by means of linked (open) data. In *Proceedings of the Electronic Lexicography in the 21st Century: Linking Lexical Data in the Digital Age. Proceedings of the eLex 2015 conference*, pp. 342–355. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd.

de la Clergerie, E., and Clément, L. 2005. MAF: a morphosyntactic annotation framework. In Z. Vetulani (ed.), *Proceedings of the 2nd Language & Technology Conference (LTC'05)*, pp. 90–99. Warsaw, Poland: De Gruyter Open.

Del Gratta, R., Frontini, F., Khan, F., and Monachini, M. 2015. Converting the parole simple clips lexicon into RDF with *lemon*. *Semantic Web* **6**(4): 387–392. Web. `http://www.semantic-web-journal.net/content/converting-parole-simple-clips-lexicon-rdf-lemon-0`.

de Melo, G. 2015. Lexvo. org: language-related information for the linguistic linked data cloud. *Semantic Web* **6**(4): 393–400. `http://www.semantic-web-journal.net/content/lexvoorg-language-related-information-linguistic-linked-data-cloud-0`.

Desipri, E., Gavrilidou, M., Labropoulou, P., Piperidis, S., Frontini, F., Monachini, M., Arranz, V., Mapelli, V., Francopoulo, G., and Declerck, T. 2012. Documentation and user manual of the META-SHARE metadata model. Labropoulou and E. Desipri. `http://www.meta-net.eu/public_documents/t4me/META-NET-D7.2.4-Final.pdf`.

Dixon-Woods, M., Agarwal, S., Jones, D., Young, B., and Sutton, A. 2005. Synthesising qualitative and quantitative evidence: a review of possible methods. *Journal of Health Services Research and Policy* **10**(1): 45–53.

Dyba, T., Dingsoyr, T., and Hanssen, G. K. 2007. Applying systematic reviews to diverse study types: An experience report. In *Proceedings of the 1st International Symposium on Empirical Software Engineering and Measurement (ESEM 2007)*. IEEE. Web. `http://doi.org/10.1109/ESEM.2007.59`

Eckle-Kohler, J., McCrae, J. P., and Chiarcos, C. 2015. lemonUby – A large, interlinked, syntactically-rich lexical resource for ontologies. *Semantic Web* **6**(4): 371–378. `http://www.semantic-web-journal.net/content/lemonuby-large-interlinked-syntactically-rich-lexical-resource-ontologies-1`.

Ehrmann, M., Cecconi, F., Vannella, D., McCrae, J. P., Cimiano, P., and Navigli, R. 2014. Representing multilingual data as linked data: the case of BabelNet 2.0. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC 2014)*, Reykjavik, Iceland. Paris, France: European Language Resources Association (ELRA), pp. 401–408.

El Maarouf, I., Alferov, E., Cooper, D., Fang, Z., Mousselly-Sergieh, H., and Wang, H. 2015. The GuanXi network: a new multilingual LLOD for Language Learning applications. In *Proceedings of the 2nd Workshop on Natural Language Processing and Linked Open Data*, Hissar, Bulgaria, pp. 42–51. Shoumen, Bulgaria: INCOMA Ltd.

El Maarouf, I., Bradbury, J., and Hanks, P. 2014. PDEV-lemon: a Linked Data implementation of the pattern dictionary of english verbs based on the *lemon* model. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics (LDL)*, pp. 88–93. Reykjavik, Iceland. Paris, France: European Language Resources Association (ELRA).

Falk, I., and Stein, A. 2016. LVF-lemon – towards a linked data representation of 'Les Verbes français'. In *Proceedings of the 10th edition of the Language Resources and Evaluation*

Conference (LREC), pp. 2401–2406. Portoroz, Slovenia. Paris, France: European Language Resources Association (ELRA).

Farrar, S., and Langendoen, D. T. 2003. A linguistic ontology for the semantic web. *GLOT International* **7**(3): 97–100.

Fellbaum, C. 1998. *WordNet*. Hoboken, New Jersey, US: Blackwell Publishing Ltd.

Fellbaum, C. 2005. WordNet and wordnets. In K. Brown *et al.* (eds.), *Encyclopedia of Language and Linguistics*, 2nd ed., pp. 665–670. Oxford: Elsevier.

Flati, T., and Navigli, R. 2013. Three birds (in the LLOD cloud) with one stone: BabelNet, Babelfy and the Wikipedia Bitaxonomy. In *Proceedings of SEMANTiCS*, pp. 10–13. Leipzig, Germany, September 4–5. CEUR-WS.org. Web. `http://ceur-ws.org/Vol-1224/paper3.pdf`.

Fonseca, A., Sadat, F., and Lareau, F. 2016. Lexfom: a lexical functions ontology model. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex - V)*, pp. 145–155. Osaka, Japan: The COLING 2016 Organizing Committee.

Forkel, R. 2014. The cross-linguistic linked data project. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics (LDL)*, pp. 60–66. Reykjavik, Iceland. Paris, France: European Language Resources Association (ELRA).

Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M., and Soria, C. 2006. Lexical Markup Framework (LMF). In *Proceedings of the 5th edition of the Language Resources and Evaluation Conference (LREC)*, pp. 233–236. Genoa, Italy. Paris, France: European Language Resources Association (ELRA).

Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., and Schneider, L. 2002. Sweetening ontologies with DOLCE. In *Proceedings of the International Conference on Knowledge Engineering and Knowledge Management*, pp. 166–181. Berlin, Heidelberg: Springer.

Gómez-Pérez, A., Vila-Suero, D., Montiel-Ponsoda, E., Gracia, J., and Aguado-de-Cea, G. 2013. Guidelines for multilingual linked data. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*, pp. 3–12. Madrid, Spain. New York, NY, USA: ACM.

Gonzalez-Agirre, A., Laparra, E., and Rigau, G. 2012. Multilingual central repository version 3.0. In *Proceedings of the 8th edition of the Language Resources and Evaluation Conference (LREC)*, pp. 2525–2529. Istanbul, Turkey. Paris, France: European Language Resources Association (ELRA).

González-Blanco, E., del Rio, G., and Martínez Cantón, C. 2016. Linked open data to represent multilingual poetry collections. A proposal to solve interoperability issues between poetic repertoires. In *Proceedings of the 5th Workshop on Linked Data in Linguistics*. Portoroz, Slovenia. Paris, France: European Language Resources Association (ELRA).

Good, J., and Hendryx-Parker, C. 2006. Modeling contested categorization in linguistic databases. In *Proceedings of the EMELD Workshop on Digital Language Documentation: Tools and Standards: The State of the Art*. Lansing, Michigan. Wayne State University, Eastern Michigan University. Web. `http://emeld.org/workshop/2006/papers/GoodHendryxParker-Modelling.pdf`.

Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Gómez-Pérez, A., Buitelaar, P., and McCrae, J. 2012a. Challenges for the multilingual web of data. *Web Semantics: Science, Services and Agents on the World Wide Web* **11**: 63–71. `http://www.sciencedirect.com/science/article/pii/S1570826811000783`.

Gracia, J., Montiel-Ponsoda, E., and Gómez-Pérez, A. 2012b. Cross-lingual linking on the multilingual web of data (position statement). In *Proceedings of the 3rd International Conference on Multilingual Semantic Web*, vol. 936, pp. 41–45. Aachen, Germany: CEUR-WS.org.

Gracia, J., Montiel-Ponsoda, E., Vila-Suero, D., and Aguado-de Cea, G. 2014. Enabling language resources to expose translations as linked data on the web. In *Proceedings of 9th Language Resources and Evaluation Conference (LREC 2014)*, pp. 409–413. Reykjavik, Iceland. Paris, France: European Language Resources Association (ELRA).

Gracia, J., Villegas, M., Gómez-Pérez, A., and Bel, N. 2016. The apertium bilingual dictionaries on the web of data. *Semantic Web Journal* **9**(2): 231–240. doi.org/10.3233/SW-170258.

Hanks, P. 2004. Corpus Pattern Analysis. In *Euralex Proceedings*, vol. 1, pp. 87–98. Lorient, France: Université de Bretagne-Sud.

Hanks, P. 2007. Pattern Dictionary of English Verbs (PDEV) – Project Page. http://pdev.org.uk. Last accessed on 02.08.2018.

Hanks, P. 2013. *Lexical Analysis: Norms and Exploitations*. Cambridge, MA, US: MIT Press.

Hanks, P., and Pustejovsky, J. 2005. A pattern dictionary for natural language processing. *Revue Française de linguistique appliquée* **10**(2), pp. 63–82.

Hayes, B., Curtiss, S., Szabolcsi, A., Stowell, T., Stabler, E., Sportiche, D., Koopman, H., Keating, P., Munro, P., Hyams, N., and Steriade, D. 2013. *Linguistics: An Introduction to Linguistic Theory*. Hoboken, New Jersey, US: John Wiley & Sons.

Hellmann, S. 2013. *Integrating Natural Language Processing (NLP) and Language Resources Using Linked Data*. Universität Leipzig. PhD Thesis. http://svn.aksw.org/papers/2013/Thesis_Sebastian/submission/public_version.pdf.

Hellmann, S., Lehmann, J., Auer, S., and Brümmer, M. 2013. Integrating NLP using linked data. In *International Semantic Web Conference*, pp. 98–113. Berlin, Heidelberg: Springer.

Ide, N., and Romary, L. 2004. International standard for a linguistic annotation framework. *Natural Language Engineering* **10**(3–4): 211–225. Cambridge, UK: Cambridge University Press.

Ide, N., and Suderman, K. 2007. GrAF: a graph-based format for. linguistic annotations. In *Proceedings of the Linguistic Annotation Workshop*, pp. 1–8. Prague, Czech Republic. Stroudsburg, PA, US: Association for Computational Linguistics.

Iqbal, R., Murad, M. A. A., Mustapha, A., and Sharef, N. M. 2013. An analysis of ontology engineering methodologies: a literature review. *Research Journal of Applied Sciences, Engineering and Technology* **6**(16): 2993–3000.

Jain, P., Hitzler, P., Yeh, P. Z., Verma, K., and Sheth, A. P. 2010. Linked data is merely more data. In *Proceedings of the AAAI Spring Symposium: Linked Data Meets Artificial Intelligence*, vol. 11, pp. 82–86. Palo Alto, California: . AAAI Publications. http://www.aaai.org/ocs/index.php/SSS/SSS10/paper/view/1130/1454.

Kemps-Snijders, M., Windhouwer, M., Wittenburg, P., and Wright, S. E. 2008. ISOcat: corralling data categories in the wild. In *Proceedings of the 6th edition of the Language Resources and Evaluation Conference (LREC)*, pp. 28–30. Marrakech, Morocco. Paris, France: European Language Resources Association (ELRA).

Khan, F., Bellandi, A., and Monachini, M. 2016b. Tools and instruments for building and querying diachronic computational lexica. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pp. 164–171. Osaka, Japan: The COLING 2016 Organizing Committee.

Khan, F., Boschetti, F., and Frontini, F. 2014. Using *lemon* to model lexical semantic shift in diachronic lexical resources. In *Proceedings of the LDL-2014 3rd Workshop on Linked Data in Linguistics*, pp. 49–53. Reykjavik, Iceland. Paris, France: European Language Resources Association (ELRA).

Khan, F., Diaz-Vera, J. E., and Monachini, M. 2016a. Representing polysemy and diachronic lexicosemantic data on the semantic web. In *Proceedings of the 2nd International Workshop on Semantic Web for Scientific Heritage co-located with 13th Extended Semantic Web Conference (ESWC 2016)*, pp. 37–46. Heraklion, Greece. Web. CEUR-WS.org. http://ceur-ws.org/Vol-1595/paper4.pdf.

Khan, F., Frontini, F., Boschetti, F., and Monachini, M. 2016c. Converting the Liddell Scott Greek-English lexicon into linked open data using lemon. In *Proceedings of the Digital Humanities 2016: Conference Abstracts*, pp. 593–596. Kraków, Poland: Jagiellonian University and Pedagogical University.

Khan, F., Frontini, F., Del Gratta, R., Monachini, M., and Quochi, V. 2013. Generative lexicon theory and linguistic linked open data. In *Proceedings of the 6th International Conference*

on Generative Approaches to the Lexicon, pp. 62–69. Pisa, Italy. Stroudsburg, PA, USA: Association for Computational Linguistics.

Kim, J. D., Cohen, K. B., and Kim, J. J. 2015. PubAnnotation-query: a search tool for corpora with multi-layers of annotation. *BMC Proceedings* **9**(5): A3. https://doi.org/10.1186/1753-6561-9-S5-A3.

Kim, J. D., and Wang, Y. 2012. PubAnnotation: a persistent and sharable corpus and annotation repository. In *Proceedings of the Workshop on Biomedical Natural Language Processing*, pp. 202–205. Montrèal, Canada. Stroudsburg, PA, USA: Association for Computational Linguistics.

Klimek, B. 2017. Proposing an OntoLex-MMoOn alignment: towards an interconnection of two linguistic domain models. In *Proceedings of the LDK Workshops: OntoLex, TIAD and Challenges for Wordnets*, pp. 68–73. Galway, Ireland. CEUR-WS.org. Web. http://ceur-ws.org/Vol-1899/OntoLex_2017_paper_6.pdf.

Klimek, B., Arndt, N., Krause, S., and Arndt, T. 2016. Creating linked data morphological language resources with MMoOn. The Hebrew Morpheme Inventory. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 892–899. Portoroz, Slovenia. Paris, France: European Language Resources Association (ELRA).

Klimek, B., and Brümmer, M. 2015. Enhancing lexicography with semantic language databases. *Kernerman Dictionary News* **23**: 5–10.

Lenci, A., Bel, N., Busa, F., Calzolari, N., Gola, E., Monachini, M., Ogonowski, A., Peters, W., Ruimy, N., Villegas, M., and Zampolli, A. 2000. SIMPLE: A general framework for the development of multilingual lexicons. *International Journal of Lexicography* **13**(4): 249–263.

Liu, H., and Singh, P. 2004. ConceptNet – a practical commonsense reasoning tool-kit. *BT Technology Journal* **22**(4): 211–226.

Mann, W. C., and Thompson, S. A. 1988. Rhetorical structure theory: toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse* **8**(3): 243–281.

McCrae, J. P., Aguado-de-Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., and Wunner, T. 2010. The lemon cookbook. Technical report, Monnet Project. http://www.lemon-model.net/lemon-cookbook.pdf.

McCrae, J. P., Aguado-de-Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., and Wunner, T. 2012. Interchanging lexical resources on the semantic web. *Language Resources and Evaluation* **46**(4): 701–719.

McCrae, J. P., and Cimiano, P. 2015. Linghub: a Linked Data based portal supporting the discovery of language resources. In *Proceedings of SEMANTiCS (Posters and Demos)*, vol. 1481, pp. 88–91. Vienna, Austria. CEUR-WS.org. http://ceur-ws.org/Vol-1481/paper27.pdf.

McCrae, J. P., Chiarcos, C., Bond, F., Cimiano, P., Declerck, T., de Melo, G., ... and Osenova, P. 2016. The open linguistics working group: developing the linguistic linked open data cloud. In *Proceedings of 10th Language Resources and Evaluation Conference (LREC 2016)*, pp. 2435–2441. Portoroz, Slovenia. Paris, France: European Language Resources Association (ELRA).

McCrae, J. P., Fellbaum, C., and Cimiano, P. 2014. Publishing and linking WordNet using *lemon* and RDF. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics*, Reykjavik, Iceland. Paris, France: European Language Resources Association (ELRA). https://pub.uni-bielefeld.de/download/2732779/2732785.

Mel'cuk, I. M. 1997. *Vers une linguistique Sens-Texte: leçon inaugurale faite le vendredi 10 janvier 1997*. Paris, France: Collège de France.

Mel'cuk, I. 1998. Collocations and lexical functions. In A. P. Cowie (ed.), *Phraseology. Theory, Analysis, and Applications*, pp. 23–53. Oxford: Clarendon Press.

Menke, P. 2016. *The Fiesta Data Model: A Novel Approach to the Representation of Heterogeneous Multimodal Interaction Data.* Norderstedt, Germany: BoD–Books on Demand.

Menke, P., McCrae, J., and Cimiano, P. 2013. Releasing multimodal data as Linguistic Linked Open Data: An experience report. In *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and Linking Lexicons, Terminologies and Other Language Data*, pp. 44–52. Pisa, Italy. Stroudsburg, PA, USA: Association for Computational Linguistics.

Miles, A., and Bechhofer, S. 2009. SKOS simple knowledge organization system reference. 18 August 2009, W3C recommendation. World Wide Web Consortium. `https://www.w3.org/TR/skos-reference`.

Millard, I., Glaser, H., Salvadores, M., and Shadbolt, N. 2010. Consuming multiple linked data sources: challenges and experiences. In: *Proceedings of the 1st International Workshop on Consuming Linked Data (COLD2010)*, pp. 1–12. Shanghai, China. CEUR-WS.org. Web. `http://ceur-ws.org/Vol-665/MillardEtAl_COLD2010.pdf`.

Montiel-Ponsoda, E., Aguado de Cea, G., Gómez-Pérez, A., and Peters, W. 2011. Enriching ontologies with multilingual information. *Natural Language Engineering* **17**(03): 283–309.

Moran, S. 2012. Using Linked Data to create a typological knowledge base. In C. Chiarcos, S. Nordhoff and S. Hellmann (ed.), *Linked Data in Linguistics*, pp. 129–138. Berlin, Heidelberg, Germany: Springer.

Moran, S., and Wright, R. 2009. Phonetics information base and lexicon (PHOIBLE). Web. `http://phoible.org`.

Nordhoff, S. 2012. Linked data for linguistic diversity research: Glottolog/langdoc and asjp online. In C. Chiarcos, S. Nordhoff and S. Hellmann (ed.), *Linked Data in Linguistics*, pp. 191–200. Berlin, Heidelberg, Germany: Springer.

Nordhoff, S., and Hammarström, H. 2011. Glottolog/Langdoc: defining dialects, languages, and language families as collections of resources. In *Proceedings of the 1st International Conference on Linked Science*, vol. 783, pp. 53–58. CEUR-WS. Web. `http://ceur-ws.org/Vol-783/paper7.pdf`.

Nuzzolese, A. G., Gangemi, A., and Presutti, V. 2011. Gathering lexical linked data and knowledge patterns from FrameNet. In *Proceedings of the 6th International Conference on Knowledge Capture (K-CAP '11)*, pp. 41–48. Banff, AB, Canada. New York, New York, USA: ACM.

Ogden, C. K., Richards, I. A., Ranulf, S., and Cassirer, E. 1923. *The Meaning of Meaning. A Study of the Influence of Language upon Thought and of the Science of Symbolism.* London, UK: Routledge.

Oltramari, A., and Vetere, G. 2008. Lexicon and ontology interplay in Senso Comune. In *Proceedings of OntoLex*, pp. 24–30. Marrakech, Morocco. Paris, France: European Language Resources Association (ELRA).

Parvizi, A., Kohl, M., González, M., and Saur, M. 2016. Towards a linguistic ontology with an emphasis on reasoning and knowledge reuse. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 441–448. Portoroz, Slovenia. Paris, France: European Language Resources Association (ELRA).

Peirce, C. S., Hartshorne, C., Weiss, P., and Burks, A. W. 1958. *Collected Papers of Charles Sanders Peirce: Edited by Charles Hartshorne and Paul Weiss: Science and philosophy and Reviews, Correspondence, and Bibliography*. Cambridge, MA, USA: Harvard University Press.

Picca, D., Gliozzo, A. M., and Gangemi, A. 2008. LMM: an OWL-DL MetaModel to represent heterogeneous lexical knowledge. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08)*, pp. 2413–2419. Marrakech, Morocco. Paris, France: European Language Resources Association (ELRA).

Polguère, A. 2006. Structural properties of lexical systems: monolingual and multilingual perspectives. In *Proceedings of the Workshop on Multilingual Language Resources and*

Interoperability, pp. 50–59. Sydney, Australia. Stroudsburg, PA, USA: Association for Computational Linguistics.

Polguère, A. 2014. From writing dictionaries to weaving lexical networks. *International Journal of Lexicography* **27**(4): 396–418.

Pradhan, S., Hovy, E. H., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. 2007. OntoNotes: a unified relational semantic representation. *International Journal of Semantic Computing* **4**(1): 405–419.

Pustejovsky, J. 1991. The generative lexicon. *Computational Linguistics* **17**(4): 409–441.

Pustejovsky, J., and Bouillon, P. 1995. Aspectual coercion and logical polysemy. *Journal of Semantics* **12**(2): 133–162.

Rubiera, E., Polo, L., Berrueta, D., and El Ghali, A. 2012. TELIX: an RDF-based model for linguistic annotation. In E. Simperl, P. Cimiano, A. Polleres, O. Corcho and V. Presutti (ed.), *The Semantic Web: Research and Applications*, pp. 195–209. Berlin, Heidelberg, Germany: Springer.

Ruimy, N., Corazzari, O., Gola, E., Spanu, A., Calzolari, N., and Zampolli, A. 1998. The european le-parole project: the Italian syntactic lexicon. In *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC 1998)*, pp. 241–248. Granada, Spain. Paris, France: European Language Resources Association.

Ruimy, N., Monachini, M., Distante, R., Guazzini, E., Molino, S., Ulivieri, M., Calzolari, N., and Zampolli, A. 2002. CLIPS, a multi-level Italian computational lexicon: a glimpse to data. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2002)*, pp. 792–799. Las Palmas, Canary Islands, Spain. Paris, France: European Language Resources Association.

Sánchez Rada, J. F., and Iglesias Fernandez, C. A. 2013. Onyx: describing emotions on the web of data. In *Proceedings of the 1st International Workshop on Emotion and Sentiment in Social and Expressive Media*, pp. 71–82. Turin, Italy. CEUR-WS.org. Web. `http://ceur-ws.org/Vol-1096/paper6.pdf`.

Saulwick, A., Windhouwer, M., Dimitriadis, A., and Goedemans, R. 2005. Distributed tasking in ontology mediated integration of typological databases for linguistic research. In *Proceedings of the 17th Conference on Advanced Information Systems Engineering (CAiSE 2005)*, Porto, Portugal. Porto, Portugal: FEUP Ed.

Schuurman, I., Windhouwer, M., Ohren, O., and Zeman, D. 2016. CLARIN concept registry: the new semantic registry. In *Selected Papers from the CLARIN Annual Conference 2015*, October 14–16, 2015, Wroclaw, Poland, vol. 123, pp. 62–70. Linköping, Sweden: Linköping University Electronic Press.

Sérasset, G. 2015. Dbnary: wiktionary as a lemon-based multilingual lexical resource in rdf. *Semantic Web* **6**(4): 355–361. `http://www.semantic-web-journal.net/content/dbnary-wiktionary-lemon-based-multilingual-lexical-resource-rdf`.

Sherif, M. A., and Ngonga Ngomo, A. C. 2015. Semantic Quran. *Semantic Web* **6**(4): 339–45. `http://www.semantic-web-journal.net/content/semantic-quran-multilingual-resource-natural-language-processing`.

Skut, W., Brants, T., Krenn, B., and Uszkoreit, H. 1998. A linguistically interpreted corpus of German newspaper text. In *Proceedings of the 10th European Summer School in Logic, Language and Information (ESSLLI'98). Workshop on Recent Advances in Corpus Annotation*, Saarbrücken, Germany, o.A. `http://www.coli.uni-sb.de/publikationen/softcopies/Skut:1998:LIC.pdf`.

Sperberg-McQueen, C. M., and Burnard, L. (Eds.) 1994. Guidelines for electronic text encoding and interchange. Technical Report, Chicago, USA; Oxford, UK: Text Encoding Initiative.

Spohr, D. 2012. *Towards a Multifunctional Lexical Resource: Design and Implementation of a Graph-Based Lexicon Model*, vol. 141. Berlin, Germany; Boston, USA: De Gruyter. `http://www.degruyter.com/view/product/179961`.

Studer, R., Benjamins, V. R., and Fensel, D. 1998. Knowledge engineering: principles and methods. *Data and Knowledge Engineering* **25**(1–2): 161–197. http://www.sciencedirect.com/science/article/pii/S0169023X97000566.

Tchechmedjiev, A., Sérasset, G., Goulian, J., and Schwab, D. 2014. Attaching translations to proper lexical senses in DBnary. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics (LDL)*, pp. 5–12. Reykjavik, Iceland. Paris, France: European Language Resources Association (ELRA).

Toral, A., and Monachini, M. 2007. SIMPLE-OWL: a generative lexicon ontology for NLP and the Semantic Web. In *Workshop on Cooperative Construction of Linguistic Knowledge Bases (AIIA 2007)*. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.100.7798.

Trippel, T. 2006. *The Lexicon Graph Model: A Generic Model for Multimodal Lexicon Development*. Saarbrücken, Germany: AQ-Verlag.

Vila-Suero, D., Gómez-Pérez, A., Montiel-Ponsoda, E., Gracia, J., and Aguado-de-Cea, G. 2014. Publishing linked data on the web: The multilingual dimension. In P. Buitelaar and P. Cimiano (ed.), *Towards the Multilingual Semantic Web*, pp. 101–117. Berlin, Heidelberg, Germany: Springer.

Villegas, M., and Bel, N. 2015. PAROLE/SIMPLE 'lemon' ontology and lexicons. *Semantic Web* **6**(4): 363–369. http://www.semantic-web-journal.net/content/converting-parole-simple-clips-lexicon-rdf-lemon-0.

Villegas, M., Melero, M., Bel, N., and Gracia, J. 2016. Leveraging RDF graphs for crossing multiple bilingual dictionaries. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 868–876. Portoroz, Slovenia. Paris, France: European Language Resources Association (ELRA).

Vossen, P., Bloksma, L., Rodriguez, H., Climent, S., Calzolari, N., Roventini, A., Bertagna, F., Alonge, A., and Peters, W. 1998. The EuroWordnet base concepts and top ontology. Technical Report. Deliverable D017 D 34:D036. WP5 Project EuroWordNet, LE2-4003.

Westerski, A., and Sánchez-Rada, J. F. 2013. Marl ontology specification, V1. 0 May 2013. Web. http://www.gsi.dit.upm.es/ontologies/marl/.

Westphal, P., Stadler, C., and Pool, J. 2015. Countering language attrition with PanLex and the Web of Data. *Semantic Web* **6**(4): 347–353. http://doi.org/10.3233/SW-140138.

Windhouwer, M. 2012. RELcat: a relation registry for ISOcat data categories. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, pp. 3661–3664. Istanbul, Turkey. Paris, France: European Language Resources Association (ELRA).

Windhouwer, M., and Wright, S. E. 2012. Linking to linguistic data categories in ISOcat. In C. Chiarcos, S. Nordhoff and S. Hellmann (ed.), *Linked Data in Linguistics*, pp. 99–107. Berlin, Heidelberg, Germany: Springer.

Wolf, F., and Gibson, E. 2005. Representing discourse coherence: a corpus-based study. *Computational Linguistics* **31**(2): 249–287.