Departamento de Inteligencia Artificial
Escuela Técnica Superior de Ingenieros Informáticos

PhD Thesis

Linguistic Linked Data for Lexicography

Author: Julia Bosque Gil

Supervisors: Dra. Asunción Gómez Pérez
Dr. Jorge Gracia del Río

September, 2019

Tribunal nombrado por el Sr. Rector Magfco. de la Universidad Politécnica de Madrid,
el día — de — de 2019.

Presidente: Dr. Presidente Presidente Presidente

Vocal: Dr. Vocal Vocal Vocal

Vocal: Dr. Vocal2 Vocal2 Vocal2

Vocal: Dr. Vocal3 Vocal3 Vocal3

Secretario: Dr. Secretario Secretario Secretario

Suplente: Dr. Suplente Suplente Suplente

Suplente: Dr. Suplente2 Suplente2 Suplente2

Realizado el acto de defensa y lectura de la Tesis el día XX de MMMMM de YYYY en
la Escuela Técnica Superior de Ingenieros Informáticos

Calificación: _____

EL PRESIDENTE

VOCAL 1

VOCAL 2

VOCAL 3

EL SECRETARIO

A mis padres y a José
A la abuela Julia y al abuelo Pascual

Agradecimientos

Ricardo Mairal called me an afternoon to recommend me that I consider the OEG to pursue a doctorate. He could not have known that his call would initiate a whole new phase in my (academic) life or that his name would be heading this document. Thank you.

This thesis would not have been possible without Asun and Óscar taking me as part of the OEG and their support throughout these years. Special thanks to the LLD group that received me back then, Asun, Lupe, Elena, and Jorge. I am very grateful to Lupe and Elena for being my companions in linguistics during all this time. Special thanks go to my supervisors, Asun and Jorge, for their help, experience, patience and support.

I would like to express my gratitude towards Ilan Kernerman and the whole K Dictionaries team for giving me the opportunity to work with their data. Many thanks go also to Philipp Cimiano and everyone at CITEC (specially Angelika, Basil and Frank) for welcoming me in Bielefeld and providing me with valuable insights for the thesis and the *lexicog* module. I extend these acknowledgements to Christian Chiarcos and everyone at ACoLi in Frankfurt for their help and feedback in the modelling of dictionaries, which infused me with a renewed motivation for the last year of the program. Special thanks here to Max, Jesse, Ilya and Nico for making me feel at home and part of ACoLi still. I would also like to thank everyone involved in the SD-LLOD datathons, and the W3C OntoLex community for their engagement in numerous telcos.

Sometimes the daily pre-doctoral life is the steepest mountain. I am grateful to everyone at the OEG for making me part of the group, our lunches, Slack, Christmas' dinners. I would need much more space to include all of you, so I will reduce this to a small set of names: thanks to Pablo for his invaluable support and that *unique* humour since my first day at UPM. Thanks to Carlos for challenging me with his conversation and touching music. To María Poveda for all her trust, expertise, boldness and oh, her

wit! To Idafen, who sent the perfect rope to help me get out of the valley of despair. To Alcázar for IT support anytime and place, as a true Time Lord. To Jarg and Ana for captaining me through the ocean of bureaucracy. To my indefatigable family of pre-docs and junior post-docs, for bringing me support and academic help, but also word plays, Hawaiian Fridays, tea corners, sugar overdoses, crochet, ambitious soccer and so, so much intensity. To Filip, Esther, José Luis, Garijo, Vila, Rico, Víctor, Miguel, Nandana, Freddy, Esteban, Edna, Paco...

Star Wars fans know that *padawans* travel the galaxy alongside *Jedi* masters. Jorge has guided me and has always been there to listen and help. I am grateful for all those hours of feedback, but specially for his respect, affection and encouragement. If I ever have a PhD student in the future, I will strive to make their initial journey through the scholarly hyperspace as academically and personally enriching as he has made mine.

To Isa, Amaranta and Max for all their friendly advice and emotional assistance, specially in the most difficult moments, when nerves and anxiety eclipse everything. To my future in-laws, Paloma(s) and Pepe, for their love, support and interest in everything I do.

I would not have entered or stayed in Academia without my parents' inspiration and support, anytime. Their views on linguistics, the scholarly world and life in general, together with their faith in me, have always been driving me and still prevent me from giving up when I feel overwhelmed. They are both the ultimate example I follow, with which I never lose my track.

Lastly, thanks to José. For being my partner in life since a long time ago, in cities far, far away, as my daily enduring mithril coat and my deflector shield. With him by my side, I feel I can surmount any obstacle. I hope to have been to him throughout these years what he has been to me. Overcoming all the difficulties we encountered since we first set foot on university hallways and finishing our PhDs at the same time reminded me of Bilbo's words at the end of *The Hobbit* (Tolkien, 1997): "*so comes snow after fire, and even dragons have their ending!*". Because, though *hic sunt dracones*, the Force is strong with these two.

Abstract

Nowadays, the number of resources that provide lexical data keeps significantly increasing as outcomes of projects in linguistics, lexicography and language technologies. However, this data is scattered throughout the Web, isolated, and often comes in a vast number of different formats and languages. To address this landscape of heterogeneous and isolated language resources, experts working in the domain of the Semantic Web have adopted approaches to linguistic data representation based on the Linked Data (LD) paradigm, giving birth to the Linguistic Linked Data (LLD) line of research. Although LLD is focused on the representation, publication and sharing of language resources, there exists no previous wide-scope exploration and assessment of the impact of the application of LLD to lexicography as a discipline: the requirements and process this involves, its practical and theoretical benefits, the challenges it raises, and the open problems on the way. Furthermore, as a required ingredient towards this exploration, guidelines to represent a wide range of lexicographic resources (as outcomes of a lexicographic compilation process) by following this new paradigm are lacking as well. In this thesis we address the application of LLD to lexicography from the looking glass of the lexicographer, the user who consults lexicographic works, or the linguist interested in lexical semantics who needs lexicographic content for their work. We detect and resolve obstacles on the way for LLD adoption in lexicography regarding the representation requirements of lexicographic works through the definition of application profiles and extensions of the *de facto* standards for LLD representation. On the basis of a set of representative resources that we convert to the Resource Description Framework (RDF), we analyse and showcase the benefits and implications of LLD for dictionary representation, both as a target format of a conversion, as well as a potential native format for lexicographic projects in the future.

Resumen

Con el incesante aumento de los recursos léxicos que surgen de numerosos proyectos en lingüística, lexicografía, y tecnologías del lenguaje, hoy en día los datos léxicos se encuentran en distintos formatos, dispersos y aislados unos de otros en la Web. Los Datos Enlazados Lingüísticos (por sus siglas en inglés, LLD) es una línea de investigación desarrollada por expertos en el campo de la Web Semántica que responde a la necesidad de estandarización en la representación de datos lingüísticos y que se basa en el paradigma de los Datos Enlazados (LD). Pese a que la línea de LLD se centra en la representación, la publicación, y la difusión de los recursos lingüísticos, no existe hasta la fecha un estudio amplio ni una valoración del impacto que tendría su aplicación a la lexicografía como disciplina: cuáles son los requisitos que cumplir en la representación de recursos lexicográficos como LLD, qué procesos habría que llevar a cabo, cuáles serían las ventajas prácticas y teóricas de este tipo de representación, los desafíos a los que daría lugar, ni los posibles problemas a los que habría que hacer frente. Asimismo, como piezas necesarias en ese estudio, destaca también la falta de guías para representar un amplio abanico de recursos lexicográficos en este nuevo paradigma. En esta tesis doctoral se investiga la aplicación de los LLD a la lexicografía desde la perspectiva del lexicógrafo, el usuario de recursos lexicográficos, o el lingüista interesado en la semántica léxica que necesita acceder a contenido lexicográfico para su trabajo. Esta tesis identifica y resuelve una serie de problemas de modelado a la hora de representar contenido lexicográfico en el formato RDF (Resource Description Framework). Mediante la definición de perfiles de aplicación y extensiones para el estándar *de facto* más utilizado en LLD, este trabajo presenta una serie de recursos lexicográficos en formato RDF que sirven para analizar y demostrar las ventajas de este paradigma para codificar información lexicográfica, tanto como formato final de un recurso tras una conversión, como como formato nativo para la creación de nuevas obras lexicográficas.

Contents

1	Introduction	1
1.1	Technological context	3
1.2	Dictionaries, lexica and terminologies	11
1.3	Structure of the document	22
1.4	Derived outcomes	23
1.5	Scholarships	25
1.6	Research stays	25
1.7	Projects	26
2	Models for the Representation of Linguistic Linked Data	27
2.1	Overview	28
2.2	Methodology of analysis	32
2.3	Core models for lexical resources	33
2.4	Level-specific and theory based models	40
2.5	Area-specific models	53
2.6	Linguistic category registries	59
2.7	Corpora- and service oriented models	64
2.8	Limitations and open challenges in LLD representation	66
3	Goals and Contributions	77
3.1	Goals	77
3.2	Assumptions	80
3.3	Hypotheses	80
3.4	Contributions	81
3.5	Restrictions	83

3.6	Research methodology	85
4	LLD in Lexicography: Challenges	91
4.1	Characteristics of bilingual dictionaries	93
4.2	Characteristics of multilingual terminologies	101
4.3	Characteristics of multilingual dictionaries	108
4.4	Characteristics of multilingual encyclopedic resources	118
4.5	Summary	120
5	The Representation of Lexicographic Resources as LLD	123
5.1	Methodology for model definitions	124
5.2	The Apertium series as RDF	125
5.3	Rayfield’s Comprehensive Georgian-English Dictionary as RDF	130
5.4	Termesp and Termcat as RDF	144
5.5	K Dictionaries Multilingual Global Series as RDF	164
5.6	The lexicalisation of Zhishi.me into Zhishi.lemon	191
5.7	Summary	193
6	<i>lexicog</i>: A Lexicography Module for OntoLex-<i>lemon</i>	197
6.1	Motivation	198
6.2	<i>lexicog</i> : a module for lexicography	205
6.3	Validating <i>lexicog</i> with K Dictionaries Global series	220
6.4	Summary	231
7	Evaluation	233
7.1	Evaluation methodology	233
7.2	Conformance of RDF data	236
7.3	Information preservation	241
7.4	Benefits of LLD for lexicography	247
7.5	Reuse of the generated outcomes by external projects	256
7.6	Queries over complementary data	257

8	Linked Data-based lexicography	265
8.1	LD and lexical systems	265
8.2	LD-native lexicography	268
8.3	Summary	274
9	Conclusion and Future Work	275
9.1	Review of contributions	276
9.2	Verification of hypotheses	280
9.3	Implications and social impact	283
9.4	Future work	284
	Bibliography	287

List of Figures

1.1	The Semantic Web Stack	6
1.2	The LLOD cloud	10
1.3	Semasiological and onomasiological access	16
1.4	The entry <i>credulous</i> (Random House Webster's)	17
1.5	The entry <i>damJere</i> (Rayfield's Georgian-English Dictionary)	18
2.1	Levels of linguistic description and their units of analysis	33
2.2	The lexical entry <i>book</i> modelled with OntoLex	38
2.3	The lexical entry <i>book</i> modelled with the OGL ontology	39
2.4	The lexical entry <i>book</i> modelled with TELIX	40
2.5	The lexical entry <i>hesitate</i> modelled with PreMOn	44
2.6	The lexical entry <i>hesitate</i> modelled with PDEV-lemon	46
2.7	The lexical entry <i>hesitate</i> modelled in Parole-Simple Clips	48
2.8	The lexical entry <i>hesitate</i> modelled in WordNet RDF (2014)	49
2.9	The lexical entry <i>hesitate</i> modelled in DBnary	57
2.10	A small section of the GOLD ontology	60
2.11	A small section of LexInfo	61
2.12	A small section of OLiA's Refence Model	62
3.1	Limitations, goals and research lines	79
3.2	Goals, hypotheses and assumptions	82
3.3	Contributions, hypotheses, goals, restrictions and assumptions	86
4.1	Merged information on the entry <i>divorce</i> from Zhishi.me	121
5.1	The <i>lemon</i> translation module	126

5.2	RDF representation of the entry <i>rouge</i> from Apertium FR-ES RDF . . .	127
5.3	The OntoLex- <i>lemonvartrans</i> module	131
5.4	The modelling of entries without parentheses in the lemma (case 1) . . .	135
5.5	The modelling of entries with parentheses in the lemma (case 2)	136
5.6	The modelling of entries with brackets in the lemma (case 3)	137
5.7	The modelling of entries with case endings in the lemma (case 5)	137
5.8	An entry with two headwords (Rayfield’s Georgian-English)	138
5.9	RDF representation of the entry <i>Adam</i> shown in Example 5.5	141
5.10	RDF representation of a portion of Example 5.10 as RDF	143
5.11	URI naming strategy for Rayfield’s Georgian-English Dictionary	143
5.12	RDF representation for the entry <i>admitancia cinética</i> (Terminesp, mono- lingual)	146
5.13	RDF representation of the entry <i>admitancia cinética</i> (Terminesp, multi- lingual)	147
5.14	RDF representation of the entry <i>anarcardo</i> (Terminesp)	149
5.15	RDF representation of the entry ‘administrador -a de xarxes’ (Termcat) .	155
5.16	The representation of translations in Termcat RDF	159
5.17	RDF representation of the entry <i>red</i> (Terminoteca RDF)	162
5.18	Phases of Terminesp, Termcat and Terminoteca RDF	163
5.19	RDF representation of the entry <i>arriesgado</i> (K Dictionaries)	167
5.20	RDF representation of the entry <i>alcachofa</i> (K Dictionaries)	172
5.21	RDF representation of the entry <i>arriesgado</i> (K Dictionaries)	175
5.22	RDF representation of the compositional phrase <i>earn a living</i> (K Dictio- naries)	176
5.23	Portion of the RDF representation of the entries <i>besuchen</i> , <i>Besuch</i> and <i>Besucher</i> (K Dictionaries)	181
5.24	A portion of the RDF representation of the entry <i>abadía</i> (K Dictionaries)	185
5.24	RDF representation of the entry <i>divorce</i> (Zhishi.me, monolingual)	192
5.25	RDF representation of the entry <i>divorce</i> (Zhishi.me, multilingual)	193
6.1	The OntoLex lexicography module (<i>lexicog</i>)	207
6.8	Usage examples and their translations in the entry <i>monastery</i> (WR English- Spanish Dictionary)	219

7.1	Extract of the results of the query shown in Example 7.9	259
7.2	Extract of the results of the query shown in Example 7.11	263

List of Tables

2.1	Group 1: Models to represent the main elements of lexical resources . . .	41
2.2	Group 2a: Models for different levels of linguistic analysis	52
2.5	Other Models	66
2.3	Group 2b: Models for branches of applied linguistics and related areas .	73
2.4	Group 3 Models	75
4.1	Features of the selected resources (Atkins and Rundell (2008))	93
4.2	Terminesp entry for the Spanish term <i>admitancia cinética</i>	104
4.3	Data provided in Terminesp and Termcat	107
4.4	Representation requirements for each lexicographic resource	122
5.1	Apertium RDF URI naming strategy	129
5.2	Terminesp RDF URI naming strategy	151
5.3	Distribution of POS categories and syntactic phrases in Terminesp Span- ish entries	152
5.4	Termcat RDF URI naming strategy	160
5.5	Distribution of entries in the Terminoteca	162
5.6	DTD elements and their RDF counterparts and modelling	168
5.7	SenseIndicator , GeographicalUsage , and their RDF mappings	172
5.8	RangeOfApplication , Register , and their RDF mappings	175
5.9	CompositionalPhrase and its RDF mapping	177
5.10	InflectionCtn , InflectionBlock , NestEntry and their RDF mappings	182
5.11	TraslationCtn and its RDF mappings	186
5.12	URI naming strategy for KD Multilingual global series RDF	189
7.1	Coverage in <i>lemon</i> , <i>OntoLex-lemon</i> and <i>LexInfo</i>	244

7.2	Comparison of # of elements in KD source data and RDF	246
7.3	Results of the query in Example 7.6	250
7.4	Results of query in Example 7.7	252
7.5	Results of the query shown in Example 7.8	253
7.6	Results of the query provided in Example 7.10	261

Chapter 1

Introduction

We have been used to dictionaries surrounding us since we started learning to consult them in primary school as part of our language courses. Nowadays, digitised dictionaries surround us in different ways by being integrated in the applications of our mobile devices, in our spell-checkers and personal assistants, or as resources we access online when writing a paper or an email in a foreign language. Dictionaries are language resources. These, according to the European Language Resource Association (ELRA)¹, are defined as follows:

“The term Language Resource refers to a set of speech or language data and descriptions in machine readable form, used for building, improving or evaluating natural language and speech algorithms or systems, or, as core resources for the software localisation and language services industries, for language studies, electronic publishing, international transactions, subject-area specialists and end users.”

The data contained in dictionaries, and other language resources such as terminologies, conveys information about the vocabulary of a language (or set of languages) or domain of knowledge, its so-called *lexicon*. Data about a lexicon is thus *lexical data*, and language resources that contain lexical data (lexica, thesauri, dictionaries) and are product of a compilation, writing and editing process are called *lexicographic resources* (more information on lexicographic resources is provided in this chapter, Section 1.2. As the number of resources that provide lexical data keeps significantly increasing as

¹<http://www.elra.info/en/about/what-language-resource/>. Last accessed on 17.08.2019.

outcomes of projects in linguistics, lexicography and language technologies, nowadays lexical data and their sources are scattered throughout the Web, and come in a vast number of different formats and languages. In addition to this heterogeneity, resources are often isolated in their data silos, described in language catalogues with varying classification, and hosted on the Web with different levels of availability, access, and license. Integrating complementary lexical data would allow to enrich resources for further consumption (both by humans or by Natural Language Processing (NLP) applications, among other potential uses) through language technology services and user interfaces, but the problematic scenario described above makes their discovery, reuse, and integration a non-trivial task.

To address this landscape of heterogeneous and isolated language resources, experts working in the domain of the Semantic Web have adopted approaches to linguistic data representation based on the Linked Data paradigm, giving birth to the Linguistic Linked Data (LLD) line of research. The Semantic Web emerges as an extension of the Web that aims to capture the knowledge about the objects and their relations encoded in the data, i.e, the meaning of the data. The Linked Data principles propose guidelines to expose, share and connect such data on the Web. In this context, LLD is focused on the representation, publication and sharing of language resources. However, there exists no previous wide-scope exploration and assessment of the impact of the application of LLD to lexicography as a discipline: the requirements and process this involves, its practical and theoretical benefits, the challenges it raises, and the open problems on the way. Furthermore, as a required ingredient towards this exploration, guidelines to represent a wide range of lexicographic resources (as outcomes of a lexicographic compilation process) by following this new paradigm are lacking as well.

This thesis is interdisciplinary and is framed in the interface between **lexicography**, the field concerned with the compilation and edition of dictionaries, and artificial intelligence (AI), specifically, the **Semantic Web and Linked Data** areas. It aims to provide a framework for the use and adoption of Linked Data technologies in the field of lexicography by addressing the gaps mentioned above through the following list of contributions:

1. An analysis of the problems of lexicographical resources representation with the current *de facto* standards for LLD representation.

2. Solutions to these problems and proposals for community initiatives to extend such models in the form of a new module for the *de facto* standard OntoLex-*lemon*.
3. New application profiles of OntoLex-*lemon*, extensions to it, and practices and methodologies for the conversion of lexicographical resources to the formats of the Semantic Web.
4. A series of datasets converted to LLD from a set of representative lexicographical resources to validate current models, the proposed application profiles and the suggested extensions.
5. An evaluation of the impact of LD technologies in lexicography and in the use of lexicographic resources to gather linguistic data (to be used by linguists, translators or non-expert users)

In this way, a lexicographer will not only find a potential motivation for the representation of their lexicographic resource in a way that maximises its discovery, reuse and integration with other sources (language resources or general knowledge bases), but also a walk-through on how to do so and what kind of problems they are likely to face. In fact, this thesis focuses on what P.H. Swanepoel described in relation to the computer age over twenty years ago, which still holds if we transpose it now to the Semantic Web and Linked Data age:

“[The new computer age] has also given birth to new kinds of dictionaries, a variety of new lexicographic methodologies and techniques, and a range of new disciplines that concern themselves with the compilation of dictionaries (in the widest possible sense of the word) and the theoretical and practical problems they pose.” (Swanepoel (1994))

As these outcomes and approach suggest, the ultimate goal of this thesis is to build a bridge between the two realms, lexicography and the Semantic Web.

1.1 Technological context

Before addressing the representation needs of lexicographic resources as LLD, in this section we introduce the basic notions of the Semantic Web and LLD.

1.1.1 The Semantic Web

The Semantic Web emerges as 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). This initiative responds to the lack of interoperability across applications and services that are developed independently and that consume Web information. It is grounded on the idea of capturing the knowledge about the objects and their relations encoded in the data, as opposed to establishing relations between documents describing this knowledge. This turns the Web into a knowledge base, in contrast to a Web of documents, where software agents can inter-operate and exchange data with each other. For this **interoperability** to be assured (in the case of linguistic data, also at the multilingual layer), data have to be exchangeable and consumable in two levels: the syntactic, concerned with the format of the data; and the semantic, concerned with the meaning encoded in the data. **Syntactic interoperability** “relies thus on specified data formats, communication protocols, and the like to ensure communication and data exchange” (Ide and Pustejovsky, 2010), whereas semantic interoperability has been stated to exist “when two systems have the ability to automatically interpret exchanged information meaningfully and accurately in order to produce meaningful results via deference [or value access] to a common information exchange model” (Ide and Pustejovsky, 2010). Syntactic interoperability between agents is achieved by relying on a common framework to represent the data: the Resource Description Framework (RDF)², which describes data through statements in the form of **triples** following the pattern *subject - predicate - object*. This simple model allows structured and semi-structured data to be mixed, exposed, and shared across different applications. The triples can be serialised in different formats such as Turtle³ or JSON-LD.⁴

If RDF accounts for the syntactic interoperability of agents that consume data represented in this format, **semantic interoperability** is achieved by the use of **ontologies** to describe the meaning encoded in the data itself. However, achieving this interoperability across language barriers is not a trivial task in itself (Vila-Suero et al., 2014a) (see Section 1.1.2). An ontology is properly defined as a “formal, explicit specification of a shared conceptualization” (Studer et al., 1998). This definition refers to an ontology

²<http://www.w3.org/RDF/>

³<http://www.w3.org/TR/turtle/>

⁴<http://www.w3.org/TR/json-ld/>

formalised as a machine-readable, explicit description of a domain common to a group of domain experts. For instance, a dataset containing film records, casting information and premiere dates could be represented in RDF by using the Movie Ontology (MO)⁵, which describes the concepts (e.g. movie, director, actor, etc.) and relations between them (e.g. a movie has a director, an actor stars a movie, etc.) that belong to this particular domain. Ontologies allow to state this knowledge in the form of statements that will be encoded as RDF triples, e.g.: `<Movie> <hasDirector> <Director>`. In this example, the actual dataset with movie records could be encoded following this mechanism: `<Alien> <hasDirector> <Ridley_Scott>`. There are also cross-domain ontologies such as the DBpedia ontology⁶ allowing to state general information and relations about people, locations, events, human artifacts, etc. Example 1.1, described in detail below, showcases the use of the DBpedia ontology to capture general information about the birthplace of a person. The meaning ontologies provide is encoded by using the Ontology Web Language (OWL)⁷ and other formalisms and schemes such as RDF Schema⁸, all of them compliant with the RDF framework.

To better illustrate the role of ontologies in the Semantic Web, Figure 1.1⁹ shows the Semantic Web Stack, which depicts the architecture of the Semantic Web as a hierarchy of languages, where each layer leverages and extends the features of the layers below (Berners-Lee, 2003; Horrocks et al., 2005). The ingredients to account for syntactic interoperability are represented at the base, with RDF, XML and Unique Resource identifiers (URIs) (see Section 1.1.2) being the core formalisms. OWL and RDF Schema provide semantic interoperability features to the base layers, while the SPARQL Protocol and RDF Query Language (SPARQL)¹⁰ serves to query the data represented as RDF with OWL and RDF-Schema.

In practice, this description of the meaning encoded in data is done by converting a resource to RDF and instantiating the ontology with the data in such a resource. Example 1.1 (in Turtle serialisation), taken from the resource DBpedia¹¹, is an RDF

⁵<http://www.movieontology.org/>

⁶<https://wiki.dbpedia.org/services-resources/ontology>

⁷<https://www.w3.org/OWL/>

⁸<https://www.w3.org/TR/rdf-schema/>

⁹Retrieved from [http://www.w3.org/2007/Talks/0130-sb-W3CTechSemWeb/#\(24\)](http://www.w3.org/2007/Talks/0130-sb-W3CTechSemWeb/#(24)), W3C. Last accessed on 17.08.2019

¹⁰<https://www.w3.org/TR/rdf-sparql-query/>

¹¹<http://dbpedia.org>

Figure 1.1: The Semantic Web Stack

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

Example 1.1: A triple stating that Francisco de Goya was born in Fuendetodos.

statement about Francisco de Goya (a famous Spanish painter) representing that “Francisco de Goya was born in Fuendetodos”. In this example, *dbr* and *dbo* are the prefixes or shorthands for the namespaces of the resource and the ontology respectively, <http://dbpedia.org/resource/> and <http://dbpedia.org/ontology/>. These namespaces are the domains where the elements *Francisco_de_Goya* and *Fuendetodos* on the one hand (from the *dbr* dataset), and the property *birthPlace* on the other (from the *dbo* ontology), are uniquely identified at a Web scale.

With this information encoded in RDF, notice that, since another triple in DBpedia states that

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

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

As in the previous example with “real live” entities (Goya, country, birthplace, etc.), lexical data can be described in RDF format as well by using domain ontologies.


```
apertiumEN:country-n-en a lemon:LexicalEntry ;  
lexinfo:partOfSpeech lexinfo:noun .
```

Example 1.2: Two triples with the same subject indicating that the word *country* from an Apertium English dictionary is a lexical entry and that it has *noun* as part-of-speech.

For instance, an Apertium dictionary (Forcada et al., 2011; Gracia et al., 2018)) in its source format describes, among other information, that the word **country** is a lemma in that dictionary, and that it has as part-of-speech the value *noun*. This information comes thus from the original resource. As the DBpedia ontology for cross-domain information and the Movie Ontology for the film domain, the *lemon* model (McCrae et al., 2012) offers elements to describe in RDF lexical items and how they relate to each other. In addition, there are numerous linguistic categories used to describe linguistic data and phenomena, so that registries gathering them have also been developed, e.g. LexInfo¹² and are used together with the *lemon* model (for details on the various models to represent LLD, see Chapter 2). This means that, in the same way we described Francisco de Goya to have Fuendetodos as birthplace in Example 1.1, Example 1.2 characterises in RDF the word *country* as a resource of the Apertium dictionary that is a lexical entry (by instantiating the *lemon* ontology), and indicates that it has *noun* as part-of-speech through the use of elements from the LexInfo category registry specially devised for this purpose.¹³ In addition, Example 1.2 illustrates how information coming from different resources (i.e., Apertium RDF¹⁴ and LexInfo), available in different Web spaces and built by different people for different purposes, have been interlinked in a rather straightforward way.

However, some clarification on the term *semantic* in *Semantic Web* may be necessary, as it is a potential place of misunderstanding for linguists and lexicographers when first encountered with the Semantic Web. The reason behind this is the interpretation of the word *semantic* in relation to the homonymous sub-field of linguistics concerned with the encoding of meaning in **natural language**. In contrast to a *semantic analysis* from the sub-field of *semantics* in linguistics, the adjective *semantic* in *Semantic Web* does

¹²<http://www.lexinfo.net>

¹³Where *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.

¹⁴<http://linguistic.linkeddata.es/apertium/>

not necessarily refer in this context to the systematic description of how the meaning of linguistic expressions emerges; that is, how it is (compositionally) built through the combination of the different linguistic units, or how linguistic phenomena affect it. Stating that `dbr:Francisco_Goya` `dbo:birthPlace` `dbr:Fuendetodos` makes the meaning of the data explicit through the use of the ontology. Even though this RDF statement describes the *semantics of the data*, it does not account for the semantics of the natural language in which that piece of knowledge is verbalised. For example, it gives no information on how this knowledge is expressed in natural language, whether we use *to be originally from* or *to be born in*, what kind of event it denotes, what semantic role *Francisco de Goya* and *Fuendetodos* would have in such a sentence as *Francisco de Goya was born in Fuendetodos*, and so on.

1.1.2 Linguistic Linked Data

The previous section dealt with semantic and syntactic interoperability between applications and services as key aspects to achieve better data exchange and integration capabilities, as well as reasoning. In addition to these two conditions, in order to solve the problem of vast data resources being isolated from one another and developed independently in their data silos, which hinders their discovery and reuse, the notion of *linking* resources comes into play. The Linked Data (LD) paradigm emerges as a series of best practices and principles for exposing, sharing, and connecting data on the Web (Bizer et al., 2009), independently of the domain. These principles state that **unique resource identifiers** (URIs) (Berners-Lee et al., 2004) 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 is transforming the Web into a vast cloud of interlinked data (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.

Dictionaries, terminologies and lexica are a subset of language resources providing lexical data. However, they come in numerous different formats, are often developed independently, and allow for different levels of access (from closed, paid resources to API-accessed ones and fully open ones, for example). In this scenario, resources are at the end isolated from each other despite that many of them contain complementary

information. Thus, a community of language technology experts and practitioners has started to adopt techniques coming from the area of LD to ease the publication, sharing, discovery, exchange and integration of linguistic data both for NLP systems and for linguistic research. Projects such as the EU’s *Linked Data as an enabler of cross-media and multilingual content analytics for enterprises across Europe* (LIDER)¹⁵ and the Open Framework of e-Services for Multilingual and Semantic Enrichment of Digital Content (FREME) project have driven major advancements in LLD. The Linguistic Linked Open Data initiative (LLOD) (Chiarcos et al., 2012a,b; McCrae et al., 2016), which is a line in LD focused on language resources specifically, highlights the *open* nature of these resources, while the term *Linguistic Linked Data* (LLD) refers, in general, to LD applied to language resources without stressing any license details about them.

The so-called LLOD Cloud¹⁶ (McCrae et al., 2016), represented in Figure 1.2¹⁷ is a visualisation of LLOD datasets available at the time of writing (June 2019) and registered at Datahub.io¹⁸, or uploaded to the general Linked Open Data Cloud.¹⁹

Figure 1.2 reflects the interest in this new paradigm among experts in the Semantic Web community. As one can see, much of this interest has been devoted to the publication of lexica, dictionaries, terminologies and knowledge bases (coloured in green tones in the diagram) to RDF, along with Universal Dependencies corpora (shaded in blue).

For this linguistic information to reach this state as LLD resource, there are several steps to undertake (Gómez-Pérez et al., 2013; Villazón-Terrazas et al., 2012).²¹ The steps for LLD generation can be broadly summarised as follows: 1) source data and

¹⁵<http://lider-project.eu/>

¹⁶ <http://linguistic-lod.org/>

¹⁷Diagram developed by John P. McCrae, Luca Matteis, Christian Chiarcos, Sebastian Hellmann and Sebastian Nordhoff. “Open Linguistics Working Group (2012)”, The Linguistic Linked Open Data cloud diagram (draft), version of June 2019, <http://linguistic-lod.org/llod-cloud>.

¹⁸<https://old.datahub.io/>

¹⁹<http://lod-cloud.net/>

²⁰Available at <http://linguistic-lod.org/llod-cloud>.

²¹For a detailed walk-through of the process of LLD generation, we refer the reader to guidelines for multilingual data generation and publication (Gómez-Pérez et al., 2013; Vila-Suero et al., 2014a), which build upon a series of general steps identified for LD publication (independently of the domain) (Villazón-Terrazas et al., 2012). Additional guidelines for LLD generation are suggested as best practices in the W3C Best Practices for Multilingual Linked Open Data community group (<http://www.w3.org/community/bpmlod/>), and the LIDER project guidelines and reference cards (<http://www.lider-project.eu/guidelines>), which cover a range of language resources.

source data model analysis, 2) URI 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, being existent ontologies and vocabularies the recommended options (Gómez-Pérez et al., 2013). Once the model to reuse is decided upon or a new model is created for a particular use case, and the URI naming strategy has been defined, the conversion to RDF (in any of its serialisations) proceeds.²² The resulting RDF resource can be now linked to external domain independent language resources, domain specific resources (e.g. terminologies) or general conceptual LD resources (e.g. BabelNet²³) through different strategies²⁴, namely, interlingual identity links, interlingual soft links, or a common index (e.g. BabelNet or DBpedia). The publication of the LLD resource and its metadata (including license information) constitutes the final step.

This thesis revisits steps 1, 2, and 3 of the methodology for LLD generation for a set of lexicographic resources, the three of them related to *linguistic knowledge representation* as RDF. Since the LLD generation process and its modelling stage depend highly on the kind of resource that is being converted and its representation needs, we provide an overview of the basic notions of lexicographic works, their structure and content, in the following section.

1.2 Dictionaries, lexica and terminologies

In the previous section we have explained the technological context which serves as framework for this dissertation. This section will delve into the kind of data we analysed and transformed in this technological context, i.e, lexica, dictionaries, and terminologies, and will also present some background information on the discipline that is concerned with the creation of lexicographic works – lexicography.

²²A list of tools and frameworks for RDF conversion of data is provided at <http://www.w3.org/wiki/ConverterToRdf>.

²³<http://babelnet.org/>

²⁴See Matteis, L. and Navigli, R. LIDER Project Reference Card *How to publish Linguistic Linked Data*.

1.2.1 Basic terminology

Depending on the specific resource, its intended use and the branch of knowledge where it has been developed or compiled, there exist nuances in the use and definitions of the terms *dictionary*, *lexicon* and *terminology*. For example, in the literature in lexicography and computational linguistics, *lexicon* is often applied to resources developed for NLP (or automatically generated as a result of an NLP effort) while *dictionary* sometimes refers to resources developed for human users, be they machine readable or not (cf. Bindi et al. (1994); Copestake (1992); Romary (2013a); Swanepoel (1994); Villegas and Bel (2015)). However, there are examples in which the terms are used similarly as well (cf. ‘valency lexicon’ and ‘valency dictionaries’ (Przepiórkowski et al., 2014)).

Throughout this work we will be using the term *dictionary* if the name of the resource states so, or to refer, in general, to resources that match the following definition from the The Oxford English Dictionary (OED, 2018) and have been developed in a lexicographical work-flow for humans or for humans and machines alike:

dictionary *1. A book or electronic resource that lists the words of a language (typically in alphabetical order) and gives their meaning, or gives the equivalent words in a different language, often also providing information about pronunciation, origin, and usage. 1.1 A reference book on a particular subject, the items of which are typically arranged in alphabetical order. 1.2 (Computing) A set of words or other text strings made for use in applications such as spellcheckers.*²⁵ [OED, 2018]

In the case of lexica, we will adjust to the following definition:

lexicon *A lexicon is a collection of information about the words of a language and about the lexical categories to which they belong. A lexicon is usually structured as a collection of lexical entries, like (“pig” N V ADJ). [...] In practice, a lexical entry will include further information about the roles the word plays, such as feature information - for example, whether a verb is transitive, intransitive, ditransitive, etc., what form the verb takes*

²⁵English Oxford Living Dictionaries (2018). [online] Available at: <https://en.oxforddictionaries.com/definition/dictionary> [Accessed 27 Sep. 2018].

(e.g. *present participle, or past tense, etc.*)²⁶ [*The Natural Language Processing Dictionary* by Bill Wilson, 2012]

Lastly, in this thesis we will refer to *terminology* both as a body of terms²⁷, following this definition from the OED,

terminology *The body of terms used with a particular technical application in a subject of study, profession, etc.*²⁸ [OED, 2018]

We will also use *terminology* to refer to a specific lexical resource with information about specialised terms (e.g. their use, their translation, their definition, their Latin denomination, examples, etc.) developed in the field of **terminology**, concerned with specialised concepts and terms. With a strong focus laid on the concepts, many terminological resources are concept-oriented, as opposed to the majority of current dictionaries at the time of writing, compiled from a semasiological perspective (lexical entry-oriented rather than concept-oriented).

Dictionaries, lexica and terminologies gather valuable information to be exploited by NLP systems in multilingual applications. In fact, systems that perform tasks such as text-to-speech synthesis or summarisation are in need of lexica, and machine translation tasks require access to multilingual lexical resources (Copestake, 1992). Indeed, the rich lexical information contained in dictionaries led over two decades ago to projects such as ACQUILEX (Copestake, 1992), focused on the semi-automatically acquisition of this information from machine-readable dictionaries in order to create richer resources for NLP purposes. But dictionaries are also a key component for other lexicographers and computational linguists that work in the creation of new resources, as well as for linguists who consult them to gather data for comparative analysis and, together with corpora,

²⁶The Natural Language Processing Dictionary (2012). [online] Available at: <http://www.cse.unsw.edu.au/~billw/dictionaries/nlpdict.html> [Accessed 27 Sep. 2018].

²⁷According to the OED, a *term* can be defined as “A word or phrase used to describe a thing or to express a concept, especially in a particular kind of language or branch of study.” [*term*. OED. <http://en.oxforddictionaries.com/definition/term>. Last accessed on 04.02.2019.]. On the definition of *term* according to different theories of terminology, see Warburton, K. (2014). *Terminology Standardisation, Terminology Management and Best Practices or What’s Going on With Terminology?* (Madrid: Asociación Española de Terminología). <http://www.aeter.org/wp-content/uploads/2014/10/Warburton-AETER14.pdf>

²⁸<https://en.oxforddictionaries.com/definition/terminology> [Accessed 1 Oct. 2018].

for hypothesis verification. This set of users grows if we include regular users of language resources, writers, language learners, translators and language service providers. With the resources scattered throughout the Web and not always registered in catalogues such as ELRA or Meta-Share²⁹, the heterogeneity of the Web makes it difficult to not only find them and integrate them in NLP systems, but even to reuse them in the creation of other lexicographical projects (or to find the license allowing so), leading to effort duplication. Since LLD emerged as a good solution, in order to convert this kind of resources to RDF we first analyse their content and structure so as to detect representation needs before undertaking the LLD modelling task.

1.2.2 Dictionaries, lexica and terminologies: structure and content

Though the lexicon of a language is shared among its native speakers, each dictionary is different from another. Atkins and Rundell (2008) provide a detailed list of the different aspects to bear in mind when planning a trade dictionary, which for our purposes serves to illustrate the variety of dictionaries and the heterogeneity in their content:

1. the dictionary's language (monolingual, bilingual – unidirectional or bidirectional –, multilingual),
2. its coverage (general language, encyclopedic and cultural material, terminology and sublanguages, specific area of language: collocations, phrasal verbs, idioms, etc.),
3. the dictionary's size (standard, concise, pocket edition) and the medium (print, electronic, web-based),
4. the dictionary's organisation (word-to-meaning or word-to-meaning-to-word),
5. the user's language,
6. the user's skills and
7. what the dictionary will be used for: decoding (understanding, translating from) or encoding (using, translating into, teaching). (Atkins and Rundell, 2008)

²⁹Meta-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)

It is important to distinguish here that, in addition to trade dictionaries, there are other kind of resources coming from the area of lexical semantics and NLP that result in (sometimes enormous) works in the form of a dictionary, a lexico-conceptual database, a lexicon, a corpus or a treebank, and which, in contrast to trade dictionaries for laymen, **belong to a theory of language**. As Trippel (2006) points out, these resources do not often have wide coverage but their information and granularity is very rich, and they serve as implementation of linguistic models of language with actual data. Here we include FrameNet (Baker et al., 1998), Melchuk’s Explanatory Combinatorial Dictionary (Mel’čuk, 2006), SIMPLE resources (from the Generative Lexicon theory) (Lenci et al., 2000), the Pattern Dictionary of English Verbs (Hanks and Pustejovsky, 2005), FunkGramKB (Periñán-Pascual and Arcas-Túnez, 2004), among other resources. Likewise, and depending on the NLP system under development, e.g. from a spell checker to a machine translation system, different lexica and treebanks are developed with varying content and structure. Thus, we find treebanks annotated with dependency structures for syntactic parsing (McDonald et al., 2013), corpus annotated with WordNet synsets (Ide et al., 2008), or lexica with links to VerbNet (Schuler, 2005), PropBank (Kingsbury and Palmer, 2002), FrameNet (Baker et al., 1998), WordNet (Miller, 1995) and OntoNotes (Aparicio et al., 2008), among many other types. Despite some of them being called *dictionaries* or *lexica*, in this thesis we will not focus on resources that were conceived in the framework of a particular theory of language or whose information is intended to serve as training for NLP systems. We limit ourselves to address resources as outcomes from a lexicographic process that also see humans (who are not necessarily experts on the area of linguistics) as potential future users.

The aspects listed by Atkins and Rundell (2008) indicate that we can expect the content encoded in the entries of a monolingual terminological dictionary to be considerably different from the content of the dictionary entries of a bilingual dictionary of German as a second language, for example. Still, a shared structure serves as backbone, which we describe in the following.

Hartmann (2001), as cited by Trippel (2006), refers to the **macrostructure** of the dictionary as the structure that defines the sequence of the lexicon entries and the access to them. Thus, the macrostructure alludes to the structure of the lexicographic resource as a whole, as opposed to the structure of its individual records, and it concerns the list of headwords, their categories, the distinctions between them and how they will

Figure 1.3: Semasiological and onomasiological access in lexicographic resources³⁰

be accessed (Trippel, 2006), the types of entries it includes, the location and arrangement of them in the lexicon (Atkins and Rundell, 2008; Trippel, 2006), the handling of homonyms, etc. Regarding the access to the content described in the entries, it is here where the distinction between *semasiological* ordering, i.e. accessing the semantics of the entry through its written representation (most dictionaries) and *onomasiological* ordering (accessing the written representation of a word through its semantics, or, rather, through the concept to which it can be mapped), comes into play.

Figure 1.3 exemplifies the two different kinds of access to lexicographical content. If the access is semasiological (case A), the lemma becomes the entry point and its semantics the access goal. When the semantics become the entry point (case B), the lemmas that verbalize that concept are accessed, in this case, *building* and *edifice*. Chapter 7, section 7.4.1, addresses the representation of dictionaries as graphs with LLD models, and will elaborate on the implications of graph structures (where each piece of information becomes a node) in the representation of the macrostructure of dictionaries.

The **microstructure** describes the content of dictionary entries (or articles) (Hart-

³⁰Definition of *building* extracted from WordNet, <http://wordnetweb.princeton.edu/perl/webwn> on 19.06.2019. Photo "BUILDING" by SJ photography, <https://www.flickr.com/photos/59707463@N00/486675282>.

mann, 2001) and their layout (Atkins and Rundell, 2008). Depending on the kind of dictionary (see aspects above), this structure or template can greatly vary, with orthography, phonetic transcription, inflected forms, variant spelling, part-of-speech and sense definitions or translations being commonly present. Etymology, cognates, derivations, examples of use, idioms, nested entries, register, domain, semantic relations, frequent collocations, lexical selection and sub-categorisation, semantic classes, encyclopedic information, among other types of data, may also be part of the description of the dictionary entry.

As an example, Figure 1.4 shows the dictionary entry of the word *credulous* as shown in the Random House Kernerman Webster’s College Dictionary.³¹ In this example, note that a model for the representation of linguistic information as LD will need to account for the written representation, the part-of-speech, syllable breaks, pronunciation, sense ordering, sense definition, usage examples, etymology, and derivations.³²

Figure 1.4: The entry *credulous* from the Random House Kernerman Webster’s College Dictionary

In contrast, the following entry from the Rayfield Georgian-English Comprehensive Dictionary (Rayfield, 2006) of the word *damJere* ‘credulous’ (Figure 1.5), in addition to the lemma, the part-of-speech, and the senses, shows different spellings, dialect information (Pshavian) and translations. Interestingly, in the case of this particular dictionary, the numbers indicate homonyms collapsed under the same entry, not numbered senses

³¹**credulous.** (n.d.) Random House Kernerman Webster’s College Dictionary. (2010). Retrieved October 9 2018 from <https://www.thefreedictionary.com/credulous>.

³²See Trippel (2006) for a thorough overview of different microstructures in a set of representative dictionaries.

of the same polysemous entry. Chapter 6 will revisit the *de facto* standard to represent LLD in terms of the issues that arise when representing the information contained in the microstructure of the dictionary.

დამჯერე, დამჯერ(ებ[ე]ლ)ი (*sst*), დამჯეროზი (*Ps*),
 დამჯეროსი (*Ps*) *pa* **1** trusting, credulous: ბავშვივით
 ~ as trusting as a child; **2** obedient

Figure 1.5: The entry *damJere* from the Rayfield Georgian-English Comprehensive Dictionary

Figures 1.4 and 1.5 show that, in human-oriented dictionaries, we easily access and interpret this information because we recognise the layout conventions that we had in paper dictionaries (such as the Georgian one), many of which have been recreated in their electronic counterparts. These layout conventions already provide some hints on the lexicon structure (Trippel, 2006). It is important to notice at this point that the issues we will be addressing in this thesis in Chapters 4, 5 and 6 pertain to the lexical content and layout decisions grounded on lexical reasons. Typographical information or wildcards to save space have remained outside the scope of this dissertation.

Lastly, a third structural level, the **mediostructure**, accounts for the relation between the individual dictionary entries and the whole dictionary by means of cross-references.

1.2.3 Dictionaries, lexica and terminologies: computational models and formats

Dictionaries and lexicographical projects, in general, come into a large number of formats. Many of them adopt a hierarchical data structure representation, mainly based on XML. A thorough analysis of all formats into which lexicographic resources are encoded is outside the scope of this thesis, but we will introduce the reader to two of the most common formalisms used to represent machine-readable dictionaries (MRD). This will place RDF, as the format of the LD initiative, in the broader context of formats which serve to encode lexical information.

```

<LexicalResource>
  <Lexicon>
    <feat att="language" val="fr"/>
    <LexicalEntry id="rouge-adj-fr">
      <feat att="partOfSpeech" val="adj"/>
      <Lemma>
        <feat att="writtenForm" val="rouge"/>
      </Lemma>
      <Sense id="rouge_rojo-adj-1"/>
    </LexicalEntry>
  </Lexicon>
  <SenseAxis id="rouge_rojo-adj-rojo_rouge-adj" senses="rouge_rojo-adj-1
    rojo_rouge-adj-r"/>
</LexicalResource >

```

Example 1.3: The French-Spanish translation *rouge* – *rojo* ‘red’ from LMF Apertium French-Spanish dictionary.

Lexical Markup Framework. In the development of resources for NLP, the **Lexical Markup Framework (LMF)** (Francopoulo and George, 2006) and its XML DTD were defined to serve as standard for the creation of lexical resources, such as MRDs, to allow for their data exchange. LMF comprises a core package, which serves to describe the basic structure of a lexical entry, and a series of extensions or modules to represent morphological, syntactic, and general multilingual information.

Example 1.3 shows an extract of the XML entry extracted from the LMF version of the Apertium bilingual French-Spanish dictionary. In blue we mark the part of the schema that captures the **microstructure** of the entry. In red, the **macrostructure** elements are highlighted, and the **mesostructure**, which is related to cross-references, is interpreted to be represented with the **SenseAxis** element providing the translation relation between senses. This last information is marked in black color.

LMF encodes lexical and grammatical information in the form of feature structures. While elements such as **LexicalEntry**, **Sense** and **Lemma** belong to the model, attributes like **partOfSpeech** or **writtenForm** belong to a data category registry such as IsoCat, nowadays DatCatInfo (see Section 2.6).³³

³³<http://www.datcatinfo.net/>

The Text Encoding Initiative (TEI) (Sperberg-McQueen and Burnard, 1994).

TEI is a set of guidelines to encode text, and the acronym refers as well to the organisation responsible for them, the TEI consortium. The TEI guidelines are based on XML, and TEI schemas provide a great range of tags to describe a given text, from collections of poetry, to historical accounts and dictionaries. In fact, TEI encompasses a module for the encoding of dictionaries³⁴, focusing on human-oriented ones, but also applicable to computational lexica. The module defines two main types of entries, one encoded with `<entry>` and the second one with `<entryFree>`. The first has restricted content to reflect the shared structure of most dictionaries, and the second allows to represent content more freely. In general, the possibility of defining or changing schemas allows to encode particularities of these dictionaries while staying TEI compliant.

Example 1.4³⁵ shows an entry of the first type, `<entry>`. To represent the microstructure, TEI reflects the hierarchical structure of dictionaries by defining a set of elements and sub-elements that group and embed others. In this example, for the entry *demigod*, this set of elements includes tags such as `entry`, `form`, `gramGrp`, `sense`, etc. from a wide range of tags to encode translations, examples, etymology, usage information, cross-references, etc. In Example 1.4, we are only showing tags that address the **microstructure** of a dictionary entry. To address the **macrostructure** of the dictionary (the overall structure), TEI provides tags such as `<superEntry>`, `<front>` or `<div>` for subdivisions of text, along with attributes as `att.entryLike` and `att.sortable` to specify the type of dictionary entry or to describe elements in lists or groups that are sortable, respectively.

³⁴P5: Guidelines for Electronic Text Encoding and Interchange, Version 3.4.0. Last updated on 23rd July 2018, revision 1fa0b54, last accessed 13.09.2018.

³⁵Extracted from <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html#DITPDE>, last accessed on 19.06.2019.

```

<entry>
  <form>
    <orth>demigod</orth>
    <pron> ... </pron>
  </form>
  <gramGrp>
    <pos>n</pos>
  </gramGrp>
  <sense n="1">
    <sense n="a">
      <def>a being who is part mortal, part god.</def>
    </sense>
    <sense n="b">
      <def>a lesser deity.</def>
    </sense>
  </sense>
  <sense n="2">
    <def>a godlike person.</def>
  </sense>
</entry>

```

Example 1.4: Excerpt of the dictionary entry *demigod* extracted from the TEI Guidelines for P5: Guidelines for Electronic Text Encoding and Interchange

TEI provides a set of XML specifications which can be customised as needed, whereas LMF does not offer this level of customisation. For this reason, efforts to serialise the LMF model (with its view of lexical information as feature structures) by using a TEI-compliant serialisation have also been proposed (e.g. Romary (2013b)).

LMF, together with its extensions, inspired part of the core of the Lexicon Model for Ontologies (*lemon*) (McCrae et al., 2012), which was designed in RDF and has become nowadays a *de facto* standard to convert lexical resources to LLD (see Chapter 2). *lemon* addresses some limitations of LMF, such as its lack of an RDF serialisation and the fact that it does not address the ontology-lexicon interface.

The **OntoLex-*lemon* model** (McCrae et al., 2017) is the effort on which this thesis is framed. Some of the data analysed in this dissertation were converted to LMF in a pre-processing step, while other adhere to custom DTDs similar to the TEI example

shown above. In Chapter 6, a module to encode lexicographic content in LLD as part of the OntoLex Community group is motivated in order to bridge, among other aspects, the gap between TEI and LLD models.

1.3 Structure of the document

This chapter has outlined the main goals of the thesis and the technological context framing it. It has also given an overview of the basic terminology used when describing lexicographical resources, and has provided an insight into their content, structure and common encoding formats.

The thesis comprises nine chapters. Chapter 2 aims to offer a broad state of the art on the current models for the representation of LLD in terms of the content they encode, in order to locate the efforts of those concerned with lexicography in a wider LLD landscape.

Chapter 3 presents the work objectives of this thesis, the followed methodology, the open research problems, the main contributions provided and the hypotheses and assumptions defined during its development.

Chapter 4 dwells on the gaps and challenges identified in the representation of lexicographic resources as LLD, focusing on the analysis of the representation needs of a series of lexicographic resources, including bilingual and multilingual dictionaries, and multilingual terminologies.

Chapter 5 provides solutions to those challenges identified in the previous chapter by suggesting application profiles of OntoLex-*lemon*, URI naming strategies, and extensions.

Chapter 6 motivates and presents the *lexicog* module, a new module for the representation of lexicographic content developed as part of the OntoLex community.

Chapter 7 elaborates on the methodology to evaluate the contributions of the thesis and the results of such an evaluation. Attention is drawn to the conformance of the output RDF resources to the models, to the coverage of the available models in contrast the proposed application profiles and extensions, and to the benefits of LLD for lexicography. This last point includes the benefits for the macro- and microstructure representation, as well as for the combination of complementary information coming from independent lexicographic sources.

In Chapter 8 we propose and analyse the notion of *LD-native dictionaries* as lexicographic resources born in LLD (*not converted* from an original format to LLD) and the problems and challenges to face still in order to make such scenario a reality.

Lastly, conclusions and future work, along with a review of the main contributions and their impact on lexicography and future lines of, are presented in Chapter 9.

1.4 Derived outcomes

The following sections lists the publications derived from the work in this dissertation.

1.4.1 Journal contributions

- Models to Represent Linguistic Linked Data. (2018) **J Bosque-Gil**, J Gracia, E Montiel-Ponsoda, A Gómez-Pérez. *Natural Language Engineering*, 24(6), 811-859. doi:10.1017/S1351324918000347

1.4.2 W3C Technical reports

- W3C OntoLex Lexicography Module Specification (*lexicog*). (2019, September) **J Bosque-Gil**, J Gracia, John McCrae, Philipp Cimiano, Sander Stolk, Fahad Khan, Katrien Depuydt, Jesse de Does, Francesca Frontini, Ilan Kernerman. At the time of writing, available at <http://jograccia.github.io/ontolex-lexicog/> and submitted for publication under the W3C namespace <http://www.w3.org/ns/lemon/lexicog>.

1.4.3 Conference contributions

1. Validating the OntoLex- *lemon* lexicography module with K Dictionaries' multilingual data **J Bosque-Gil**, D Lonke, J Gracia, I Kernerman. Accepted in eLex 2019: Smart Lexicography, October, 1-3, Sintra.
2. Challenges for the Representation of Morphology in Ontology Lexicons. B Klimek, J.P McCrae, **J Bosque-Gil**, M Ionov, J.K Tauber, C Chiarcos. Accepted in eLex 2019: Smart Lexicography, October, 1-3, Sintra.

3. Toward Linked Data-native Dictionaries. (2017) J Gracia, I Kernerman, **J Bosque-Gil**. In *Proceedings of eLex 2017: Lexicography from scratch*, September, 19-21, Leiden.
4. The OntoLex-Lemon Model: Development and Applications. (2017) JP McCrae, **J Bosque-Gil**, J Gracia, P Buitelaar, P Cimiano. In *Proceedings of eLex 2017: Lexicography from scratch*, September, 19-21, Leiden.
5. Terminoteca RDF: a Gathering Point for Multilingual Terminologies in Spain. (2016) **J Bosque-Gil**, E Montiel-Ponsoda, J Gracia, G Aguado-de-Cea. In *Proceedings of TKE 2016: the 12th International conference on Terminology and Knowledge Engineering*, June 22-24, Copenhagen.
6. Zhishi.lemon: On Publishing Zhishi. me as Linguistic Linked Open Data. (2016) Z Fang, H Wang, J Gracia, **J Bosque-Gil**, T Ruan. In *Proceedings of the International Semantic Web Conference (ISWC)*, p. 47–55.
7. Applying the OntoLex Model to a Multilingual Terminological Resource. (2015) **J Bosque-Gil**, J Gracia, G Aguado-de-Cea, E Montiel-Ponsoda. In *Proceedings of the Extended Semantic Web Conference (ESWC) (Satellite Events)*, p. 283–294.

1.4.4 Workshop contributions

1. Modelling Multilingual Lexicographic Resources for the Web of Data: The K Dictionaries Case. (2016) **J Bosque-Gil**, J Gracia, E Montiel-Ponsoda, G Aguado-de-Cea. 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
2. Towards a module for lexicography in OntoLex. (2017) **J Bosque-Gil**, J Gracia, E Montiel-Ponsoda. In *Proceedings of the 1st Workshop on the OntoLex Model (OntoLex-2017)*, Galway.

1.4.5 Demos

- Towards the Integration of Multilingual Terminologies: an Example of a Linked Data Prototype. (2015) E Montiel-Ponsoda, **J Bosque-Gil**, J Gracia, G Aguade-

de-Cea, D Vila-Suero. In *Terminology and Artificial Intelligence (TIA)*, November 4-6, Granada.

1.4.6 Magazine contributions

- Linked Data in Lexicography. (2016) **J Bosque-Gil**, J Gracia, A Gómez-Pérez. Kernerman Dictionary News, 19-24.

1.5 Scholarships

This work has been supported by a grant from the Spanish Ministry of Education, Culture and Sports through the Formación del Profesorado Universitario program (FPU), from September 2015 to September 2019. From 2014 to 2015, this work was supported by a grant from the Universidad Politécnica de Madrid’s Oficina de Transferencia Tecnológica in the context of the LIDER project (FP7 project reference number 610782 in the topic ICT-2013.4.1: Content analytics and language technologies.)

1.6 Research stays

1. 14.05.18–14.08.18. Research stay at the **Applied Computational Linguistics Group (ACoLi) of Goethe University Frankfurt**, supervised by Prof. Dr. Christian Chiarcos and framed in the Linked Open Dictionaries (LiODi) project of the host institution. During this stay we analysed the representation requirements of the Rayfield Georgian-English Comprehensive Dictionary (Rayfield, 2006), proposed an application profile of OntoLex-*lemon* and converted a portion (consisting of nouns) to RDF following the proposed model.
2. 01.06.17–31.07.2017. Research stay at the **Semantic Computing Group, Cognitive Interaction Technology Excellence Center (CITEC), University of Bielefeld, Germany**. In this stay we motivated and initiated the development of the *lexicog* module of OntoLex-*lemon* with the supervision of one of the co-chairs of the W3C Ontology-Lexica community group, Prof. Dr. Philipp Cimi-ano. The stay was funded by the Spanish Ministry of Education, Culture and Sports through the FPU program of grants for research stays.

1.7 Projects

The following projects and networks framed the work presented in this thesis:

1. Legal Knowledge Graph for Multilingual Compliance Services (LYNX), funded by the European Union’s Horizon 2020 research and innovation programme under grant agreement No 780602.
2. Linked Open Dictionaries (LiODi) at the Goethe University Frankfurt, funded by the German Ministry for Education and Research. Rayfield’s Comprehensive Georgian-English dictionary (see Chapters 4 and 5) was data provided in the context of this project during the research stay at Goethe University Frankfurt.
3. ReTeLe - Red de recursos para tecnologías de la lengua (red de excelencia). Ref:TIN2015-68955-REDT.
4. Austria-Israel bilateral EUREKA research and development LDL4HELTA – Linked Data Lexicography for High-End Language Technology Applications. K Dictionaries, Tel Aviv and Semantic Web Company, Vienna. K Dictionaries’s Multilingual Global series Spanish, German and English dictionaries (version of 2016) were data provided in the context of this project.
5. Linked Data as an enabler of cross-media and multilingual content analytics for enterprises across Europe (LIDER), a European FP7 project reference number 610782 in the topic ICT-2013.4.1: Content analytics and language technologies.

Chapter 2

Models for the Representation of Linguistic Linked Data

In this chapter we present the state of the art on models for the representation of language resources (LRs) as LLD. As lexicographic resources vary in content and granularity, we will underline the different mechanisms developed to represent data pertaining to each level of linguistic description. Furthermore, we will delve into the various formalisms proposed to account for lexical description in sub-disciplines of linguistics and related areas. We will lay emphasis on the work performed in lexicography in each of its flavours by considering the works being devoted to convert synchronic and diachronic resources to LLD from different linguistic perspectives. In the process, we will both highlight the gaps encountered in these efforts that make of this task a difficult endeavour, as well as draw attention to the lack of models and guidelines to specifically capture lexicographic content in all its granularity.

In addition, this chapter aims to serve itself as a contribution of this thesis and it is presented in Bosque-Gil et al. (2018). 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 the authors' knowledge there still lacks a comprehensive review of the state of the art on the models or ontologies that support the representation of LLD 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.

2.1 Overview

Linking linguistic information from heterogeneous sources that contain complementary information allows to create a graph of LR^s 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 (Chiarcos et al., 2013, 2012c; Declerck et al., 2012; El Maarouf et al., 2015; Flati and Navigli, 2014; González-Blanco et al., 2016; Hellmann et al., 2013; McCrae et al., 2014; Villegas et al., 2016) and an extensive number of efforts are now devoted towards the conversion of LR^s to RDF. These works apply LLD for language documentation efforts (Nordhoff, 2012), to facilitate dialectal and etymological studies (Abromeit et al., 2016a; Chiarcos and Sukhareva, 2014a), lexical analysis (Khan et al., 2014), morphological studies (Klimek et al., 2016), as well as querying and reuse of lexicographical data (Declerck and Wandl-Vogt, 2015)³⁶, 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 et al. (2012a,b, 2013); Chiarcos and Sukhareva (2015a); Hellmann et al. (2013), just to name a few). These applications are possible thanks to the use of shared conceptualisations and terminologies, such as the ones we revisit throughout this chapter, and common formats.

In addition to the LLOD cloud, an initiative which emerged as the interest in adopting LD among language technology experts grew (see Section 1.1.2), some recent advancements in LLD have also been driven by the activities developed within the framework of international projects such as LIDER³⁷ or FREME³⁸, among others. Workshops, datathons and conferences such as the Multilingual Semantic Web Workshop³⁹, the Linked Data in Linguistics Workshop (LDL)⁴⁰, the Workshop on Knowledge Extraction and Knowledge Integration (KEKI)⁴¹, the Summer Datathon on Linguistic

³⁶Details on its application to a range of dictionaries and terminologies as part of this thesis are presented in Chapter 5 and Bosque-Gil et al. (2016b).

³⁷<http://lider-project.eu/>

³⁸<http://www.freme-project.eu/>

³⁹<http://msw4.insight-centre.org/>

⁴⁰<http://ldl2018.linguistic-lod.org/>

⁴¹<http://keki2016.linguistic-lod.org/>

Linked Open Data⁴², the Conference on Language, Data and Knowledge (LDK)⁴³, the NLP&DBpedia Workshop Series⁴⁴, among other initiatives, have encouraged interdisciplinary contributions and community gathering, and provide a perfect scenario to establish new collaborations along these lines of work. Recently, the H2020 Project Prêt-à-LLOD aims to leverage LLD technologies to create ready-to-use multilingual data, which involves the development of new methodologies for building data value chains across a variety of sectors.⁴⁵ Specially focused on lexicographic data, the ongoing project European Lexicographic Infrastructure (ELEXIS)⁴⁶ aims at introducing common standards and conversion tools, as well as to link existing lexicographic resources to allow for their integration in the form of new data to be consumed by NLP applications.

In order to support the representation of the range of linguistic information contained in LRs, 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, individuals) needed to represent the linguistic information of a LR and to account for its structure and internal/external connections (e.g. lexical entry, syntactic frame, morph, root, antonym, etc.). Such models will be used to represent the linguistic facts contained in LRs. 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.

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 in terms of representation content and granularity. Many of them are complementary to each other and can be used simultaneously to represent the different linguistic features provided in LRs, and lexicographic resources in particular. The classification that we provide aims to serve as a broad framework for thinking about these models, allowing experts to make

⁴²<http://datathon2017.retele.linkeddata.es/>, <http://datathon2019.linguistic-lod.org/>

⁴³<http://ldk2017.org/>, <http://2019.ldk-conf.org/>

⁴⁴<http://nlpdbpedia2015.wordpress.com/>, <http://nlpdbpedia2016.wordpress.com/>

⁴⁵<http://www.pret-a-llod.eu/>

⁴⁶<https://elex.is/objectives/>

informed decisions about the models that best cover their needs. 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.

As part of this thesis, *this chapter highlights the gaps encountered in the literature in the conversion of lexical data to LLD and reflects the lack of standardised models and guidelines to specifically capture lexicographic content*. To this aim, we focus on the observed resource-specific problems and general tendencies detected throughout the literature.

The LLOD cloud offers a classification of LRs 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, lastly, 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.

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. This group, **Group 1**, encompasses ontology lexica models, such as *lemon* (McCrae et al., 2012) and *OntoLex-lemon*⁴⁷, and initiatives such as TELIX (Rubiera et al., 2012) or the Oxford Global Languages Ontology (OGL) (Parvizi et al., 2016).
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**. This group encompasses two subgroups:

⁴⁷<http://www.w3.org/2016/05/ontoLex/>

- (a) **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, 2012; Moran and Wright, 2009), the MMoOn ontology for morphology (Klimek, 2017), and the 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 (Pustejovsky, 1991).
 - (b) **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* (Abromeit et al., 2016a; Chiarcos and Sukhareva, 2014a) for etymological data, along with models used in contrastive linguistics and lexicography.
3. **Catalogues** of linguistic data categories. This group contains catalogues of linguistic data categories that cover the whole linguistic domain and are widely used in conjunction with models of Group 1, Group 2a and 2b (concrete examples are provided in Section 2.6). In this group we revisit ontologies of linguistic description, category and language identifier registries (GOLD (Farrar and Langendoen, 2003), LexInfo (Cimiano et al., 2011a), ISocat⁴⁸ (Marc Kemps-Snijders and Wright, 2008; Windhouwer and Wright, 2012), Lexvo (de Melo, 2015) and the OLiA Ontologies of Linguistic Annotation⁴⁹(Chiarcos and Sukhareva, 2015a)), among other resources.
 4. Models geared towards the representation of **corpora and service-oriented** models. These last models are analysed in detail here, since these contributions include many annotation-related (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.

⁴⁸Previously at <http://www.isocat.org> and nowadays the Data Category Repository (DCR) at <http://www.datcatinfo.net/#/>

⁴⁹<http://acoli.cs.uni-frankfurt.de/resources/olia/>

This chapter is organised as follows. Sections 2.3 and 2.7 focus on each group in detail. Section 2.3 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 LAF (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.⁵⁰ Lastly, we provide a discussion in Section 2.8.

2.2 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 2.1. This arrangement of the linguistic description levels is widely found in reference books in linguistics, 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 chapter 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, commonly addressed in theoretical linguistics, and some of the sub-fields in applied linguistics and areas related to it.

Group 1 models mainly represent information pertaining to the 2nd-4th levels in Figure 2.1 (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 covered by them. We analyse and compare the contributions in each of these groups in the following sections.

⁵⁰Likewise, other models that do not provide linguistic descriptions as LD are outside of the scope of this thesis. Further, techniques, tools, and best practices for building ontologies, as the ones used in LLD, are not examined in this chapter. On this topic, we refer the interested reader to the literature in ontology engineering (e.g. reviews of methodologies for building ontologies by Corcho et al. (2003); Iqbal et al. (2013)).

Figure 2.1: Levels of linguistic description and their units of analysis

2.3 Core models for lexical resources

In this section we first revisit work on the representation of lexical content as graph outside the context of LLD, as well as standardisation initiatives for the creation and annotation of NLP resources. Following, we examine those LD-based models that address the representation of core elements of lexical resources.

2.3.1 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, 1998b, 2005), and Polguère’s lexical systems (Polguère, 2006; Polguère, 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 and George, 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., 2011a) 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 2.3.2. In the context of the ISO’s Technical Committee TC37 SC4⁵¹, we briefly refer also to the **Linguistic Annotation Framework (LAF)** (Ide and Romary, 2004) and its XML serialization, the **Graph Annotation Format (GrAF)** (Ide and Suderman, 2007), the **Morpho-syntactic Annotation Framework (MAF)** (de la Clergerie and Clément, 2005) and the **Syntactic Annotation Framework (SynAF)** (Declerck, 2006).

In addition to LMF, other formats suggested to address the syntactic heterogeneity in linguistic resources for NLP are **TEI** (Sperberg-McQueen and Burnard, 1994) (see Section 1.2.3) and **TBX** (ISO DIS 30042, 2007) for terminologies.

2.3.2 LD-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.

⁵¹<http://www.iso.org/committee/297592.html>

Simple Knowledge Organization System. The Simple Knowledge Organization System or SKOS (Miles and Bechhofer, 2009) is a W3C Recommendation⁵² that is widely used to represent thesauri and taxonomies (the Food and Agriculture Thesaurus AGROVOC⁵³, the GEMET Thesaurus of the European Environment Agency⁵⁴, the Multilingual Thesaurus of the EU (EUROVOC)⁵⁵, 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).⁵⁶ This extension allows to reify lexical labels of resources and provides properties to relate them (e.g., an acronym relation can be a sub-property 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.⁵⁷

lemon and OntoLex-lemon *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-*lemon*, 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 modularize it. Both *lemon* and OntoLex-*lemon* belong to the group of ontology lexica models, i.e, models for the lexicalization of an ontology. The criteria guiding the development of *lemon* and OntoLex-*lemon* 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.,

⁵²<https://www.w3.org/TR/skos-reference/>

⁵³<http://aims.fao.org/standards/agrovoc/linked-open-data>

⁵⁴<http://www.eionet.europa.eu/gemet>

⁵⁵<http://eurovoc.europa.eu/drupal/>

⁵⁶<http://www.w3.org/TR/skos-reference/skos-xl.html>

⁵⁷See <http://www.w3.org/2016/05/ontolex/#skos-xl>

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 lexicalize 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 and Vetere, 2008), which follows the DOLCE 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-lemon* 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 2.4 presents the extensions to *lemon* and *OntoLex-lemon* that have been proposed in the literature.

The Oxford Global Languages Ontology (OGL). With the conversion of lexicographic resources with fine-grained lexical description to RDF, new modelling problems arise, and the lack of mechanisms to represent certain lexicographic annotations moves authors to design extensions to *OntoLex-lemon* (Section 2.4) or, recently, to the creation of a new ontology, partially from scratch, targeted at the conversion of lexicographic resources. The *Oxford Global Languages Ontology (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, 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. `LexicalEntry` in *lemon* vs. `LexicalEntry` 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. Gracia et al. (2018), see Chapter 5) or the exploitation of translation relations among senses to obtain indirect translations through a pivot language (Gracia et al., 2018; Villegas et al., 2016).

TELIX. Lastly, 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-lemon*, and the OGL ontology. Just as corpus oriented models (Section 2.7), 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 2.6). 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-lemon* 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 2.2, 2.3 and 2.4 show the different representations of the lexical entry *book* according to *OntoLex-lemon*, the OGL ontology, and TELIX, respectively.⁵⁸ Example 2.1 provides the RDF Turtle serialization of the entry shown in Figure 2.2 and also illustrates the *subject-predicate-object* model that was introduced

⁵⁸Throughout 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.

Figure 2.2: The lexical entry *book* modelled with OntoLex

earlier in Section 1.1.1. The *book* example is inspired by examples provided in the literature (Parvizi et al., 2016; Rubiera et al., 2012) and the OntoLex Specification, which in turn have been adapted for comparative purposes. As shown in Figure 2.3, in the 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-*lemon* (instantiated in Figure 2.2), in contrast, forms are not shared by lexical entries, and inflectional word forms are viewed as forms of the lexical entry. In OntoLex-*lemon* there is no **entry** element to track the provenance, but it relies on properties available in other vocabularies, such as `dc:source`, to represent this information. Due to its lack of elements such as **Entry**, the OntoLex-*lemon* representation does not keep track of the original representation in the lexicographic resource.

Lastly, since TELIX is oriented towards the representation of annotations, Figure 2.4 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.


```

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

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

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

```

Example 2.1: Turtle RDF serialization of the example in Figure 2.2

Figure 2.3: The lexical entry *book* modelled with the OGL ontology

Figure 2.4: The lexical entry *book* modelled with TELIX

Table 2.1 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.

2.4 Level-specific and theory based models

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

2.4.1 Models for different levels of linguistic analysis

In the following, we revisit the ontologies and extensions to the models presented in Section 2.3.2 developed to address linguistic content from each of these levels and their interfaces by following the order presented in Figure 2.1.

Model	Related area	Applicable to
MLR model (Spohr, 2012)	Lexicography	Dictionaries and lexica
SKOS (Miles and Bechhofer, 2009)	Terminology	Terminologies, thesauri
<i>lemon</i> (McCrae et al., 2012)	Ontology engineering, NLP, lexicography	Ontologies and lexica, corpora (with a corpus model)
OntoLex- <i>lemon</i> ^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

Table 2.1: Group 1: Models to represent the main elements of lexical resources

^a<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, 2012; Moran and Wright, 2009) 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 **General Ontology for Linguistic Description** (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 data sets yet.

Morphology. In addition to the classes, properties and individuals available in linguistic data category registries (Section 2.6), the **Multilingual Morpheme Ontology**⁵⁹ (MMoOn) (Klimek, 2017; Klimek et al., 2016) was recently made available with

⁵⁹<http://mmoon.org/>

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, e.g. the Bantu Language Model (Eckart et al., 2019). At the time of writing, the OntoLex community is working on a new module for the representation of morphological data, based on MMoOn, which accounts for information ranging from endings to derive whole paradigms, to the decomposition and generation of lexical entries in compliance to other OntoLex-*lemon* modules.⁶⁰ 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:Form` elements 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 of the lexical entry and the set of transformations that would systematically generate a series of inflected forms of that lexical entry. This module is available as the **Lemon Inflectional Agglutinative Morphology (LIAM)** ontology.⁶¹

Syntax. The syntax domain is mainly addressed by the annotations schemes present in linguistic data category registries, which are covered in detail in Section 2.6. 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 (Chiarcos et al., 2012a; Skut et al., 1998).

Syntax-Semantics Interface. The representation of the interface between syntax and semantics is receiving much attention by the community. The **Predicate Model for Ontologies** (PreMOn) (Corcoglioniti et al., 2016) and its extension to model NomBank⁶², PropBank⁶³, VerbNet⁶⁴, and FrameNet⁶⁵ are put forward as complements to

⁶⁰<http://www.w3.org/community/ontolex/wiki/Morphology>

⁶¹<http://lemon-model.net/liam>

⁶²<http://nlp.cs.nyu.edu/meyers/NomBank.html>

⁶³<http://verbs.colorado.edu/~mpalmer/projects/ace.html>

⁶⁴<http://verbs.colorado.edu/~mpalmer/projects/verbnet.html>

⁶⁵<http://framenet.icsi.berkeley.edu/>

lemon to represent predicate models and their mappings (SemLink⁶⁶) as LD. The model, just as MMoOn for morphology, responds in an OntoLex-*lemon*-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-*lemon* and SKOS. An extract of the entry *hesitate* modeled with PreMOn is illustrated in Figure 2.5. 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 conceptualization mapping (which in the actual data takes two conceptualizations, `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 (WordNet⁶⁷, VerbNet, OmegaWiki⁶⁸ in German and English and Wiktionary⁶⁹ in English and German) had previously been converted to *lemon* in *lemonUby* (Eckle-Kohler et al., 2015). The vocabulary *ubyCat* was proposed to extend *lemon* with data structures found in UBY resources (in LMF) and to link the data 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.

⁶⁶<http://verbs.colorado.edu/semLink/>

⁶⁷<http://wordnet.princeton.edu/>

⁶⁸<http://www.omegawiki.org/>

⁶⁹<http://www.wiktionary.org/>

Norms and Exploitations (TNE) (Hanks, 2013; Hanks and Pustejovsky, 2005). 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 et al., 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 TNE) give rise to modelling problems.

The lack of direct connection from a syntactic frame to a lexical sense in *lemon* and *OntoLex-lemon* (link only through the class `OntoMap`) is also encountered in recent work in the conversion of other verbal inventories, such as “Les Verbes Français” (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 2.6 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 created 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 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 realizations. 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

Figure 2.6: The lexical entry *hesitate* modelled with PDEV-lemon

primitives information). Thus, 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 TNE 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 Generative Lexicon Theory (GL) (Pustejovsky, 1991). As a result of the PAROLE (Ruimy et al.) 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 2.6) (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 templates 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 no 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 (Gratta et al., 2013). An example of this dataset for the entry *esitazione* ‘hesitation’ is provided in Figure 2.7. 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.

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 Generative Lexicon qualia roles (formal, constitutive, agentive, 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’čuk, 1997; Mel’čuk, 1998) have also been proposed. The recent Lexical Functions Ontology Model (*lexfom*) (Fonseca et al., 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 vs. syntagmatic), semantic classes of lexical functions; features, constituents and government patterns of lexical functions, and means to link *lemon*

Figure 2.7: The lexical entry *hesitate* modelled in Parole-Simple Clips

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-lemon*, WordNet 2.0 had been previously converted to RDF, relying in a custom data model.⁷³ In the conversion 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 2.8 with a sense of the entry *hesitate* linked to a synset as reference. Through

⁷³<http://www.w3.org/TR/wordnet-rdf/>

Figure 2.8: The lexical entry *hesitate* modelled in WordNet RDF (2014)

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-*lemon*.⁷⁴ 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 (CILI) (Bond et al., 2016).

The Global WordNet Formats to publish WordNets⁷⁵ reuse the entities from the WordNet ontology as RDF and JSON-LD, along with the OntoLex-*lemon* vocabulary. Another ontology developed in order to convert a lexico-semantic resource is in fact the EuroWordNet Multilingual Central Repository (MCR) Ontology⁷⁶, generated from the LMF versions of WordNet Lexica included in the Multilingual Central Repository (MCR)⁷⁷ (Gonzalez-Agirre et al., 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.

With respect to OntoLex-*lemon* specifically and related to the realm of description

⁷⁴See <http://wordnet-rdf.princeton.edu/about>.

⁷⁵<http://globalwordnet.github.io/schemas/>

⁷⁶<http://github.com/martavillegas/EuroWordNetLemon>

⁷⁷<http://adimen.si.ehu.es/web/MCR>

logics (DL), a technique has recently been developed to extend an `ontolex:LexicalConcept` with DL definitions in a way that the natural language definition is transformed to the DL-definition and related to the domain ontology properties in ontology lexica (Gritz, 2018).

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 (LMM) (Davide Picca 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 DOLCE (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 Ontology⁷⁸, imported by *OntoLex-lemon*, and includes also pragmatics-inspired elements such as the class **LinguisticAct**.

Discourse and information structure. The Ontologies of Linguistic Annotation (OLiA), explained in more detail in Section 2.6, 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 NERD⁷⁹ or the Grounded Annotation Framework (GAF)⁸⁰). 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

⁷⁸<http://www.ontologydesignpatterns.org/cp/owl/semiotics.owl#>

⁷⁹<http://nerd.eurecom.fr/ontology>

⁸⁰<http://groundedannotationframework.org/>

on the annotation schemes of the Rhetorical Structure Tree Bank (RSTB)⁸¹, grounded on the Rhetorical Structure Theory (Mann and Thompson, 1988), OntoNotes (Pradhan et al., 2007), and the Penn Discourse Tree Bank⁸² 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 RSTB 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 2.6).

Table 2.2 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 type of resources to which they are applied. The structure of this table differs from Table 2.1 to reflect the fact that the systems described in this section deal with specific levels of linguistic analysis.

⁸¹<http://www.isi.edu/~marcu/discourse/Corpora.html>

⁸²<http://www.seas.upenn.edu/~pdtb/>

Model	Level	Linguistic Theory/Approach	Applicable to
PHOIBLE Model (Moran, 2012; Moran and Wright, 2009)	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 (Nuzolese 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

Table 2.2: Group 2a: Models for different levels of linguistic analysis

Model	Level	Linguistic Theo- ry/Approach	Applicable to
SIMPLE OWL Ontology (Toral and Monacchini, 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
<i>lexfom</i> (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
Linguistic Meta-Model (LMM) (Davide Picca and Gangemi, 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

Group 2a: Models for different levels of linguistic analysis (continuation)

2.5 Area-specific models

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 applied linguistics 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, lexicography, etc.), 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, cross-linguistic studies, and, specially, lexicography.

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 et al., 2016c). 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 OWL⁸³ for the conversion of historico-philological data (Khan et al., 2016c). An interface to create datasets based on *lemonDIA* is also provided (Khan et al., 2016a).

In relation to historical linguistics and moving into lexicography, the European Network of e-Lexicography (ENeL) fostered work towards the conversion of a range of dictionaries, some of which contain historical data, to LLD based on the *OntoLex-lemon* model. To this aim, new classes and properties are defined to capture informations which the *lemon* and *OntoLex-lemon* 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 ENeL (Declerck and Wandl-Vogt, 2015). An extension to *lemon* to represent etymological information of lexical entries has also been proposed (*lemonet*) (Chiarcos and Sukhareva, 2014a). In the past years, a new revisited version of *lemonet* builds upon the properties proposed in the modelling of the etymological WordNet⁸⁴ in order to undertake the conversion of the Tower of Babel (Starling)⁸⁵ with properties such as `lemonet:cognate` and `lemonet:derivedFrom` as subproperties of `vartrans:lexRel` from the *vartrans* module of *OntoLex-lemon* (Abromeit et al., 2016a). This last work has been launched in the context of the Linked Open Dictionaries (LiODi) project⁸⁶, 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. Recently, a

⁸³<http://www.w3.org/TR/owl-time/>

⁸⁴icsi.berkeley.edu/~demelo/etymwn/

⁸⁵<http://starling.rinet.ru/>

⁸⁶<http://acoli.cs.uni-frankfurt.de/liodi/home.html>

new extension for *OntoLex-lemon*, *lemonEty*, has been proposed to capture hypotheses about the history of lexical entries, the lexemes that play a role in that etymology and their ordered arrangement (Khan, 2018).

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., 2016d) and *polyLemon* (Khan et al., 2016c) 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 Lexicon⁸⁷ 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-lemon* have also been proposed, for instance, for the conversion of K Dictionaries (KD) German dictionary of the KD’s Global series (Klimek and Brümmer, 2015).⁸⁸ 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. Recent works on the representation of semantic shifts in inflectional morphology have also highlighted the gaps in *OntoLex-lemon* that concern the relation between an `ontolex:LexicalSense` and an `ontolex:Form` in those cases in which a sense only allows for a certain form (e.g. in plural) (Gromann and Declerck, 2019), which is information commonly given in dictionaries as annotations of the sense (see Chapter 6 for details on the *lexicog* module, aimed at bridging these gaps). The conversion from TEI to RDF of dialectal dictionaries of Arabic for their subsequent integration using *OntoLex-lemon* has been addressed as well (Declerck and Mörtz, 2016). The focus here has been set 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.

However, lexicographic resources do not only come in the form of dictionaries or lexica. The representation of thesauri as LLD has also been addressed. The new model *lemon-tree* (Stolk, 2017, 2019) overcomes the shortcomings in *OntoLex-lemon* for the

⁸⁷<http://www.tlg.uci.edu/lsg/>

⁸⁸See Chapter 5 for the proposed modelling of the Spanish, German and English dictionary of this same series, and Chapter 6 for the application of the *lexicog* module to this same data.

representation of the different levels in thesauri topical systems and provides the property `isSenseInConcept`, distinct from `ontolex:isLexicalizedSenseOf`, to capture the difference between lexicalization and categorisation in topical thesauri.

With regards to the representation of the multilingual aspect in lexicographic resources, and lexical resources in general, a translation module for *lemon* was proposed (Gracia et al., 2014a) – the *lemon translation module* (see Section 5.2.1). This module, reused in resources such as the Apertium RDF series (Gracia et al., 2018), inspired later the *vartrans* module of *OntoLex-lemon*. 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, Wiktionary⁸⁹ was converted to *lemon* in the DBnary resource (Sérasset, 2015; Tchechmedjiev et al., 2014) 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 a Wiktionary page, etc. Figure 2.9 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 et al., 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 (`plx:Language`, `plx:LanguageVariety`).

⁸⁹<http://www.wiktionary.org/>

Figure 2.9: The lexical entry *hesitate* modelled in DBnary

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 2.3), prove that the mapping of proprietary formats to a standard framework like *lemon* or *OntoLex-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-lemon*, in fact, though devised from a lexicographic perspective, were originally intended to lexicalize 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 and Mörrth (2016); Declerck and Wandl-Vogt (2015); Klimek and Brümmer (2015); Parvizi et al. (2016), etc.). Details on this are presented in Chapters 6, 7 and 8.

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 2.6) and PHOIBLE (introduced in Section 2.4.1), we point out the Typological Database System (TDS) 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 (CCLD) (Forkel, 2014) promotes the creation of a LD infrastructure to integrate and publish typological data sets (PHOIBLE, WALS⁹⁰, WOLD⁹¹, Glottolog⁹², afbo⁹³, 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 2.3, and its extension SKOS-XL, best practices have been defined for the generation of LLD from terminological resources in the Term Base eXchange (TBX) format.⁹⁴ This has been the case of the conversion of IATE⁹⁵ and the European Migration Network (EMN)⁹⁶ datasets to RDF using OntoLex-*lemon* and a specific vocabulary put forward for that purpose (Lider TBX Ontology) (Cimiano et al., 2015a). This vocabulary encodes header information, reliability codes, or information related to transactions, for example.

Table 2.3 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 applied.

⁹⁰<http://wals.info/>

⁹¹<http://wold.clld.org/>

⁹²<http://glottolog.org/>

⁹³<http://datahub.io/km/dataset/clld-afbo>

⁹⁴<http://www.tbxinfo.net/>

⁹⁵<http://iate.europa.eu/>

⁹⁶<http://www.emn.ie/>

2.6 Linguistic category registries

This section 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 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 General Ontology for Linguistic Description (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 (SUMO)⁹⁷ 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, tense) and phonetic layers (Figure 2.10). In contrast to ontologies such as DOLCE (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 was in turn based on ISO 12620:2009⁹⁸, which establishes the specifications of data categories (the data model) and the management of such a registry (Marc Kemps-Snijders and Wright, 2008). Importantly, it was not a formal ontology but a structured collection of terms (Eckle-Kohler et al., 2015), and ontological relationships were not 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

⁹⁷<http://www.adampease.org/OP/>

⁹⁸<http://www.iso.org/standard/37243.html>

GOLD

GOLD 2010																													
issues	versions	xml	owl/rdf	gold community	help																								
Properties	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z			
Properties										Grammar Unit										Perfective Aspect									
abbreviation										Grapheme										Perfect Tense									
A										Greater Paucal Number										Perlative Case									
Abessive Case										Greater Plural Number										Permissive Modality									
Abilitative Modality										Group										Personal Passive Voice									
Ablative Case										H										Personal Pronoun									
Absolute Antipassive Voice										Habitual Aspect										Person Property									
Absolute Case										has Citation										Phasal Aspect									
Abstract										has Complex Value										Phoneme									
Accusative Case										has Constituent										phonemic Rep									
Acoustic Property										has Constraint										Phonetic Property									
acoustic Realization										has Entry										phonetic Rep									
Actional Force										has Example										Phonological Property									
Active Voice										has Feature										Phonological System									
Adessive Case										has Form										Phrase									
Adjectival										has Glosses										Physical Abilitive Modality									
Adjective Phrase										has Grammar										Place Property									
adjunct										has Lexical Item										Plain Adjective									
Adjunct Subordinate										has Lexical Unit										Plain Middle Voice									
Adposition										has Meaning										Plural Number									
Adverbial										has Morphological Constituent										Plus A T R									

Figure 2.10: A small section of the GOLD ontology in its web interface

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 for 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-Köhler et al., 2015; Schuurman et al., 2016).

Another linguistic data category registry is LexInfo (Cimiano et al., 2011a), specifically developed to address the lexicon-ontology interface and widely used together with *lemon* and *OntoLex-lemon* (see Ehrmann et al. (2014); El Maarouf et al. (2014); Gracia et al. (2018); Sérasset (2015), 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, val-

Figure 2.11: A small section of LexInfo in the ontology editor *Protégé*

ues and properties derived partially from ISOcat⁹⁹, and currently its elements capture information from the morphosyntactic (see Figure 2.11), syntactic, syntactic-semantic, semantic and pragmatic levels of linguistic description.

With an architecture different from LexInfo and ISOcat, the Ontologies of Linguistic Annotation (OLiA) (Chiarcos and Sukhareva, 2015a) 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

⁹⁹<http://www.lexinfo.net/ontology/2.0/lexinfo.owl>

Figure 2.12: A small section of OLiA's Reference Model in the ontology editor *Protégé*

this architecture is to abstract from the specific annotation schemes and to provide a common ontological ground across the different models to allow for the generalization, 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 et al., 2016b) (Figure 2.12).

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, 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 *language*, 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. Some approaches have addressed the shortcomings of the authoritative list of language codes by ISO 639 and resources such as Glottolog, specially in relation to the encoding of lesser lesser-known languages spoken by minorities and also for historical stages of languages (Gillis-Webber et al., 2019a). A model for representing information about languages, MoLA has been recently been proposed to specify language tags, lects, language and dialect families, their region or their time period (Gillis-Webber et al., 2019b).

Table 2.4 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.

2.7 Corpora- and service oriented 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 structural aspects of corpora or characteristics of NLP services. Although the analysis of these models falls outside the scope of this chapter, it is worth mentioning them in order to enrich the context of this chapter. 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-lemon*) and linguistic data category catalogues (e.g., *LexInfo*, *OLiA*). Examples of efforts here are the NLP Interchange Format (NIF)¹⁰⁰ (Hellmann et al., 2013), which can be used with the Open Annotation Data Model¹⁰¹, now superseded by the Web Annotation Data Model¹⁰², to represent corpora and annotations 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 et al., 2015).

Additional corpora models are the annotation layer of TELIX (Rubiera et al., 2012), the Fiesta/Mexico data model (Menke, 2016; Menke et al., 2013) 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

¹⁰⁰See ontologies and specifications at <http://persistence.uni-leipzig.org/nlp2rdf/>

¹⁰¹<http://www.openannotation.org/spec/core/>

¹⁰²<http://www.w3.org/TR/annotation-model/>

structure in general. As examples of corpora that reuse these models we mention The Universal Dependencies Treebank series in RDF, which use the NIF 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. Recently, the Ligt vocabulary (Chiarcos and Ionov, 2019) has been proposed to represent linguistic examples as text with interlinear glosses (IGT) as LD. With its base classes derived from NIF, it provides elements to encode the relations between documents, morphemes, and their annotations in IGT, e.g. `ligt:Tier` to capture a set of annotations that share the same segmentation.

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 Fernandez, 2013) is aligned with the WordNet-Affect taxonomy¹⁰³ and the Emotion Markup Language (EmotionML)¹⁰⁴ in order to encode emotions and their relations to the lexical items lexicalizing 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 2.5 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.

¹⁰³<http://www.gsi.dit.upm.es/ontologies/wnaffect/>

¹⁰⁴<http://www.w3.org/TR/emotionml/>

Model	Domain	Applicable to
NIF (Hellmann et al., 2013)	NLP, corpus linguistics	Corpora
POWLA (Chiarcos, 2012)	NLP, corpus linguistics	Corpora, treebanks
Fiesta/Mexico (Menke, 2016; Menke et al., 2013)	NLP, corpus linguistics	Multimodal corpora
Semantic Quran (Sherif and Ngonga Ngomo, 2015)	Historical linguistics, digital humanities	Corpora, The Quran
Marl (Buitelaar et al., 2013; Westerski and Sánchez-Rada, 2013)	NLP (Sentiment analysis, opinion mining)	Lexica, lexical databases
Onyx (Sánchez Rada and Iglesias Fernandez, 2013)	NLP (Sentiment analysis, opinion mining)	Lexica, lexical databases
MLSA (Declerck and Krieger, 2014)	NLP (Sentiment analysis, opinion mining)	Lexica, lexical databases
Ligt (Chiarcos and Ionov, 2019)	NLP, corpus linguistics, cross-lingual studies	Corpora with interlinear glosses

Table 2.5: Other Models

2.8 Limitations and open challenges in LLD representation

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., *OntoLex-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.

Given the landscape of models presented in this chapter, we identify a series of limitations in the state of the art for LLD representation and a list of open challenges

to address.

2.8.1 Limitations

1. **A 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 (Generative Lexicon-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-lemon vartrans* module).

2. Despite the multiple works addressing the conversion of **lexicographical resources to LLD** coming from historical linguistics, cross-lingual studies, and general lexicography, the mapping of lexicographical content to a standard framework as *OntoLex-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. There exist **no broad guidelines to convert a wide range of lexicographic resources as outcomes of a lexicographic compilation process to LLD**. This has called for the proposal of models and small extensions, contributing to the scenario mentioned in (1). Consequently, there exists **no wide-scope exploration and assessment of the impact of the application of LLD to lexicography as a discipline**: the requirements and process this involves, its practical and theoretical benefits, the challenges it raises, and the open problems on the way.
3. 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 find 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.
4. 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**, besides Universal Dependency treebanks, 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.

5. **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-lemon 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. However, several challenges arise when trying to **automatically establish links among resources in different languages** (Gracia et al., 2012a,b), 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* vs. Fr. *rivière* or *flueve*; *president* or *prime minister* vs. Sp. *presidente*, etc.), or even when **addressing specific features of a multilingual resource** (for instance, multilingual examples in a dictionary).
6. Despite past and on-going projects¹⁰⁵, the literature shows that **there is still scarce assessment of the impact of the use of LLD-based language resources in empirical linguistics**.
7. **The application of the LD paradigm to areas related to linguistics unmentioned in this survey 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

¹⁰⁵Among them, Linked Data Lexicography for High-End Language Technology (LDL4HELTA, <http://ldl4.com/>), LIDER, the European Lexicographic Infrastructure (ELEXIS, <http://elex.is/>), Linked Open Dictionaries (LiODi, <http://www.acoli.informatik.uni-frankfurt.de/liodi/home.html>) and explore.AI - Exploring Austria's culture through the language glass (<http://www.oeaw.ac.at/acdh/projects/exploreat/>).

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.

2.8.2 Open challenges

Even though the scenario offered above shows 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:

1. 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 (Chiarcos et al., 2016a; McCrae and Cimiano, 2015) and the recent project Prêt-à-LLOD (see Section 2.1) aims at paving the way towards its solution.
2. Secondly, 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 (Anjomshoaa et al., 2014; Jain et al., 2010; Millard et al., 2010), 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.

3. 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 to conduct 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 chapter reveal, gradually more and more granularity is being added onto the different levels through extensions, new ontologies and new use cases.

The above mentioned lists of limitations and challenges can be summarised into three key aspects:

1. In relation to modelling efforts, data pertaining to **some levels of linguistics** analysis, and specially lexico-semantic data, causes **modelling challenges still**, particularly when a resource to convert to RDF is grounded on a specific theory of language. As regards the areas related to linguistics, the number of efforts devoted to model **lexicographic resources** (including synchronic and diachronic resources) are increasing in the recent years, with experts often proposing custom extensions to meet the representation needs of a resource. This often leads to a **proliferation of independently developed vocabularies**.
2. With respect to the **application and adoption** of the LD paradigm, its application to some areas related to linguistics (and sub-disciplines of linguistics) is still **largely unexplored**. Likewise, the implications of LD for linguistic research (beyond NLP purposes), involving the **use of LLD resources by linguists**, has not been widely addressed yet. However, on-going projects are paving the way for this to become a reality.
3. The current **lack of a solid ecosystem of tools and services** that consume LLD (or that give access to it through user interfaces) partly outshines the benefits

and the potential use of these technologies by language experts unfamiliar with the Semantic Web.

In Chapter 3, we will address the first two aspects by dividing them in a series of sub-limitations specifically related to lexicography.

Model	Related area	Applicable to
lemonDIA (Khan et al., 2014)	Diachronic linguistics	Lexica, lexico-semantic databases
Dictionary Extensions by (Declerck and Wandl-Vogt, 2015)	Lexicography, historical linguistics, cross-lingual studies, dialectology	Lexica
lemonet (Chiarcos and Sukhareva, 2014a)	Lexicography, historical linguistics, etymology	Etymological lexica, dictionaries and databases
lemonet (Abromeit et al., 2016a)	Lexicography, historical linguistics, etymology	Etymological Lexica, dictionaries and databases
lemonLSJ and polyLemon (Khan et al., 2016c,d)	Lexicography	Dictionaries
lemonEty (Khan, 2018)	Lexicography, historical linguistics, etymology	Lexica, lexico-semantic databases
K Dictionaries Vocabulary (Klimek and Brümmer, 2015)	Lexicography	Global Series K Dictionaries (monolingual)
Extensions by (Gromann and Declerck, 2019)	Lexicography, lexical semantics	Lexica, lexico-semantic databases

Table 2.3: Group 2b: Models for specific branches of applied linguistics and related areas

Model	Related area	Applicable to
<i>lemon</i> -tree (Stolk, 2019)	Lexicography	Topical thesauri
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., 2015a)	Terminology	Terminologies in TBX

Group 2b: Models for specific branches of applied linguistics and related areas (continuation)

Model	Level	Domain
ISOcat (Marc Kemps-Snijders and Wright, 2008)	Morphosyntax, syntax, (lexical) semantics	Terminology, lexicography, translation, sign language linguistics
CLARIN CR (Schuurman et al., 2016)	Morphosyntax, syntax, (lexical) semantics	Terminology, lexicography, translation, sign language linguistics
LexInfo (Cimiano et al., 2011a)	Morphosyntax, syntax, (lexical) semantics, pragmatics	Terminology, lexicography, translation, NLP
OLiA (Chiarcos and Sukhareva, 2015a)	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
MoLA (Gillis-Webber et al., 2019b)	N/A	Descriptive linguistics, cross-lingual studies

Table 2.4: Group 3 Models

Chapter 3

Goals and Contributions

3.1 Goals

The ultimate goal of this thesis is to research the application of the Linguistic Linked Data (LLD) paradigm to Lexicography. More specifically, we aim to achieve the following goals:

- G1. To advance in the state of the art on models for the representation of LLD in order to support the representation of 1) explicit lexical information encoded in the annotations of lexicographic records and 2) implicit lexical information reflected on structural aspects.
- G2. To facilitate the transformation of lexicographic resources to LLD by providing new application profiles for the *de facto* standard (OntoLex-*lemon*), extensions, methodologies, design decisions and a migrated set of representative examples in RDF.
- G3. To assess the impact of LLD in the lexicographic workflow and in the use of lexicographic resources for data gathering and consultation by linguists and lexicographers.

In Chapter 2, we summarised the detected limitations of the State of the Art in three general categories: (1) difficulty in the modelling of lexico-semantic data, and data coming from a range of lexicographic resources, leading the proliferation of independent vocabularies, (2) scarce exploration of the adoption of the LLD paradigm in linguistics

and related areas, and (3) lack of a solid ecosystem of tools that consume LLD and showcase its potential.

For the three above mentioned objectives to be attained towards the end goal of this thesis, we address limitations (1) and (2) and subdivide them in the following list:

- L1. There exists no previous wide-scope exploration and assessment of the impact of the application of LLD to lexicography as a discipline: the requirements and process this involves, its practical and theoretical benefits, the challenges it raises, and the open problems on the way.
- L2. There exist no broad guidelines to convert a wide range of lexicographic resources as outcomes of a lexicographic compilation process to LLD.
- L3. Despite past and on-going projects¹⁰⁶, the literature shows that there is still scarce assessment of the impact of the use of LLD-based lexicographic resources by linguists and lexicographers for consultation and data gathering.

To overcome these limitations in the literature, the goals mentioned above will be aligned along four main research lines:

- RL1 Identification of problems and challenges in the application of LLD to lexicographic resources.
- RL2 Proposal of modelling solutions and motivation of community initiatives to solve such problems.
- RL3 Conversion of lexicographic resources into LLD.
- RL4 Theoretical and query-based evaluation of the implications of LLD-based resources for lexicography, and for consultation and data gathering by linguists and lexicographers.

Figure 3.1 summarises the mapping between state of the art limitations, the research lines, and the defined goals.

¹⁰⁶see Section 2.8

Figure 3.1: State of the art limitations and their alignment with the defined goals and research lines

3.2 Assumptions

The goals mentioned above assume the following statements to be true. These assumptions are presented before the definition of the hypotheses that would guide us throughout the development of this thesis:

- A1. Semantic Web standard languages such as RDF and OWL are expressive enough to model linguistic data and phenomena documented in lexical resources.
- A2. The *OntoLex-lemon* model, initially conceived to associate lexical information to ontologies, provides a solid and extendable foundation for the representation and exchange of lexical data extracted from lexicographic resources in the Web, as it will be used in this thesis.
- A3. Most lexicographic resources share a set of features in their description of lexical data. While the content and approach can greatly vary in them, the set of common features can be extracted to develop general models to adequate them to LLD representation.

3.3 Hypotheses

In the following we define the list of hypotheses that will be evaluated in this thesis, which are aligned with the goals and assumptions mentioned in the previous section:

- H1. Current *de facto* standards for the representation of linguistic information as LLD and the available best practices for the generation of multilingual LLD provide mechanisms to capture lexical data contained in a range of NLP and human-oriented multilingual terminologies and NLP-oriented lexica, but fall short of preserving lexical data from lexicographic records (explicit, in the form of annotations or implicitly encoded through microstructural aspects) in mono- and multilingual general human-oriented trade dictionaries.
- H2. New modelling solutions and application profiles of the *de facto* standard¹⁰⁷ and extensions to it are able to represent both the microstructure of the lexicographic

¹⁰⁷In an analogy to the Dublin Core Metadata Initiative Application Profiles (<http://dublincore.org/documents/profile-review-criteria/>), which require the objectives and scope of an application, its functional requirements, a data model of the entities described by the application, and descriptions

record as well as annotations found in mono- and multilingual general human-oriented trade dictionaries. Such modelling solutions and applications are able to preserve explicit lexical data in lexicographic entries and implicit lexical information reflected on structural aspects during their conversion into RDF.

H3. The LD paradigm brings benefits to the lexicographic work:

H3.1. By virtue of being based on graphs and URIs, the LD paradigm resolves the limitations of tree-based formats for the linking of lexicographic data in a resource by enforcing the reuse of its elements and the instantiation of typed relations.

H3.2 By virtue of being based on graphs and URIs, the LD paradigm allows for onomasiological and semasiological access on the same lexicographic resource.

H3.3 It is possible to reuse and integrate the content of lexicographic resources, if represented as RDF, in NLP systems for further lexicographic enrichment.

H4. Lexicographic resources represented as LLD allow linguists and lexicographers to formulate single queries that combine and filter lexical data from different sources to obtain data for their research, which would not be possible if the resources were not linked and did not share a common representation format.

Figure 3.2 summarises the mapping between the defined goals, the assumptions and the formulated hypotheses.

3.4 Contributions

This section lists the contributions originated from this dissertation.

C1. A survey of the state of the art on models to represent LLD in terms of the level of linguistic description they address (Chapter 2).

that detail these classes and properties, throughout this work we will refer by *application profile of a model or de facto standard* to any model to represent data from a *specific resource* that reuses classes and entities of an external model, and which also specifies the kinds of data that these classes and properties will instantiate in that specific application scenario.

Figure 3.2: Mapping of the goals, hypotheses and assumptions

- C2. Analysis of the challenges and limitations of the current *de facto* standard for the representation of lexicographic resources, the OntoLex-*lemon* model: identification of open issues not previously reported in the literature (Chapter 4).
- C3. New modelling solutions and applications of OntoLex-*lemon*, extensions to it, and design decisions for a set of representative resources: multilingual terminologies, bilingual and multilingual dictionaries, and multilingual lexico-conceptual databases (Chapter 5).
- C4. Motivation and requirements for a new module for the OntoLex-*lemon* model to represent lexical data and structural aspects from lexicographic resources grounded on lexical reasons (Chapter 6).
- C5. Resource contribution (Chapter 5): A set of technologies and lexicographical resources converted to LLD (and analysed as part of C2), comprising the following RDF datasets: Apertium dictionaries (Forcada et al., 2011)¹⁰⁸, K Dictionaries Multilingual Global Series Spanish Dataset¹⁰⁹, Terminesp¹¹⁰, Termcat (terminologies from the domains Telecommunications, Internet and Clothing Industry)¹¹¹, Zhishi.me (Niu et al., 2011)¹¹² and an excerpt from the Rayfield’s Comprehensive Georgian-English Bilingual Dictionary (noun classes) (Rayfield, 2006).
- C6. Theoretical analysis of the benefits, challenges and implications of LLD for lexicography as source format for lexicographic resources (*LD-native dictionaries*, Chapter 8).

3.5 Restrictions

Our contributions are limited with regard to the aspects mentioned in the following:

¹⁰⁸RDF dataset accessible at <http://linguistic.linkeddata.es/apertium/> (Gracia et al., 2018)

¹⁰⁹RDF data accessible through the Lexicala API at <http://www.lexicala.com/resources#dictionaries> (Bosque-Gil et al., 2016b)

¹¹⁰Original resource: <http://www.wikilengua.org/index.php/Wikilengua:Terminesp>

¹¹¹Source datasets available at <http://www.termcat.cat/es/TerminologiaOberta/>, RDF datasets for Terminesp and Termcat accesible at <http://linguistic.linkeddata.es/terminoteca/> (Bosque-Gil et al., 2015, 2016c)

¹¹²Source data accesible at <http://zhishi.me/>, RDF dataset available at <http://old.datahub.io/dataset/zhishi-lemon> (Fang et al., 2016)

- R1. Despite the assumption that most lexicographic resources share a set of features in their description of lexical data that can be extracted to develop models to account for them (see A3), the representation needs of each lexicographic resource are different and will depend thus also on varying criteria, from the intended use of the resource to the set of conditions for its conversion established by data providers. This means that no model will be able to be **exhaustive** for them all.
- R2. All the work carried out to attain the ultimate goal of this thesis is limited and dependant on the set of lexicographic resources, in their respective languages, studied for its achievement. There is no guarantee that the results, analysis and experiences reported in this work will hold for all sub-branches of lexicography, all kinds of lexicographic resources, in every language.
- R3. We will not focus on resources that were conceived in the framework of a particular theory of language or whose information is intended to serve as training for NLP systems. We limit ourselves to address resources as outcomes from a lexicographic process that also see humans (who are not necessarily experts on the area of linguistics) as potential future users.
- R4. Only lexicographic entries in the following languages were analysed: Spanish, German, English, Georgian, and Chinese. Entries in the following list of other languages where analysed in cases in which 1) the resource was a multilingual resource with one of the above languages as source/target and/or 2) the micro-structure of the lexicographic entry or format was shared with entries in one of the languages above: Italian, Portuguese, Japanese, Norwegian, Swedish, Latin, French, Catalan, Dutch. Thus, we have addressed the representation of data in five languages directly and in fourteen indirectly.
- R5. The analysis of Rayfield’s Georgian-English Bilingual dictionary was focused solely on nouns due the complexity of other entries (specially verbal and adjectival), the limited time as research stay project, and the dependency on the assistance of expert linguists fluent in Georgian to analyse the data.
- R6. The adoption, and hence, impact, of LLD both in lexicography and in research in empirical linguistics assisted by lexicographic resources highly depends on the ease of use of RDF-technologies by non-experts in the graph query language SPARQL

Protocol and RDF Query Language (SPARQL). The development of user interfaces on top of SPARQL, as well as their current lack, affects directly the overall benefits of the implementation of these technologies. Technological support for distributed or user-interface-based querying of LLOD-datasets is out of the scope of this thesis.

Table 3.3 summarises the contributions and their alignment with the defined goals, hypotheses, restrictions and assumptions.

3.6 Research methodology

In this thesis we have followed the life-cycle of incremental models. In each of the steps (identification of requirements, model development, refinement, standardisation, impact assessment and building of theoretical foundations for future work), we followed the “divide and conquer” approach, which is reflected in a series of tasks as road-map towards completion of each step.

1. Analysis of the **state of the art** on models to represent LLD and open problems and limitations.
2. **Identification and analysis of representation requirements** in lexicographic resources **not previously detected** in the literature. For each of the resources analysed and converted in this thesis, we manually inspected its data model in order to 1) select a series of existing ontologies to reuse that address the same level of linguistic description (see Section 2.4.1), and 2) detect potential mismatches (see Section 5.1) between the elements of the data model and those of the selected ontologies to reuse. The **information preservation requirements** were also discussed and defined with the data owners (if applicable) at this stage. For each lexicographic resource (see Chapter 4), and incrementally as more resources were analysed, this allowed us to compile **a list of representation needs and a general list of problematic issues** for their fulfilment as LLD with current mechanisms (see Section 6.1).
3. Iterative **model development and manual evaluation**. For each resource, we followed the work methodology proposed for LLD generation (Gómez-Pérez

Figure 3.3: Contributions and their alignment with the defined hypotheses, goals, restrictions and assumptions

et al., 2013; Vila-Suero et al., 2014a). In the case of the conversion of the K Dictionaries Global series, the application profile of the selected ontologies and the URI naming strategy definition were discussed firstly with data owners and illustrated with randomly selected entries that were manually converted to RDF or depicted graphically as instantiations of ontology elements. This step was repeated iteratively until data owners agreed with the proposed representation. In the case of Rayfield’s Georgian-English Bilingual dictionary, the model and URI naming strategy was discussed iteratively with an expert empirical linguist fluent in Georgian.

4. **Implementation: RDF generation.** In those cases in which the generation of the RDF was assumed by the authors (exception to this is the conversion of K Dictionaries Global Series, with Semantic Web Company performing this task in 2016 and K Dictionaries in 2019), the RDF was generated using the tool OpenRefine¹¹³ with its LD plug-in.¹¹⁴
5. **Evaluation.** Verification of the implementation and validation of the resulting RDF through **queries** performed locally on Apache Jena Fuseki.¹¹⁵ The results obtained from these queries allowed us to identify **errors** at several steps of the process: 1) in the modelling, and 2) in the implementation of mappings using the OpenRefine tool. Exceptions, i.e., lexicographic records that did not match the expected outcomes, were also identified here as well, and were due to either errors in the data or to rare entries whose structure and content was not prototypical of the resource and hence not addressed as part of (2).
6. **Refinement and publication:** After the evaluation, the proposed models were improved to **address the identified errors**. From this phase on, steps 3-6 were repeated until the evaluation yielded satisfactory results. Following, the resource was published – except for those cases in which this step was performed by a third-party for internal consumption (K Dictionaries’ case).

¹¹³<http://openrefine.org/>

¹¹⁴Available under the Digital Repository of Ireland’s repository on Github: <https://github.com/stkenny/grefine-rdf-extension>

¹¹⁵<http://jena.apache.org/documentation/fuseki2/>

7. **Extension:** as a result of (2) and steps (3-6), we had compiled a list of problematic issues for lexicographic resources represented as RDF and proposed a series of modelling solutions for those specific cases to serve as example for future conversions of lexicographic resources to RDF. This step was devoted to **analysing further lexicographic resources to assert whether the issues encountered in (2) and solved in (3-6) were applicable to other resources as well, and whether these additional resources posed further modelling challenges**. The results of this step served to motivate and lead the development of a new module to represent lexicographical data as part of the W3C Ontology-Lexica community group.

8. **Standarisation.** This phase involved the **development of a module for Lexicography under the umbrella of the W3C Ontology-Lexica community group** through bi-weekly telcos from July 2017 to October 2018. This step followed the methodology of **incremental models** in its own, but phase (8.1), identification of representation requirements, took as input by the output of phase (7).
 - (a) Analysis and identification of **representation requirements** in lexicographic resources. In addition to those representation requirements identified as a result of the work in this thesis until this phase, the OntoLex community provided **further dictionary entries and recent use cases** from the literature in which OntoLex-*lemon* proved not to be exhaustive.¹¹⁶

 - (b) Iterative **model development** and manual evaluation. On the basis of the data provided for discussion, the challenges identified up to the date, and the solutions provided for the resources analysed in this thesis, **a small new vocabulary was proposed as an extension to OntoLex-*lemon***. For each lexicographic entry provided for discussion in the community, the proposed vocabulary was instantiated manually and the resulting RDF was discussed with OntoLex community members (and data providers).

¹¹⁶The complete list of gathered issues is available at <https://www.w3.org/community/ontolex/wiki/Lexicography>.

- (c) **Refinement and publication.** After the manual evaluation and discussion, the proposed vocabulary was **gradually improved** to address the identified errors and inconsistencies until **agreement** was reached in the OntoLex community. Following, the module was **documented**, encoded in OWL and submitted for **publication** under a W3C namespace.
9. **Impact assessment.** This phase involved the testing of hypotheses. Four aspects were analysed: **coverage** of the developed contributions in contrast to available models for the representation of LLD; **information preservation** after the conversion to RDF; **implications** of the representation of lexicographic data as LLD; and **reuse** of the converted resources for data gathering and external projects.
 10. Building of theoretical foundations for future implementations. The last step of the research methodology delved into the theoretical benefits, drawbacks, implications and open challenges of **RDF as a means to create new lexicographical projects**, i.e., the **lexicographic resource is born as LLD** (LD-native resources).

Chapter 4

LLD in Lexicography: Challenges

This chapter analyses the gaps and challenges identified in the representation of lexicographic resources as LLD focusing on a series of representative families of resources, namely: bilingual dictionaries, multilingual terminologies, multilingual dictionaries, and lexicalisations of multilingual knowledge bases. For each of these families, we have taken one or two illustrative resources. Overall, we address the representation as LLD of a total number of six resources, covering directly five languages and indirectly fourteen.¹¹⁷ The analysis of their data and source model was performed by answering the following questions:

1. What source data needs to be preserved in the resulting RDF resource?
2. How does the source data model encode this information? (structures, tags, guidelines, etc.)
3. Which representation needs are not straightforwardly mapped into the *de facto* standard OntoLex-*lemon*?

This chapter is thus aligned with the first step in LLD generation (Gómez-Pérez et al., 2013), namely, source data and data model analysis, and will be followed by the development of modelling solutions, extensions, and application profiles of OntoLex-*lemon* (Chapters 5 and 6).

¹¹⁷We are not taking into account all language pairs covered by the Apertium family of dictionaries, as in this work we *build on* the effort devoted to its conversion to RDF (Gracia et al., 2018). The modelling of Apertium RDF serves thus as reference for the work presented in Chapters 4 and 5, but it is not a contribution of this thesis on its own.

In Section 1.2.2, we introduced the different aspects to bear in mind for the compilation of trade dictionaries in order to illustrate the considerable variety across them. According to Atkins and Rundell (2008):

- ‘the dictionary’s language (monolingual, bilingual – unidirectional or bidirectional –, multilingual),
- its coverage (general language, encyclopaedic and cultural material, terminology and sublanguages, specific area of language: collocations, phrasal verbs, idioms, etc.),
- the dictionary’s size (standard, concise, pocket edition) and the medium (print, electronic, web-based),
- the dictionary’s organisation (word-to-meaning or word-to-meaning-to-word),
- the user’s language,
- the user’s skills and
- what the dictionary will be used for: decoding (understanding, translating from) or encoding (using, translating into, teaching).

We have focused on **five** of these aspects: **language**, **coverage**, **organisation**, **medium**, and **intended use**, in order to guide the selection of the resources analysed in this thesis. Table 4.1 describes the selected resources in terms of these five features. Regarding the **language**, we consider bilingual and multilingual resources.¹¹⁸ In terms of **coverage**, the selection includes general language resources (Apertium dictionaries, K Dictionaries, Rayfield’s Georgian-English dictionary), terminological (Termesp and Termcat), and encyclopaedic ones (Zhishi.me). Since terminologies and most traditional dictionaries also differ in their **organisation**, being the former concept-oriented and the latter lemma-oriented (see Section 1.2.2), the heterogeneity in organisation is also well represented in our selection. With respect to the **medium**, all dictionaries except from Rayfield’s Georgian-English dictionary are adapted to the electronic format, and their

¹¹⁸Chapter 6 exposes a series of challenges for LLD representation that are also common to lexicographic resources of monolingual nature.

Feature/ Resource	Apertium	Rayfield's	Termesp	Termcat	K Dictionaries	Zhishi.me
Language	Bilingual	Bilingual	Multilingual	Multilingual	Multilingual	Multilingual
Coverage	General	General	Domain-specific	Domain-specific	General	Encyclopaedic
Organisation	Semasiol.	Semasiol.	Onomasiolog.	Onomasiolog.	Semasiolog.	Sema- and onomasiol.
Medium	Electronic, structured	Digitized, schemaless	Electronic, structured	Electronic, structured	Electronic, structured	Electronic, structured
Intended use	NLP systems and human users	Human users	Human users, reusable by NLP systems	Human users, reusable by NLP systems	Human users, reusable by NLP systems	NLP systems

Table 4.1: Description of the selected resources according to five of the features of trade dictionaries identified by Atkins and Rundell (2008).

content is structured according to an annotation schema, whereas Rayfield’s Georgian-English dictionary is a rendering of the print dictionary that inherits features from paper lexicography. Lastly, as regards the **intended use** of the resource, the Apertium dictionaries and Termcat terminologies provide data to be integrated in NLP applications (albeit also consumable by humans), whereas Termesp or K Dictionaries data also present annotations targeted specifically at humans users. In contrast, Rayfield’s Georgian-English dictionary is intended to solely be consulted by human readers.

4.1 Characteristics of bilingual dictionaries: the Apertium dictionaries and Rayfield’s Comprehensive Georgian-English dictionary

There is a growing body of literature in bilingual (e-)lexicography and the issues bilingual (or *translation*) dictionaries pose (see, for example, Adamska-Sałaciak (2014); Piotrowski (1994); Yong and Peng (2007), among many others). In this section we focus on two bilingual dictionaries and some of their salient features as representative examples of bilingual lexicographic resources to convert to RDF.

Bilingual dictionaries contain translations from one language into another.¹¹⁹ They can be **unidirectional** (translation from source language (L1) to target language (L2))

¹¹⁹Piotrowski points out, however, that some authors describe bilingual dictionaries rather as types of monolingual dictionaries, since these translations provide “definitions” ((Benson, Benson, Ilson 1986) as cited by Piotrowski (1994)) and synonyms ((Landau 1984), as cited by Piotrowski (1994)).

or **bidirectional** (if they provide L1-L2 as well as L2-L1). Even though they contain translations, Zgusta (1971) identified some special cases of bilingual dictionaries in which they rather resemble a multilingual one (Piotrowski, 1994), for instance, philological bilingual dictionaries, which provide a thorough description of the lexicon of a dead language in another language. Here we focus on two bilingual dictionaries of living languages, but dictionaries for Ancient Greek or Sumerian, for example, give rise to further representation challenges in RDF, including pointers to literature or archives (see Khan et al. (2017) for Ancient Greek and Chiarcos et al. (2018) for challenges in the work with Sumerian data in the LLD context). In bilingual dictionaries, in addition to the translation of the source entry (if existent in the target language) and its grammatical features, dictionary entries may also include equivalents (“[a] target-language expression which serves as an explanation in the senses of the entries” (Piotrowski, 1994)), a target language - definition of the sense of the source-language, further comments or glosses, examples, meaning discriminators (or sense indicators to distinguish senses), selectional restrictions, domain information, or semantic categories for the translations.

With this diverse content, the advantages of representing bilingual dictionaries (and translations in general) as LLD have already been explored in the literature (Gracia et al., 2014b, 2018; Khan et al., 2017). One of the benefits of LLD for the representation of (multilingual) lexical data is the potential of interlinking complementary data initially developed in isolation, and lexicographic data specifically. In addition, LLD allows to encode the translations of bilingual dictionaries in a reusable manner, making them available for consumption by third party applications. Likewise, by virtue of being represented as part of a language graph, LLD supports the use of a language as pivot to obtain indirect translations (Gracia et al., 2018). Chapter 2 mentioned several works in the current literature devoted to the conversion of bilingual dictionaries to RDF, including the series of dictionaries Apertium (Gracia et al., 2018), and the efforts in the representation of the Liddell-Scott Greek-English lexicon and the Lewis-Short Latin-English (Khan et al., 2017, 2016b).

The Apertium family of dictionaries was developed to be integrated in a open-source platform for machine translation. The Apertium RDF series (Gracia et al., 2018) was an effort to represent the data of the Apertium family as LLD. Due to the major modelling focus provided in that work, namely, the translations explicitly encoded in the source data, Gracia et al. (2018)’s approach serves as basis for the analysis provided

in this whole chapter. This decision was taken on the grounds that all lexicographic resources analysed in this thesis are multilingual and include translations in their data. We outline the features of the Apertium family as a starting point in the following section.

The second bilingual dictionary described in this section is Rayfield’s Comprehensive Georgian-English Dictionary (Rayfield, 2006). It is a paper dictionary, aimed at human consumption, and it is intended to provide English equivalents for the entire lexical corpus of the Georgian language: ancient, classical and modern, as well as literary, colloquial and dialectal (Rayfield, 2006).

These two dictionaries come from very disparate backgrounds and were developed with different purposes and target consumers in mind. However, they do share their bilingual nature and some of their representation needs. In later sections we will stress some features of multilingual dictionaries aimed at human consumption, which brings them close to Rayfield’s dictionary, as well as the characteristics of multilingual terminologies setting them closer to the Apertium family.

4.1.1 Characteristics of the Apertium family of dictionaries

The Apertium¹²⁰ family of bilingual dictionaries belongs to an open-source platform for machine translation and comprises up to 40 language pairs, among them, under-resourced languages which pose problems for statistical machine translation due to lack of data (Forcada et al., 2011; Gracia et al., 2018). These dictionaries were compiled for their use integrated in the system as letter transducers (Forcada et al. (2011), as cited by Gracia et al. (2018)), with source language entries as input and its equivalent target language entries as output. A set of the Apertium family was converted to the ISO standard LMF (Francopoulo and George (2006), see Section 1.2.3) as part of the METANET4U Project¹²¹(cf. Gracia et al. (2018)), and the data analysed for their conversion to RDF thus adheres to such a model.

Example 4.1 presents an extract of the LMF Apertium source data showing the French-Spanish translation *rouge* – *rojo* ‘red’.

¹²⁰ http://wiki.apertium.org/wiki/Main_Page

¹²¹ <http://www.meta-net.eu/projects/METANET4U/>

```

<LexicalResource dtdVersion="16">
  <Lexicon>
    <feat att="language" val="fr"/>
    <LexicalEntry id="rouge-adj-fr">
      <feat att="partOfSpeech" val="adj"/>
      <Lemma>
        <feat att="writtenForm" val="rouge"/>
      </Lemma>
      <Sense id="rouge_rojo-adj-l"/>
    </LexicalEntry>
  </Lexicon>
  [...]
  <Lexicon>
    <feat att="language" val="es"/>
    <LexicalEntry id="rojo-adj-es">
      <feat att="partOfSpeech" val="adj"/>
      <Lemma>
        <feat att="writtenForm" val="rojo"/>
      </Lemma>
      <Sense id="rojo_rouge-adj-r"/>
    </LexicalEntry>
  </Lexicon>
  <SenseAxis id="rouge_rojo-adj-rojo_rouge-adj" senses="rouge_rojo-adj-l
    rojo_rouge-adj-r"/>
</LexicalResource >

```

Example 4.1: The French-Spanish translation *rouge* – *rojo* ‘red’ from LMF Apertium FR-ES dictionary

A bilingual dictionary in Apertium is represented as a *LexicalResource* with two lexica (for source and target languages) and a series of *SenseAxis* elements. LMF is based on feature structures and attribute-value pairs. In each lexicon, the entries are described with a part-of-speech value, and they are given a sense (or meaning) and an ID following the pattern *lemma* - *part-of-speech* - *language tag*. The *SenseAxis* element in the lexical resource also receives an ID and encodes the equivalency or translation relation that links the source (or input, marked with the ‘-r’ ending, for *right*) and target senses (or output, marked with the ‘-l’ ending, for *left*). Note that sense IDs

include their source and target respectively, so senses depend on the existence of a target equivalent for them.

The **challenging** aspect in the representation of the Apertium series as RDF was essentially that **translations** were explicitly described in their data as relations between senses. Prior to the Ontolex-*lemon* model, *lemon* did not include a mechanism to represent translations explicitly, which required a new module geared specifically at translation representation, the *lemon* translation module (Gracia et al., 2014a), developed in the context of the LIDER project.

The representation needs of Apertium are taken as the **basis** for the work on lexicographic resource modelling presented in this thesis for several reasons:

1. The option of **modelling translations explicitly** as part of the modelling of a bilingual dictionary is crucial, as all the resources we analyse in this thesis are multilingual: they share with the Apertium family the information about source and target equivalents of dictionary entries.
2. The configuration of Apertium LMF lexical resources include two separate lexica that are ‘linked’ together at the sense level via *SenseAxis* elements. We consider this a design decision to hold for all lexical resources analysed, i.e., that **independent lexica for both source and target language exist**, and are linked at the sense level through translations, even though other dictionaries are not originally presented as such.
3. Apertium LMF lexica only include the following information about the entry: **lemma, part-of-speech, and translation** (for a single sense).

These are the basic representation needs **shared by all of the analysed resources**. The modelling of Apertium thus constitutes **a solid ground** upon which to build other models and application profiles of Ontolex-*lemon*, as more representation needs gradually emerge from the analysis.

4.1.2 Characteristics of the Comprehensive Georgian-English Bilingual Dictionary by Donald Rayfield

Rayfield’s Comprehensive Georgian-English Dictionary is intended to provide an English equivalent for the entire lexical corpus of the Georgian language, including ancient,

classical and modern, as well as literary, colloquial and dialectal Georgian (Rayfield, 2006), although the different layers are not separated in the organisation of the dictionary as independent lexica. The dictionary is aimed at human consumption by mainly English-speakers interpreting Georgian text or discourse, and, as the editor states, it is only secondarily conceived for the production of English sentences by Georgian speakers (Rayfield, 2006). The data contains fine-grained annotations of the meaning and use of headwords, with the list of abbreviations and conventions being described in the introductory section of the volume.

The analysis of the data was performed by the author, assisted by Jesse Wichers Schreur, an empirical linguist expert in Caucasian languages and fluent in Georgian. This work was executed in the context of a research stay in the Linked Open Dictionaries (LiODi) project at the Applied Computational Linguistics (ACoLi) research group at Goethe Universität Frankfurt. The provided dictionary data was encoded in text format (.txt) with only italics markup indicating the annotations (morpho-syntactic, usage restrictions, dialect, domain, etc.) inherited from the paper format. The paper version was used to interpret the annotations (for example, the copyright symbol - © - represents a mark of colloquial use) in the text version. A simple example noun entry, *amfsoni*, translated to English as ‘companion, comrade’, is presented below in its original encoding.

ამფსონი <i>n † ©</i> companion, comrade; crony, pal

Example 4.2: Dictionary entry for the word *amfsoni*

The annotations in Example 4.2 indicate that the entry is a noun (*n.*), whose use was colloquial (©) at it is now obsolete (†). It also has two senses, separated by semicolons: the companion one, and the crony, pal one. For each sense, alternative translations are introduced separated by commas.

This dictionary has very specific guidelines provided by the editor regarding the use of parentheses, brackets and case endings (marked with (-..)) (details on these guidelines and conventions are provided in Rayfield (2006)). The information encoded through the use of punctuation is key for the RDF representation of the resource, since it refers in many cases to spelling variants or different word forms (in the case of verbs). In Example 4.3, shown below, brackets indicate a vowel which is normally elided when case or plural suffixes begin with a vowel.

აბგარ-აბჟ[ა]რი <i>n †</i> weaponry <i>see</i> აბჯარ-აბგარი

Example 4.3: The entry *abgar-abj[a]ri*

In contrast, parentheses refer to characters that may be omitted, to verbal forms, or to case endings (shown in this order in Example 4.4).

დაფუყ(ებ)ული <i>pp</i> deflated; disappointed

დაფქვავს (დაფქვა, დაუფქვავს) 1 will grind sth into flour; 2 will blurt sth (<i>everything</i>) out

დახუჭობანა (-ას) <i>n</i> hide-and-seek

Example 4.4: Three different uses of parentheses throughout the dictionary, illustrated in the entries *dapug(eb)uli*, *dapkvavs* and *dakhuch'obana (-as)*

In addition, the entry is often headed by two headwords separated by a comma. These are often different orthographic variants of the same headword. In Example 4.5, the headword is a noun belonging to Old Georgian (to 1100 AD) with two different orthographic variants:

გძიდუ, გძილი <i>n</i> <i>†</i> partition

Example 4.5: The entry *gdzide, gdzili*

Along with these conventions regarding punctuation marks, we also find more information in the entries that made the conversion of this dictionary to RDF a challenging task. In the entry structure, this data comprises multiword expressions, cross-references (indicated by <i>see</i>), homonyms (enumerated) but collapsed under the same entry, translation alternatives (separated by comma) and different senses (separated by a semi-colon). Some entries allow for more than one part-of-speech as well.

For each sense, a range of different annotations can apply, as with the entry as a whole. In addition to the annotations concerning the dating period, temporal qualifiers, part-of-speech, and register shown in the previous examples, some entries contain abbreviations and marks for domain, context of use, dialect, register, connotations (e.g. derogative), etc. In cases in which there is no translation to English, the entry provides

a definition in English for that specific sense. Examples and multi-word expressions are commonly offered as well. Some additional information, from clarifications of use to Latin denomination (e.g. (*Convolvulus tricolor*)) also occur in italics.

The first entry shown in Example 4.6 has a part-of-speech, a temporal qualifier, and also a definition in the first sense. In its second sense, it provides a translation, together with a multiword expression associated with it, which in turn has a translation. The second entry provides the domain, botany, along with a dialect marking (*Ig*, Ingolo dialect), the translation, and a cross-reference. Note that it also indicates in brackets a thematic vowel which can be omitted.

ადლი <i>n †</i> unit of measure $\approx$ 1 metre; yardstick: თავისი
ადლით ზომავს measures (<i>sb</i>) by one's own standard

ათაღ[ა]რი <i>n</i> <i>bot Ig</i> oriental hornbeam <i>see</i>
ჯაგნცნილა

Example 4.6: Dictionary entries for the words *adli* and *atagh[a]ri*

The modelling of this dictionary as LLD posed a challenging task due to several reasons:

1. Its use of punctuation marks to encode both morphological and orthographic information, inherited from **paper lexicography**. The **lack of a formal annotation scheme to map directly to OntoLex-lemon for RDF representation** also complicates the task, since there is certain amount of heterogeneity in the range of information and structure of the entries, as opposed to a formal DTD. The conventions used throughout the work are explained in the prologue to the dictionary as a guide to the human user, and hence need to be interpreted by the modeller during the pre-processing of the data in order to establish appropriate mappings to RDF vocabularies.
2. The wide range of information provided in the resource, which is comparable to that of the K Dictionaries (see Section 4.3 in the coming pages for a thorough analysis), raises **the question whether all lexical information provided in the resource is accounted for by OntoLex-lemon**, together with the use of external vocabularies. The **representation needs** of this resource (taking

only nouns into consideration) that we addressed involved translations, dialectal information, temporal qualifiers, domain, alternative forms, synonyms, polysemy, register, dialects, and related multiword expressions.

Due to the complexity of the resource, the provided annotation and the fact that part of our converted resources are terminologies with a majority of nouns in them, we focused for our use case on just nouns (see Section 3.5, *Restrictions*), which in this dictionary includes proper nouns and abbreviated forms.

Rayfield’s Comprehensive Georgian-English dictionary is a good example of a bilingual dictionary aimed at human consumption which shows numerous **features inherited from paper lexicography**. The source data we studied also contrasts with other dictionaries already developed in e-lexicographic projects where there is a rich annotation in XML (e.g. LMF or a proprietary DTD) that facilitates their mapping to RDF models. Also, this dictionary raises the question of **the scope of the representation**: since it is **marked for human consumption**, no other distinction beyond italics and punctuation marks is made in terms of the different fields of information. The information is left for interpretation by the human user, being equally marked the register or the part-of-speech, in contrast to the resources we will find in the next sections. Whether to make these distinctions explicit (to be reused in the future) or not in the RDF version as a result of an interpretation process, instead of just recorded as string values in comments, labels or notes, is then left to the LD expert making the conversion and the intended use of the resulting resource.

4.2 Characteristics of multilingual terminologies: **Terminesp** and **Termcat**

Terminological resources are concept-oriented lexical resources: lexical because they contain data about a specialised lexicon, for example, the lexicon used in a particular branch of knowledge; and concept-oriented because they aim at describing the concepts used in a specialised field and the terms denoting these concepts. Terminological resources can be viewed as knowledge organisation systems for a specific domain, which, according to the definition of the Networked Knowledge Organization Systems Working

Group¹²² includes all types of schemes for organising information and promoting knowledge management (Hodge, 2000). They range from terms lists, including authority files, glossaries, dictionaries, and gazetters, to relationship lists where concept-based terms are related to one another, by, for example, hyperonymy, equivalence, or association relations (thesauri resources) (Hodge, 2000).

The variety of terminological resources leads to an heterogeneity in their content as well. Domain specific lexica include information about the use of terms, and often also their translation, their definition, and their grammatical information (e.g. part-of-speech, gender if applicable, etc.). Examples, scientific denominations, hyponyms and hypernyms, usage notes, links to specialised documentation from which the term was extracted, sources, media files, etc. is also recurring information in terminologies.

Terminologies are frequently developed independently and following non-standard formats. By converting them to LLD, their data is exposed and accessible in a standardised manner (through RDF and SPARQL respectively). Linking them to other external resources facilitates their discovery, reuse and integration, but this not only enables humans and software agents an easier access, but linked multilingual lexical resources represent a potentially relevant source of knowledge for NLP-based applications developed in the fields of machine translation, content analytics, multilingual information extraction, word sense disambiguation or ontology localisation (Bosque-Gil et al., 2016c). The importance of integrating terminologies and other types of language resources is proven also by the high number of projects that aim at combining language resources as LLD and provide a single access point. This has already been explored with the creation and publication of thesauri, vocabularies and terminology repositories, especially in the environmental and geological domain (Albertoni et al., 2014; Caracciolo et al., 2013; Rüther et al., 2010), as well as in the financial (Neubert, 2009) and linguistics fields (Chiarcos and Sukhareva, 2015b; Cimiano et al., 2011b; Kemps-Snijders et al., 2008).

As representative resources, for this thesis we chose the Terminesp database and Termcat terminological resources. Terminesp¹²³ is a multilingual terminological Spanish database created by the Spanish Association for Terminology, AETER, by extracting

¹²²At its initial meeting at the ACM Digital Libraries'98 Conference (see Hodge (2000)).

¹²³<http://www.fundeu.es/tema/terminesp/>

the terminological data from UNE (*Una Norma Española*¹²⁴, ‘a Spanish norm’) documents (similar to ISO Standards that, in turn, were produced by committees of experts in different domains) produced by AENOR (Asociación Española de Normalización y Certificación). Termcat¹²⁵, the Catalan Terminological Centre, gathers terminologies in Catalan divided by domain, and are available through the portal *Terminologia Oberta*.¹²⁶ We deemed these resources as representative because of their multilingual nature and, altogether, they gather many of the annotations mentioned above, from scientific denominations to translations, definitions, terminological norms, semantic relations and abbreviations.

4.2.1 Terminesp

The conversion of Terminesp to RDF had already been addressed in the literature (Gracia et al., 2014b) before building upon it for this work. In that effort, Terminesp data served as a validating example of a terminological resource to which the *lemon* translation module would apply. The analysis we present here and the modelling solutions to account for the data explained in the next chapter are detailed in Bosque-Gil et al. (2015) and Bosque-Gil et al. (2016c).

The database amounts to more than thirty thousand terms for which definitions, norms and translations from Spanish to a variety of languages are provided, along with scientific denominations in Latin. Norms are defined over a range of domains, from aeronautics and electro-technical engineering to fruit nomenclature. An entry in Terminesp consists of the definition of the term, the norm from which the term is extracted, the norm title, and, if available, the translation of the term to one or several different languages: German, French, Italian, Swedish, and English. Some definitions and definition notes in these languages are included as well. The entries themselves range from simple nouns and adjectives to complex nominal, prepositional and adjectival phrases.

¹²⁴A UNE document is a document issued by the Spanish Association for Terminology. It provides technical specifications based on obtained technological results and technological development. These norms must be approved by a recognised normalisation organisation. See “Normalización”. Asociación Española de Normalización y Certificación. p. 3. Available at <http://www.aenor.es/DescargasWeb/normas/que-es-una-norma-es.pdf>, last accessed 21.08.2019.

¹²⁵<http://www.termcat.cat/>

¹²⁶http://www.termcat.cat/en/Terminologia_Oberta/.

¹²⁷http://www.wikilengua.org/index.php/Terminesp:admitancia_cin%C3%A9tica

Norm	UNE_21302-801
Norm title	Vocabulario electrotécnico: capítulo 801. Acústica y electroacústica
Spanish term	admitancia cinética
German term	Bewegungsadmittanz
English term	motional admittance
French term	admittance cinétique, admittance motionnelle
Definition	En un transductor, diferencia entre la admitancia eléctrica en carga y su admitancia eléctrica cuando está mecánicamente bloqueado.
Note	NOTA - Esta definición es válida principalmente para transductores con acoplamiento por transformador.

Table 4.2: Terminesp entry for the Spanish term *admitancia cinética*¹²⁷

Table 4.2 shows a record for the entry *admitancia cinética* ‘motional admittance’.¹²⁸ The entry is headed with the norm document ID from which it was extracted, followed by the title of that norm. After the translations, in this case, into four languages (other entries include Latin denominations and translations into Italian and Swedish as well), the entry lists the definition. Occasionally, the editors included a note which specifies in which conditions the specific definition holds and any other relevant information for its interpretation.

Entries such as the previous one raised modelling doubts so as to how exactly could we model definitions and definition notes in a way that reflects that the note string is related to the definition but not to be confused with the definition string itself, as well as how to record a terminological norm (norm codes) or standard and its entry in the most reusable way across terminologies.

Another important issue is the scientific naming of Terminesp terms. The entries that provide a Latin term come mainly from the botanical domain, denoting most of them fruits and vegetables, e.g. Sp. *uva* ‘grape’, Lat. *Vitis vinifera* Linnaeus. The questions that arise when dealing with scientific denominations are a) whether to

¹²⁸En. *Norm*: Electrotechnical vocabulary: Chapter 801. Acoustics and electroacoustics. *Definition*: Referred to a transducer, the difference between the electrical admittance on charge and its electrical admittance when it is mechanically locked. *Note*: NOTE - This definition applies mainly to transducers with transformer coupling. Extracted from Bosque-Gil et al. (2015).

treat them as translations, sticking to the fact that we are connecting terms in two different languages, b) whether to treat them as terminological variants on the grounds that they are international scientific denominations: in this sense, scientific names are considered a specific type of variants subject to domain and register rather than to any other factor. Also, they are internationally accepted over scientific communities and can appear in texts written in any language, provided that the register and the domain are adequate for their use. In this case, Latin denominations are viewed as register variations and a kind of sense relations. Another option could be to consider the relation between scientific and common terms as a special case of synonymy (Soergel et al., 2006) or as a relation to be captured by `sameAs` properties (which might be refined into `hasScientificTaxonomicName` and `isScientificTaxonomicNameOf` (Sini et al., 2009)) in resources that encode taxonomic knowledge, for instance.

In sum, the publication of **Terminesp** as RDF involves the following aspects:

1. The representation of **translations**, as we saw in the case of Apertium.
2. The representation of features specific to terminological resources such as:
 - **Norm** information
 - **Latin** denominations
 - **Definition notes**
3. Coverage of other lexical information detected in some of the entries of the resources analysed in next sections (see Section 4.3), such as **definitions**.
4. The augmentation of the database with part-of-speech information: in order to link this resource as RDF to other resources, we plan to rely on a common URI naming strategy for the conversion of all of them to RDF (details on this are provided in Chapter 5). However, the designed URI naming strategy requires the use of the part-of-speech of terms in the URI, and hence the augmentation of the database with this information becomes crucial.

4.2.2 The representation of Termcat as RDF

Termcat¹²⁹ is the Catalan Terminological Centre. At its portal for open¹³⁰ terminologies, *Terminologia Oberta*, there is a set of multilingual terminologies in Catalan divided by domain, most of them available in XML, HTML and/or PDF format. Since the overlap of Termcat entries with those of Terminesp was a key factor in order to show the potential of LLD in future steps of this work, we selected domains that are shared by both terminologies: *Internet*, *Telecommunications*, and *Electronics*. The analysis we provide here was published in Bosque-Gil et al. (2016c).

These Termcat domain lexica include translations from Catalan to Spanish, English, and, if available, to French, as well as synonyms, abbreviated forms, acronyms, and obsolete or dismissed denominations (the latter not modelled in the current work). Morphological information on part-of-speech and gender is also provided.

An example of a simple entry in a Termcat lexicon is presented in Example 4.7. The term *administrador -a de xarxes* ‘network administrator’ (masc. and fem.) is described in the Catalan domain lexicon *Internet i Societat de la Informació*, record 32. The part-of-speech of both the main term and the synonym (*remissió* ‘remission’) are provided in the attribute *categoría* ‘category’, and translations receive an *equivalent* value in the *tipus* ‘type’ attribute. Note that, in the XML source, two terms are in Catalan, with the first one being of type “principal” and the second one “remissió”, cross-reference. The *jearquia* ‘hierarchy’ attribute is always “terme pral.”, or principal, preferred term. In the dictionaries we analysed, the values of this hierarchy attribute can either be synonyms or abbreviations, while in the rest of Termcat dictionaries they can refer to re-directions, commercial denominations, obsolete denominations, or absolute synonyms.¹³¹

As Terminesp entries, Termcat records are concept-oriented, i.e., a polysemous term will be recorded in (at least) two different entries. The selected terminologies share with Terminesp the fact that they include translations, but, in contrast to Terminesp, they lack definitions and include synonyms. Also, while Terminesp lacks morphosyntactic information, Terminesp includes part-of-speech and gender values, initialisms, and abbreviations for each term. This comparison is presented in Table 4.3.

¹²⁹<http://www.termcat.cat/>

¹³⁰The licence of these terminologies is Attribution-NoDerivs 3.0 Unported (CC BY-ND 3.0). More information at <http://creativecommons.org/licenses/by-nd/3.0/deed.en>

¹³¹See the list of abbreviations used in Termcat, available at Termcat.ABREVIACIONSUTILITZADES, v. 26denovembrede2010.<http://arxiu.termcat.cat/criteris/abreviacions-utilitzades.pdf>

```

<fitxa num="32">
  [...]
  <denominacio llengua="ca" tipus="principal" jerarquia="terme pral."
    categoria="n m, f">administrador -a de xarxes</denominacio>
  <denominacio llengua="ca" tipus="remissio" jerarquia="terme pral."
    categoria="n m, f">gestor -a de xarxes</denominacio>
  [...]
  <denominacio llengua="en" tipus="equivalent" jerarquia="terme pral."
    categoria="">network administrator</denominacio>
  [...]
</fitxa>

```

Example 4.7: XML of the entry ‘administrador -a de xarxes’

Resource /Feature	Termesp	Termcat
Lemma	✓	✓
POS	×	✓
Gender	×	✓
Definition	✓	×
Notes to definitions	✓	×
Translations	✓	✓
Initialisms and abbreviations	×	✓
Synonyms	×	✓

Table 4.3: Data provided in Termesp and Termcat

The **modelling issues** that Termcat raises concern different aspects:

1. Semantic relations (**synonyms**) and term types (**abbreviations and initialisms**)
2. The future **integration with other terminologies**. For example, some aspects to address were whether *lemon* lexica will be extracted from the terminologies, following Apertium’s lexica-inside-lexical-resources approach, and the resource is understood to include concepts that are lexicalised by those entries, or whether the

resource itself should be treated as lexicon containing entries in different languages.

4.3 Characteristics of multilingual dictionaries: the case of K Dictionaries

Multilingual dictionaries are very close to bilingual dictionaries in that they raise the same representation needs regarding translations. They also have a source language, but they offer a set of target languages instead of a single one. This was the case of the multilingual terminologies explained in the previous section, *Terminesp* and *Termcat*, which had Spanish and Catalan as source languages respectively, and a series of target languages different in each resource. Multilingual dictionaries, as dictionaries, are still subject to other categorisations with regard to their size (standard, concise, etc.), the medium (print, electronic), the language of the target user and whether the dictionary will be used for decoding or for encoding (see Section 1.2). Depending on these parameters, the contents of multilingual dictionaries vary to better accommodate different needs. However, they share some features with monolingual and bilingual dictionaries: grammatical data, pronunciations, senses, sense disambiguators, usage notes, domain annotations, definitions, usage examples, etymology, derivations, along with (related) multi-word expressions, proverbs, compounds, etc. On top of this, they contain structural information related to groupings, sub-sense hierarchies, and ordering, as well as mediostructure features such as cross-references.

The challenge here lies in the representation of this information in a LD-compliant way, which involves a careful analysis of the (proprietary) formats in which the data are encoded, and the definition of mappings to common models in order to ensure that the information in the original resource is preserved in its RDF counterpart.

Examples of works that show the increasing interest in the LD community to convert (multilingual) lexicographic resources to RDF, not necessarily terminologies, include the efforts undertaken as part of the COST Action IS1305 European Network of e-Lexicography (ENeL)¹³² for a range of dictionaries (dialectal, bilingual, monolingual, historical, etc.) (Declerck and Wandl-Vogt, 2015), the already mentioned Apertium RDF (Gracia et al., 2018) effort, the work on Caucasian languages resources and etymological data carried out as part of the LiODi project (Abromeit et al., 2016b; Chiarcos

¹³²<http://www.cost.eu/actions/IS1305>

and Sukhareva, 2014b; Chiarcos et al., 2017), and the conversion of Wiktionary to RDF into as DBnary (Sérasset, 2015; Tchechmedjiev et al., 2014). It is important to highlight here the recent European Lexicographic Infrastructure (ELEXIS)¹³³, which aims to create an infrastructure to enable access to high quality lexical data by interconnecting existing lexicographic resources via common standards and conversion tools, among other aspects.

The Spanish dictionary of K Dictionaries (KD) multilingual Global series is an illustrative example of a rich multilingual digital dictionary compiled both for machine and human consumption: the lexical description is fine-grained and the resource shares numerous features with other dictionaries and resources discussed in the previous sections (Apertium series, Rayfield’s Georgian-English dictionary, Termesp, Termcat), including some additional aspects (e.g. a proprietary DTD, and round-tripping requirements for the conversion - see later in this same Section 4.3) not yet addressed.

This dictionary belongs to a set of multilingual, multi-layered dictionaries for 25 European and Asian languages developed by KD.¹³⁴ Each dictionary has a monolingual layer, in this case, Spanish, which is augmented with translation equivalents from bilingual versions for that language.¹³⁵

The work carried out as part of this thesis with KD’s data was part of the Linked Data Lexicography for High-End Language Technology project (LDL4HELTA)¹³⁶ between KD and Semantic Web Company, whose aim was to combine the lexical resources of KD from multilingual and pedagogical lexicography to develop new products and services to enable cross-lingual search, among other goals. Our role in the project was to develop a prototype to model the above mentioned Spanish dictionary as LD that would also be applicable to the English and German datasets, although not exhaustive for these two languages. The results of this work are elaborated in Bosque-Gil et al. (2016b).

The dictionaries from the Global series are rich lexicographical resources in a XML format with a proprietary DTD. The source data was provided by KD, along with the style guidelines for editors, the DTD itself, and additional documentation on specific tags used in the files. The Spanish dataset includes translation equivalents of headwords

¹³³<http://elex.is/>

¹³⁴<http://www.lexicala.com/>

¹³⁵<http://www.lexicala.com/resources#dictionaries>

¹³⁶<https://ldl4.com/>

to Brazilian Portuguese, Norwegian, Japanese and Dutch. Together with the Spanish dictionary, KD facilitated the English and German datasets to explore shared features so as to identify common representation needs. The associated guidelines were consulted throughout the exploratory analysis of the original data to understand the use of each tag in the DTD.

An entry in these datasets encodes grammatical and phonetic information about the headword at the *headword container* level (with the **HeadwordCtn** tag), although geographical usage restrictions and pragmatic aspects may be listed here as well, e.g. register. The grammatical information can be provided both with data about gender (in the Spanish and German datasets), number, subcategorisation, or inflection of headwords, being each type of data indicated with its own DTD tag. A snippet of the headword container of the word Sp. *abadía* ‘abbey’ is given in Example 4.8.

```
<Entry HomNum="" hw="abadía" identifier="EN00000019" pos="noun">
  <DictionaryEntry identifier="DE00000020" version="1">
    <HeadwordCtn>
      <Headword>abadía</Headword>
      <Pronunciation>aβa'ðia</Pronunciation>
      <PartOfSpeech value="noun" />
      <GrammaticalGender value="feminine" />
    </HeadwordCtn>
    [...]
  </DictionaryEntry>
</Entry>
```

Example 4.8: Extract of the entry *abadía* ‘abbey’

The senses of the headword follow this first container and each of them encapsulates its definition, translations equivalents of the headword, usage examples and their translations, along with antonyms, synonyms, selectional and usage restrictions (these restrictions are marked with the **RangeOfApplication** tag), register, or sense disambiguators, for instance. Example 4.9 shows the sense SE00007353 for the Spanish entry *arriscado*. This entry, in its first sense, means ‘rugged, craggy’. However, in the sense shown below, it means ‘bold’, and it can only be predicated of a person (marked with the tag **RangeOfApplication**) in the formal register (encoded with the tag **Register** with value *formal*.)


```

<Entry HomNum="" hw="arriscado" identifier="EN00004786" pos="adjective">
  <DictionaryEntry identifier="DE00005471" version="3">
    [...]
    <SenseBlock>
      [...]
      <SenseGrp identifier="SE00007353" version="3">
        <Register value="formal" />
        <RangeOfApplication>persona</RangeOfApplication>
        [...]
        <Definition>que va por lugares peligrosos o difíciles de
        escalar</Definition>
        [...]
      </SenseGrp>
    </SenseBlock>
  </DictionaryEntry>
</Entry>

```

Example 4.9: An extract of the senses of the entry *arriscado* ‘rugged, craggy’, ‘bold’ (if referred to a person in the formal register)

Example 4.10 shows a synonym for *abadía* in Spanish, its definition, *monasterio con territorio propio* ‘monastery with (its) own territory’ and the *translation container* with the Dutch translation for *abadía*: *abdiĳ*. Morphological, syntactico-semantic and pragmatic information about the translation may be provided as well (see **GrammaticalGender** in the example).

In addition to headword translations, senses comprise usage examples and their translations to other languages. Example 4.11 shows the translation into Dutch of a usage example of the first sense of *abadía*, «la Abadía de Westminster» ‘Westminster Abbey’.

Moreover, some senses include a **CompositionalPhrase** element that encodes idioms, collocations and frequent combinations. These groupings, in turn, have their own definition and set of senses with the corresponding translations, usage examples, and translated examples. Example 4.12 shows that one of the senses of the entry *agente* ‘agent’ provides the **CompositionalPhrase** *agente comercial* ‘commercial agent’, with its own definition.

```

<SenseBlock>
  <SenseGrp identifier="SE00000039" version="1">
    <Synonym>convento</Synonym>
    <Definition>monasterio con territorio propio</Definition>
    <TranslationCluster identifier="TC00000097" text="monasterio con territorio
      propio" type="def">
      <Locale lang="nl">
        <TranslationBlock>
          <TranslationCtn>
            <Translation>abdijs</Translation>
            <GrammaticalGender value="feminine" />
          </TranslationCtn>
        </TranslationBlock>
      </Locale>
    <...>
  </TranslationCluster>
  <...>
</SenseGrp>
<...>
</SenseBlock>

```

Example 4.10: Extract of the translations of the headword *abadía* ‘abbey’

```

<ExampleCtn type="sid" version="1">
  <Example>la Abadía de Westminster</Example>
  <TranslationCluster identifier="TC00000098" text="la Abadía de Westminster"
    type="exmp">
    <Locale lang="nl">
      <TranslationBlock>
        <TranslationCtn> <Translation>de Abdijkerk van
          Westminster</Translation>
        </TranslationCtn>
      </TranslationBlock>
    </Locale>
  <...>
</TranslationCluster>
</ExampleCtn>

```

Example 4.11: XML of the entry *abadía* ‘abbey’ with a usage example and its translation into Dutch

```

<DictionaryEntry identifier="DE00002242" version="1">
  [...]
  <HeadwordCtn> <Headword>agente</Headword> [...]</HeadwordCtn>
  <SenseBlock>
    [...]
    <SenseGrp identifier="SE00002802" version="1">
      [...]
      <CompositionalPhraseCtn version="1">
        <CompositionalPhrase>agente comercial</CompositionalPhrase>
        <Definition>persona que vende para una empresa por una
          comisión</Definition>
        [...]
      </CompositionalPhraseCtn>
    </SenseGrp>
    [...]
  </SenseBlock>
</DictionaryEntry>

```

Example 4.12: An extract of the entry *agente* ‘agent’. Sense SE00002802 includes a description of the compound *agente comercial* ‘commercial agent’.

Given that the target languages in these dictionaries include Japanese (and other languages with non-Latin alphabets in the rest of the Global series), there is also information about an `AlternativeScripting` in the translation section concerning those headwords or examples, with the transliteration of the translation equivalent. Example 4.13, related to Example 4.11, shows the translation of «La Abadía de Westminster» ‘Westminster Abbey’ to Japanese, with the transliteration in Latin alphabet (rōmaji).

While dictionary entries, indicated with the tag `DictionaryEntry`, are the building blocks of the dataset, there is also a structure above them, grouping them together in some cases: the `NestEntry` tag. The use of this tag depends on the guidelines for each dictionary, with slight differences among them. Example 4.14 shows a nest entry grouping the German dictionary entries *besuchen* ‘visit (v.)’, *Besuch* ‘visit’ (n.) and *Besucher* ‘visitor’, all of them sharing the same root.

```

<ExampleCtn type="sid" version="1">
  <Example>la Abadía de Westminster</Example>
  <TranslationCluster identifier="TC00000098" text="la Abadía de
    Westminster" type="exmp">
    [...]
    <Locale lang="ja">
      <TranslationBlock>
        <TranslationCtn>
          <Translation>ウェストミンスター寺院 (じいん) </Translation>
          <AlternativeScripting type="kanji"></AlternativeScripting>
          <AlternativeScripting type="romaji">Wesutominutaa
            jiin</AlternativeScripting>
        </TranslationCtn>
      </TranslationBlock>
    </Locale>
  </TranslationCluster>
</ExampleCtn>

```

Example 4.13: Extract of the entry *abadía* ‘abbey’, with the translation of the usage example «La Abadía de Westminster» ‘Westminster Abbey’ to Japanese. The transliteration to Latin alphabet (rōmaji) is also provided.

Tags like the ones we showed in the previous examples include attributes with a given value, for example, `<GrammaticalGender value="feminine" />`. These values, just as *feminine* in the example, either come from a list of predefined values or they are manually inserted by the editor of the dictionary into what is called a *free value* attribute. While the predefined values are a close list whose members can be mapped in the modelling step to the categories available in the linguistic data category registries, free value attributes can be any string provided by the editor and do not necessarily have a counterpart in a linguistic data category registry. Example 4.15 shows an extract of the entry *barbería* ‘barber’s shop’. Its sense SE00009794, whose definition in English can be translated to ‘profession of the person who performs haircuts and shaves men’, includes a subject field annotation with a free value *occupations*. Here *occupations* does not belong to the list of subject field values already pre-defined in KD’s DTD.

```

<Entry HomNum="" hw="besuchen" identifier="EN00002666" pos="verb">
  <NestEntry>
    <DictionaryEntry identifier="DE00003297" version="1">
      <HeadwordCtn>
        <Headword>besuchen</Headword>
        [...]
      </HeadwordCtn>
      [...]
    </DictionaryEntry>
    <DictionaryEntry identifier="DE00003298" version="1">
      <HeadwordCtn>
        <Headword>Besuch</Headword>
        [...]
      </HeadwordCtn>
      [...]
    </DictionaryEntry>
    <DictionaryEntry identifier="DE00003299" version="1">
      <HeadwordBlock>
        <HeadwordCtn>
          <Headword>Besucher</Headword>
          [...]
          <GrammaticalGender value="masculine" />
          [...]
        </HeadwordCtn>
        <HeadwordCtn>
          <Headword>Besucherin</Headword>
          [...]
          <GrammaticalGender value="feminine" />
          [...]
        </HeadwordCtn>
        [...]
      </HeadwordBlock>
      [...]
    </DictionaryEntry>
  </NestEntry>
</Entry>

```

Example 4.14: A nested entry grouping the German dictionary entries *besuchen* ‘visit (v.)’, *Besuch* ‘visit’ (n.) and *Besucher* ‘visitor’, all of them sharing the same root.

```

<DictionaryEntry identifier="DE00007137" version="2">
  <HeadwordCtn>
    <Headword>barbería</Headword>
    [...]
  </HeadwordCtn>
  <SenseBlock>
    [...]
    <SenseGrp identifier="SE00009794" version="2">
      <SubjectField freeValue="occupations" value="notPredefined" />
      <Definition>oficio del que corta el pelo y afeita a los hombres</
        Definition>
      [...]
    </SenseGrp>
  </SenseBlock>
</DictionaryEntry>

```

Example 4.15: Extract of the entry *barbería* ‘barber’s shop’, with a free value tag

Furthermore, one of the conditions for this modelling prototype was that information had to be preserved in order to guarantee the reconstruction of the original XML entries from their RDF representation, allowing for a round trip from the original format to the LD version and back. This implies that data which do not encode linguistic information also had to be considered in the analysis step, as they serve towards this reconstruction. For example, homographs received an attribute to maintain them in an ordered list, the **homographID**: e.g. blow (noun) **homID**=1 and blow (verb) **homID**=2, as separate entries.

During the analysis, a number of modelling doubts arose which were related to the use of certain tags (e.g. synonyms, cross-references) and to annotation differences across the Spanish, German and English dictionaries. Some of these discrepancies might be due to the fact that languages require different structures and elements to describe them (e.g. we only found the tag **Case** in the German dataset), whereas other discrepancies, such as, for instance, viewing a transitive verb and its reflexive use as two senses of the same entry or as two dictionary entries, are structural differences which require further analysis to discern whether they are motivated by the editor’s preferences or motivated by the results of corpus analysis. In addition, some errors were detected on the editors’ side (e.g. the incorrect use of a tag or its free value attribute), but this is frequent in complex dictionaries with several versions, as has been already pointed out (cf. Declerck and Wandl-Vogt (2015)).

In sum, the task of defining a model proposal for the representation of these datasets in RDF was a **challenging** one due to **three** broad aspects:

1. These dictionaries provide information not addressed by *OntoLex-lemon*, for example, **usage examples**, **translations of examples**, **selectional restrictions as they occur in a dictionary** (not as domain and range restrictions on properties of an ontology), along with **inflectional groupings**, **nested entries**, **compositional phrases**, **restrictions of the usage** (e.g. geographical, register, etc.) at both the headword and the sense level, etc.
2. The comparison between the tags and values in the proprietary DTD and LexInfo would lead to two different scenarios:
 - (a) **Mismatches** between the DTD values of a tag and LexInfo classes (e.g. predefined values of a tag in the DTD are individuals from a class in LexInfo that is not compatible with the tag in KD)
 - (b) **Different level of granularity** in the predefined values in the DTD and the individuals in LexInfo (e.g. KD's DTD **possessive** vs. **lexinfo:possessiveAdjective**, **lexinfo:possessiveDeterminer**, etc.)
3. The design decisions taken prior to the model definition phase involved the **preservation of all information available in the XML**, which is intrinsically related to the creation of elements that must be both *OntoLex-lemon* and LexInfo compliant, as well as applicable to the three datasets, and which should allow for the reconstruction of the source XML back from the LLD version. This translates into the preservation of identifiers of entries and dictionary entries, aspects related to the display of the data in the interface, groupings, number of homographs, etc. The need to represent structural information was one of the motivations for the development of the *OntoLex-lemon* Lexicography Module (see Chapter 6) to minimise information loss in the conversion of lexicographic resources to LLD.

Due to these three aspects, the Spanish, English and German dictionaries of KD's multilingual Global series are illustrative examples of resources whose representation needs build upon the ones previously explained with respect to Apertium, Terminesp, and Termcat, and add further complexity due to the richness of their description and the requirement to preserve all source data in their RDF representation.

4.4 Characteristics of multilingual encyclopedic resources: the case of Zhishi.me

So far, the resources analysed in this thesis are language resources in their original format that we planned to convert to LLD. To this aim, their analysis was driven by the need to detect representation requirements to meet in the LLD model (based on *OntoLex-lemon*), following the second step in the methodology for LLD generation (Gómez-Pérez et al., 2013). However, LLD language resource can also be *generated* by adding linguistic information (or making it explicit) to encyclopaedic resources (for example, Wikipedia) or to their LD version (e.g. DBpedia (Auer et al., 2007)).

The *lexicalisation* of an encyclopaedic resource already in LD format (and its underlying ontology) involves the linguistic description of its concepts, properties and individuals. More specifically, it provides information about which lexical entries in natural language denote the elements that constitute the encyclopaedic resource, and what their morphosyntactic properties, their syntactic behaviour, and their semantic relations to other entries are, etc. The benefits of ontology lexicalisation, along with the lexicalisation of those resources that populate an ontology, have already been explored (Buitelaar et al., 2011). One of these benefits is that it enables knowledge acquisition from text (e.g. morphological and syntactic variants of an entity). Furthermore, multilingual knowledge access allows to know which lexical entries in different languages refer to the same ontology entity, facilitating querying (Buitelaar et al., 2011).

The guidelines for LLD generation of multilingual dictionaries¹³⁷ that were published by the W3C BpMLoD Community Group are focused on BabelNet (Ehrmann et al., 2014; Navigli and Ponzetto, 2010). BabelNet is a multilingual encyclopaedic dictionary and ontology aggregating content from WordNet, Wikipedia, OmegaWiki, the Open Multilingual Wordnet¹³⁸, Wikidata¹³⁹ and Wiktionary. Some of the points to address during the modelling step, as stated in the guideline, involve the mapping between the lexical entry and its senses in the resource, and the addition of definitions. Other common linguistic data in these resources, for instance, grammatical information and examples of use, are shared by multilingual dictionaries and resources presented in the previous sections. As most of the literature has been focused on languages with a high

¹³⁷<https://www.w3.org/2015/09/bpmlod-reports/multilingual-dictionaries/>

¹³⁸<http://compling.hss.ntu.edu.sg/omw/>

¹³⁹https://www.wikidata.org/wiki/Wikidata:Main_Page

presence in the LLOD cloud, such as English, modelling data in other languages poses additional problems, as discussed in Section 4.1.2 in relation to Georgian, and as we will show later in this section in the case of Chinese. In addition to the encoding of linguistic data, the organisation of the resource also plays a role in the modelling. For example, DBnary (Sérasset, 2015), the RDF version of Wiktionary, had to account for Wikipedia pages that bundle together different lexical entries (see Section 2.5).

Zhishi.me¹⁴⁰ (Niu et al., 2011) is an encyclopaedic resource that gathers RDF triples from the three largest Chinese encyclopaedic Web sites: Baidu Baike¹⁴¹, Hudong Baike¹⁴² and Chinese Wikipedia¹⁴³. Its lexicalisation, including its modelling, generation and linking to BabelNet and DBpedia to identify translations, is addressed in detail in Fang et al. (2016). In this thesis we will concern ourselves solely with the representation needs to consider during the modelling step as part of that endeavour.

In Zhishi.me, a sample input label query for 离婚 ‘divorce’ redirects to a disambiguation page with different possible concepts denoted by the given input label, with a list of links to the different knowledge bases (Baidu, Hudong and Chinese Wikipedia). The option to merge content in one single page as desired by the user is provided as well, shown in Figure 4.1. In some cases, the label might be ambiguous and this leads to several pages from each database. In our example, 离婚 leads to six pages in total: four pages coming from the Chinese Wikipedia, one coming from Baidu and one from Hudong. The content describing the concept *divorce* includes an abstract, its written representation in Madarin Chinese, alternative forms in simplified Chinese, translations to English, its pin-yin transliteration, a series of definition or explanations, and a list of internal links and metadata.

The lexicalisation of Zhishi.me using *lemon* was **challenging** due to the fact that this resource differs from the others analysed in this chapter in its data on **dialectal variants, different alphabets and transliteration options, and the need to keep track of the links to disambiguation pages coming from the encyclopaedic resources**. After the work with the previous mentioned multilingual resources, the suggested representation of translations and alternative forms provided already a solid basis, while the guidelines to convert BabelNet mentioned above served

¹⁴⁰<http://zhishi.me/>

¹⁴¹<http://baike.baidu.com/>

¹⁴²<http://www.baike.com/>

¹⁴³<https://zh.wikipedia.org/>

as walk-through as well. The analysis of Zhishi.me proved that many of the representation needs of its lexicalisation were in fact shared by other multilingual lexicographic resources, so that general approaches can be developed to account for them.

4.5 Summary

Lexicographic resources have different sets of features for their description of lexical data that depend on varying criteria. As we mentioned in the introduction to this chapter, the analysis we conducted was aimed at providing answers for the following set of questions for a given lexicographic resource:

1. What source information needs to be preserved in the resulting RDF resource?
2. How does the source data model encode this information? (structures, tags, guidelines, etc.)
3. Which representation needs are not straightforwardly mapped into the de facto OntoLex-*lemon* standard?

Table 4.4 provides a summary on the features of each resource that are challenging to address as part of the next step in the LLD generation methodology: model definition and URI naming strategy.

zhwiki:离婚
hudong:离婚
baidu:离婚 (脱离夫妻婚姻关系)

Index

- zhishi:abstract
- infobox
- dcterms:subject
- zhishi:relatedImage
- zhishi:internalLink
- zhishi:externalLink
- zhishi:relatedPage

zhishi:abstract

离婚是指夫妻双方通过协议或诉讼的方式解除婚姻关系,终止夫妻间权利和义务的法律行为。按照我国《婚姻法》的规定,如感情确已破裂,调解无效,应准予离婚。夫妻“感情确已破裂”是判决离婚的法定条件。解除婚姻关系,后专指通过法律程序解除夫妻关系。《晋书愍怀太子遹传》:“初,太子之废也。妃父王衍表请离婚。”南朝宋刘义庆《世说新语贤媛》:“贾充前妇是李丰女,丰被诛,离婚徙边。”宋陆游《老学庵笔记》卷三:“郭知运以离婚为逐客。”曹禺《北京人》第二幕:“他,他疯了,他要离婚。”印度法院准许一对夫妻用网络电话办理离婚,他们因而成了全球第一对以网络电话离婚的夫妻。

>>TOP

infobox

外文名	divorce; get a divorce; break a marriage
拼音	li hun
解释	解除婚姻关系
法定条件:	感情确已破裂
解释:	解除婚姻关系
中文名	离婚
法定条件	感情确已破裂
外文名:	divorce

Figure 4.1: The merged information for the entry 离婚 extracted from Baidu, Hudong and Chinese Wikipedia and offered by Zhishi.me.

Resource / Feature	Headword	Morphosyntactic and phonetic description	Alternative Forms and/or scripts	Definitions or notes to definitions	Sense annotations (e.g. register, connotation, semantic selection, etc.)	Examples	Translations (of headwords and/or examples)	Multi-word expressions, idioms, related to headword	Semantic relations	Hierarchical structures	Structures above the lexical entry level	Cross-references
Apertium	Yes	Low	No	No	No	No	Yes	No	No	No	No	No
Rayfield	Yes	High	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
TermineSP	Yes	No	No	Yes	Yes	No	Yes	No	No	No	No	No
Termcat	Yes	Medium	No	No	Yes	No	Yes	No	Yes	No	No	No
K Dictionaries	Yes	High	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Zhishi.me	Yes	No	Yes	Yes	No	No	Yes	No	No	No	No	No

Table 4.4: Summary of the representation requirements for each lexicographic resource

Chapter 5

The Representation of Lexicographic Resources as LLD: Modelling Solutions and Practices

The second step in the generation of LLD according to the methodology proposed by Gómez-Pérez et al. (2013) consists in the definition of a model to account for the representation requirements detected in the first step, which was focused on the analysis of the source data and source data model (Chapter 4). The present chapter proposes modelling solutions, application profiles of OntoLex-*lemon*, and practices to encode a range of structured data provided in lexicographic resources as RDF. Following the same structure outlined in Chapter 4, in this chapter we resume the conversion to RDF of a series of representative lexicographic resources and offer mechanisms to meet their representation needs.

Some of the solutions proposed in this chapter involve the creation of an *ad hoc* vocabulary for the specific resource. In Chapter 6 we motivate and lead the development of a module for lexicography under the umbrella of the OntoLex community group, based on some of the challenges identified in the previous chapter and the solutions we propose in this one. This frames **both Chapter 5 and Chapter 6 as contributions to advance in the second phase of LLD generation: modelling and URI naming strategy definition.**

5.1 Methodology for model definitions

The steps that we identified and followed in the definition of models to cover the representation needs of the six resources described in the previous chapter are listed below:

1. Identification and **selection of available vocabularies** to reuse (e.g. OntoLex-*lemon*, its modules, linguistic category registries such as LexInfo, etc.). The criterion guiding this selection is the **widespread use** of the vocabulary in the LLD community and their recommendation according to the BPMLOD community group.
2. **Definition of (equivalence) mappings** from the source data model to the selected vocabularies. By *mapping*, we adhere to the definition of *ontology mapping* given by Su (2002), as cited by Ramos and Gómez-Pérez (2008): “Given two ontologies A and B, mapping one ontology with another means that for each concept (node) in ontology A, we try to find a corresponding concept (node), which has the same or similar semantics in ontology B and vice versa.”
3. Analysis of these vocabularies in terms of **compatibility with the data model** of the resource by focusing on potential mappings, and conceptualisation and explanation mismatches between the two models (Visser et al., 1997). Some of the questions to address here were, among others: *when two elements share a label, do they share the concept they represent and the definition?*, *are there any concepts covered in both models sharing the definition but whose labels are not identical?*, *are shared concepts defined at the same level of abstraction?*, *are the attributes (or properties) applicable to the same shared elements?*, *if two concepts are shared, do they also share their scope?*, etc. The **scenarios** identified after this manual analysis were the following:
 - a. **Shared concepts and definitions in both source model and RDF model**. Labels may not be identical but the correspondence has been manually identified.
 - b. Shared concept and label, but **difference in class hierarchy**, or shared concept/property and label but **different types of attributes**/domain and ranges.

- c. Concept **non-existent** in RDF model.
- 4. If the available vocabularies are not exhaustive to cover the representation needs of the resource as the source data model does (cases b and c), **development of a custom vocabulary** to which to map the problematic entities from the source data model.
- 5. Development of an **application profile of the model(s) selected in Step (1)** that reuses entities from external linguistic data categories registries and that also includes the vocabulary created *ad hoc* (if Step (4) was applicable).

Alongside the model definition comes the URI naming strategy definition. A URI naming strategy proposes a pattern for the definition of the URIs of the elements that will instantiate the proposed model. For each converted resource, we provide a URI naming strategy table that aims at preventing collisions across elements and has a series of implications for the linking of resources.

5.2 The modelling of the Apertium series as RDF

Apertium RDF instantiates the *lemon* translation module.¹⁴⁴ Since this module would inspire the Variation and Translation module of OntoLex-*lemon* years later (*vartrans*), we revisit its main components before introducing Apertium RDF in the later subsection.

5.2.1 The *lemon* translation module

The *lemon* translation module (Gracia et al., 2014a) is an extension of *lemon* (see Section 2.3.2) to explicitly represent translations as relations across `lemon:LexicalSense` elements. Figure 5.1 summarises the classes and properties of the module. Given two instances of `lemon:LexicalSense`, the reified relation `Translation` takes one as source and one as target to encode the directionality of the translation. There is a small vocabulary of translation categories that can be used to further specify the type of translation.¹⁴⁵ This serves to distinguish senses that are semantically equivalent from those that are only pragmatically equivalent or those that are semantically equivalent

¹⁴⁴<http://linguistic.linkeddata.es/def/translation-content/index.html>

¹⁴⁵<http://purl.org/net/translation-categories>

and also loan translations. **Translations**, in turn, are grouped in translation sets (**TranslationSet**) extracted from a specific resource.

Figure 5.1: The *lemon* Translation Module¹⁴⁶

5.2.2 The Apertium series as RDF

In the previous chapter we introduced the Apertium family of bilingual dictionaries (Forcada et al., 2011). The challenging aspect in the representation of the Apertium series as RDF was essentially that **translations were explicitly encoded in the data as relations between senses**. Prior to the *OntoLex-lemon* model, *lemon* did not include a mechanism to represent translations explicitly, which required a new module geared specifically at translation representation – the *lemon* translation module explained above. The modelling of the Apertium RDF dictionaries was addressed in Gracia et al. (2018). **The solutions and practices regarding translations offered for the five remaining resources of this chapter are based on the Apertium RDF application profile of *lemon*.**

Example 4.1 presented in Chapter 4 concerned the French entry *rouge*. The entry included data about its part-of-speech, and its translation to Spanish *rojo*, which was

¹⁴⁶Figure extracted from the module documentation available at <http://linguistic.linkeddata.es/def/translation-content/index.html>.

encoded with a **SenseAxis** element. Figure 5.2, following the approach proposed by Gracia et al. (2018), represents how this information was mapped to the *lemon* model and the *lemon* translation module.

Figure 5.2: RDF representation of the entry *rouge* from Apertium FR-ES RDF

In LMF Apertium source data, entries in different languages belong to different **Lexicon** elements, as the path **Lexicon/LexicalEntry** indicates in the source files. Following this same configuration, Figure 5.2 shows that LMF entries are directly mapped in RDF to **lemon:LexicalEntry** elements, each of them belonging a different **lemon:Lexicon**. This correspondence was straightforward in this case because the *lemon* model inherits some classes and properties from LMF, with **LexicalEntry** and **Lexicon** being two of them. In addition, a LMF sense is mapped to a **lemon:LexicalSense**, which, however, does not have an ontological reference in this specific application profile of *lemon*. As the focus in this effort was the translation relations across senses from a given resource, ontological references remained unspeci-

fied. A `lemon:Form` represents a realisation of a lexical entry, and lemmas are mapped to the written representation of such forms. The values provided in the LMF data had a correspondence with LexInfo’s `lexinfo:PartOfSpeech` individuals (e.g. `lexinfo:adjective`, `lexinfo:noun`, etc.), and thus they are reused to capture part-of-speech data. Figure 5.2 shows that the data which previously was structured with `SenseAxis` elements, the translation relation itself, is now captured through the use of the *lemon* translation module. As the French sense was the source of the translation (or the left side of the transducer in the original Apertium apparatus, marked in the data with *-l* in the sense IDs) and the Spanish sense the target (or the output of the transducer, marked with *-r*), the `tr:Translation` relation respects this directionality through the use of the `tr:translationSource` and `tr:translationTarget` properties. By reifying the translation relation between those senses, as opposed to using another mechanism (e.g. the property `lexinfo:translation`), any other information regarding that specific translation can be added, for instance, the type of translation or its source.

On the left side of Figure 5.2 there are two sets of resources, each linked to either lexical entry. These represent the original Apertium files from which the data was extracted. Given a file, by gathering those entries in the target language and adding them to a lexicon in that same language (in this example, Spanish), the lexicon dynamically grows as more resources are converted to RDF. This same result is achieved both with those original dictionaries in which Spanish is the source language as with those in which it is the target. For this reason, the sources of a lexical entry are various: all those original LMF files in which the entry occurs either as source or as target in a translation. In this figure, note that there is a resource with ID 17090, which acts as both source of *rouge* and *rojo*, i.e, this resource is the original Apertium LMF FR-ES lexicon.

Table 5.1 details the URIs of the elements that instantiate the *lemon* and translation module classes, e.g. `:lexiconFR/rouge-adj-fr` as an instance of `lemon:LexicalEntry`. The table shows that in Apertium there are different patterns, depending on the element.

The **advantages of this URI naming strategy** are two:

1. The URIs are **transparent**, which allows humans an easier access and navigation through the RDF source to query and/or validate.

Element	Pattern	Example
Lexical Entry	<Lexicon>/<lemma>-<part of speech> -<language tag>	:lexiconFR/ rouge-adj-fr
Form	<Lexicon>/<lemma>-<part of speech> -<language tag>-form	:lexiconFR/ rouge-adj-fr-form
Lexical Sense (source)	<TranslationSet>/<lemma of source entry> _<lemma of target entry>-<part of speech>-<language tag>-sense	:tranSetFR-ES/ rouge_rojo-adj-fr-sense
Lexical Sense (target)	<TranslationSet>/<lemma of target entry> _<lemma of source entry>-<part of speech>-<language tag>-sense	:tranSetFR-ES/ rojo_rouge-adj-es-sense
Translation	<TranslationSet>/<URI of source sense> -<URI of target sense without the tranSet path>-trans	:tranSetFR-ES/ rouge_rojo-adj-fr-sense-rojo_ rouge-adj-es-sense-trans

Table 5.1: Apertium RDF URI naming strategy

- By using the lemma of the entry (and its part-of-speech and language code) instead of relying on internal identifiers, **the lexica in different languages are dynamically augmented** with all entries extracted in that language. This means that there is no difference between the entry **rouge-adj-fr** extracted from the ES-FR lexicon and the entry **rouge-adj-fr** extracted from FR-EN: they are the same resource. This linking or, more specifically, *merge*, at the level of the lexical entry is thus caused by the use of this particular URI naming strategy. As a **disadvantage**, this approach **does not consider homonymous entries**, since it solely relies on lemma, language and part-of-speech tag in order to create a resource.

Apertium RDF provides the grounds for the application profiles of OntoLex-*lemon* and the extensions to it we suggest in this thesis. More specifically, we adopt the configuration of dynamically growing monolingual *lemon* lexica relating to each other through translation relations in all of the dictionaries mentioned in the previous chapter. For the advantages mentioned above, Apertium RDF’s URI naming strategy is adopted as well with the remaining resources, with minor variations depending on the dictionary or terminology. For each resource, we provide a table detailing the URI naming strategy and its deviations from the original one suggested for Apertium RDF.

5.3 The modelling of the Rayfield Comprehensive Georgian-English Dictionary

Approaching the modelling of Rayfield’s Comprehensive Georgian-English Dictionary involved moving from machine-readable resources (e.g. Apertium, Termcat) to a resource inheriting features from paper lexicography and exclusively oriented towards human users. For example, this dictionary resorts to punctuation marks to encode both morphological and orthographic information on the assumption that the user will interpret this encoding, described in the editor’s preface, along with the used list of symbols and abbreviations. Since the only mark-up provided in the original source file (text) was italics, each annotation marked in italics was interpreted according to the list of abbreviations and editor’s preface, together with the assistance of Jesse Wichers Schreur, a linguist expert in Georgian linguistics.

Since we adhere to the OntoLex-*lemon* model for the conversion of the data, section 5.3.1 provides an overview of its *vartrans* module to capture translations. These module will be used in the remaining resources discussed in this chapter. Following this section, we describe the modelling of Rayfield’s dictionary for two layers: the monolingual and the multilingual.

5.3.1 The OntoLex-*lemon* vartrans module

As mentioned in Chapter 2, OntoLex-*lemon* is the resulting work of the continued efforts made by the W3C Ontology Lexica Community group during the past three years to build a rich model to represent the lexicon-ontology interface. It is largely based on the *lemon* model and, along with the extensions to it, integrates work of the various Community members.

The OntoLex *vartrans* module, shown in Figure 5.3, was developed to record variation relations across entries in the same or different languages. The intuition behind it is to capture two kinds of relations: those across senses and those across lexical entries. Variation relations across senses are of semantic nature and include terminological relations and translation relations. In contrast, relations across lexical entries can also

¹⁴⁷Figure extracted from the Final Model Specification of OntoLex (Lexicon Model for Ontologies: Community Report, 10 May 2016) <https://www.w3.org/2016/05/ontolex/#variation-translation-vartrans>.

Figure 5.3: The OntoLex-lemon vartrans module¹⁴⁷

encode lexical variation concerned with the surface form of a term. In the case of Rayfield’s dictionary, Termesp RDF and K Dictionaries RDF, we focus on the first kind of relations, variation relations across senses, to represent translations. In Termcat RDF, we also turn to LexInfo to represent lexical relations between an initialism, an acronym or an abbreviation and the entries that are their expanded form, while also capturing this information with `ontolex:Form` elements.

The *vartrans* module frames translation relations as a special type of lexico-semantic variation across the entries of different lexica, more specifically, as relations that hold across the senses of those entries. The translation component goes back to the *lemon* translation module. In this apparatus, the *vartrans* module uses a pivot class `vartrans:Translation` to represent translations among lexical entries as relations across `ontolex:LexicalSense` elements that can point to the same ontology concept (although not necessarily). One of the main differences between the *vartrans* and the *lemon* translation modules is the conception of a class to encompass variation relations, the `vartrans:LexicoSemanticRelation` class. Its sub-classes denote the two types relations mentioned above, those that hold across lexical entries (`vartrans:LexicalRelation`) and those that relate lexical senses (`vartrans:SenseRelation`). The pivot class `var-`

`trans:Translation` is thus a `vartrans:SenseRelation`, and so are the relations across terminological variants (`vartrans:TerminologicalVariant`). The resources we address in this work contain translation relations of the `directEquivalent` category, but other translation categories (e.g. `culturalEquivalent`, for culture-dependent concepts; `lexicalEquivalent`, for literal translations of the source term, etc.) are supported as well and can be included from the same external ontology used with the *lemon* translation module. In addition to translation relations across their senses, lexical entries in different languages can be directly related through the property `vartrans:translatableAs`.

```
@prefix ontollex: <http://www.w3.org/ns/lemon/ontollex#> .
@prefix vartrans: <http://www.w3.org/ns/lemon/vartrans#> .
@prefix dct: <http://purl.org/dc/terms/>.

:tuberculosis a ontollex:LexicalEntry ;
  ontollex:lexicalForm :tuberculosis_form ;
  ontollex:sense :tuberculosis_sense.
:tuberculosis_form ontollex:writtenRep "tuberculosis"@en .
:tuberculosis_sense ontollex:reference <http://dbpedia.org/resource/
  Tuberculosis>.

:phthisis a ontollex:LexicalEntry ;
  ontollex:lexicalForm :phthisis_form ;
  ontollex:sense :phthisis_sense.
:phthisis_form ontollex:writtenRep "phthisis"@en .
:phthisis_sense ontollex:reference <http://dbpedia.org/resource/Tuberculosis>;
  dct:subject <http://dbpedia/resource/Medicine> .

:phthisis_diachronic_var a vartrans:TerminologicalVariant;
  vartrans:source :phthisis_sense ;
  vartrans:target :tuberculosis_sense ;
  vartrans:category isocat:diachronic.
```

Example 5.1: Diachronic variation between the terms *phthisis* and *tuberculosis* represented with the *vartrans* module

With respect to terminological variants, the *vartrans* module allows for the encoding of dialectal (diatopic), register (diaphasic), chronological (diachronic), stylistical (dias-
tratic) as well as dimensional variation across entries. In this way and in the same fashion as translations, two lexical entries are linked to their respective lexical senses, and these are related through the pivot class `vartrans:TerminologicalVariant`, with the

property `vartrans:category` allowing for the specific type of terminological variation under analysis to be included as well. For cases in which the directionality of the translation is not considered, that is, which term is source or target in the translation is not relevant, the property `vartrans:relates` (similar to the former `tr:translationSense` in *lemon*) links the two senses to the element acting as pivot. Example 5.1¹⁴⁸ shows the fact that there has been a diachronic variation between the terms *phthisis* and *tuberculosis*, being the latter one the one used nowadays. This shift is captured by representing the two senses as source and target respectively.

5.3.2 Modelling monolingual information in a nominal entry of Rayfield’s dictionary

The monolingual information addressed in this prototype of Rayfield’s Georgian-English dictionary concerns the representation of the **lemma, its part-of-speech, different forms, domain, temporal qualifiers, register, connotations, grammatical number, dialect and multi-word expressions** in a way that distinguishes each piece of information in the RDF graph, and reuses as much as possible already existent entities from external vocabularies created to encode these data (see Section 2.6). In the following, we elaborate on each of these aspects and its representation with RDF mechanisms.

Punctuation and orthography. In order to achieve a correct mapping from the original text format to the selected RDF models, the first step was to ensure that the distinctions encoded with punctuation marks, which are highly dependant on the features of Georgian, were interpreted and reflected in the outcome of the conversion. This also involved recording at the same time the original source data as presented in the dictionary for traceability and future improvement purposes.

After a thorough analysis of nominal entries, we detected that lemmas in this dictionary occur with different types of patterns:

1. A single lemma, without any punctuation mark

¹⁴⁸Example extracted from the Lexicon Model for Ontologies: Community Report, 10 May 2016. <https://www.w3.org/2016/05/ontolex/#variation-translation-vartrans>. Last accessed 04.09.2019.

2. A single lemma, with parentheses
3. A single lemma, with brackets
4. A single lemma followed by different wordforms in parentheses. However, this scenario is not present in the case of nouns and is thus left for future work on this conversion.
5. A single lemma followed by a case ending in parentheses
6. Two lemmas concatenated by commas, with each one being any option from 1-3 and 5.

Example 5.2 shows an example of each case following the order presented above. Figures 5.4 – 5.8 depict the RDF modelling of five of these cases, as case (4) was not attested in nouns. In case 1 (Figure 5.4), we treat the lemma as the written representation of a single `ontolex:Form`, following OntoLex. In contrast, the parentheses of case 2 (Figure 5.5) indicate that a lemma can occur with or without the characters in parentheses. OntoLex specifies that orthographic distinctions, as opposed to grammatical distinctions, are recorded by instantiating more than one `ontolex:writtenRep` value, whereas grammatical distinctions would lead to several `ontolex:Form` elements. Even though case 2 (and case 6, see below in Figure 5.8) seem to just encode orthographic variants, the dictionary also presents cases in which one of these variants receives an annotation as preferred or not (marked with `sst`, if the short-stemmed variant is the most used variant; or `lst`, for *long stem*, if the long-stemmed variant is the one preferred). Modelling each variant as an `ontolex:Form` would allow to later specify which one is the canonical one. If modelled as two different values of `ontolex:writtenRep` associated with a single form, a way to capture this distinction would be to turn to a second property `skos:prefLabel` to record the preferred option at the `ontolex:LexicalEntry` or `ontolex:Form` level. However, creating two `ontolex:Form` instances leaves the door open to add any other descriptive features of that specific variant (e.g. whether it is an archaic orthographic variant, for instance) at the cost of deviating from the OntoLex-*lemon* core.

Case 3, depicted in Figure 5.6, exemplifies the RDF modelling of a lemma with characters in brackets. Brackets indicate a vowel which is normally elided when case or plural suffixes begin with a vowel. Since in these entries we lack the inflected forms in

- (1) აბაჟური <i>n</i> lampshade
- (2) ავეცი(ა)ობა <i>n</i> spiteful language, nasty remark
- (3) აბჯროს[ა]ნი <i>n</i> man wearing armour
- (4) აზვავებს (დააზვავა, დაუზვავებია) sends sth (<i>pile of earth/snow/hay etc</i>) falling/sliding/scattering
- (5) აზრთელაპია (-ას) <i>n</i> censurer, caviller; wet-blanket
- (6) აკუმერი, აკუმორი <i>n</i> obstetrician

Example 5.2: Six different examples of lemma patterns across noun entries in Rayfield's Georgian-English Bilingual dictionary

Case 1

Figure 5.4: The modelling of entries without parentheses in the lemma (case 1)

which that vowel is omitted, the lemma of the entry (without the marked vowel) is taken to be the sole written representation provided in the entry. Similarly, case 5, depicted in Figure 5.7, provides a case ending in parentheses, but not the full inflected form. The lemma of the entry is interpreted to be the written representation without the case ending information, which is nonetheless recorded as label of the entry for traceability purposes.

Lastly, Figure 5.8 shows the modelling of entries that are headed by two lemmas concatenated by comma. The approach followed here is the same as the one followed for case 2, grounded on the same reasons.

Figure 5.5: The modelling of entries with parentheses in the lemma (case 2)

Number and part-of-speech. This information was straightforward to model according to the *OntoLex-lemon* specification. As regards part-of-speech, this dictionary includes both common (*n.*) as well as proper nouns (*N.*), which is distinguished with the use of `lexinfo:noun` and `lexinfo:properNoun` at the level of the `ontolex:LexicalEntry`. In addition, some nouns are further described as deverbal nouns (*vn*), and hence they have been represented as instances of `http://purl.org/olia/olia.owl#VerbalNoun` with part-of-speech `lexinfo:noun`. In some cases, nominal entries denote a collective or are used in plural, which in the original resource is indicated with the abbreviation *ct* and *plu*, respectively. We account for this information with the use of `lexinfo:plural` and `lexinfo:collective` at the `ontolex:Form` level. The entry shown in Example 5.3 shows this annotations, along with information on register, temporal qualifiers and dialect (see below).

Register, temporal qualifiers and dialect. Example 5.3 shows the entry *st'ilni*, with a single sense with translations into English ‘teddy-boys’ and ‘dandies’ (multiple senses are marked with semi-colon, whereas translation equivalents for a single senses are divided by commas, see Section 4.1.2), referring to members of the British sub-culture

Case 3

Figure 5.6: The modelling of entries with brackets in the lemma (case 3)

Case 5

Figure 5.7: The modelling of entries with case endings in the lemma (case 5)

from the 50s. This entry includes annotations indicating, in order, part-of-speech (*n*), number (plural, *plu*), dialect or transliterations from a specific language (Russian, *Rus*), register (colloquial, ©) and a temporal qualifier that specifies the term is an obsolete form (†).

The information on **dialects** provided in Rayfield's dictionary covers Fereidan (*Fe*),

Case 6

Figure 5.8: The modelling of two lemmas concatenated by comma in the dictionary entry as two alternative `ontolex:Form` elements

(7) სტილნი <i>n plu Rus © †</i> teddy-boys, dandies

Example 5.3: An example of an entry used in plural, with register, temporal qualifiers and dialect/language annotations

Gudamaqruli (*Gd*), Gurian (*Gu*), Ingolo (*Ig*), Imeretian (*Im*), Javakhetian (*Ja*), Judaic (Kutaisi) (*Ju*), Kartlian (*Ka*), Kiziqi (*Ki*), Kakhetian (*Kk*), Khevsur (*Kv*), Lechkxumi (*Le*), Meskhian (*Me*), Mokhevi (*Mo*), Mtiuli (*Mt*), Rachian (*Ra*), Tianeti (and Ertso) (*Ti*) and Tush (*Tu*). The only transliterated language is Russian (*Rus*). Dialect information in *OntoLex-lemon* is indicated with a language tag at the written representation level, e.g. `@en-GB` for British English, following the Best Current Practice 47¹⁴⁹, which specifies the use of language tags and subtags. While some of these annotations have a tag counterpart (e.g. Judeo-Georgian, `jge`; Russian, `ru`), most of these dialects are

¹⁴⁹<http://www.rfc-editor.org/rfc/bcp/bcp47.txt>

not included in the Internet Assigned Numbers Authority (IANA) subtag registry.¹⁵⁰ Glottolog (see Section 2.6) does provide a unique and internal code for some of them, e.g. `guri1246` for the Gurian dialect) but the list is not exhaustive either. Due to these limitations, this information has been recorded by a generic property, `rdfs:comment`, for later processing following recent proposals to encode language and dialect information of lexical entries (Gillis-Webber et al., 2019a,b).

Register annotations include markers for colloquial and slang use, as well as poetic, vulgar and obscene. Whenever the individuals provided in Lexinfo matched (e.g. `lexinfo:slangRegister` as mapping from the tag ©), we represented this information by reusing `lexinfo:register` and the appropriate individual at the `ontolex:LexicalSense` level. Notice that this annotation is not specifically linked to the sense in the original entry (it could be linked to the lexical entry as a whole), but we consider register annotations to describe the context in which a word is used in a specific sense, and hence we encode this at the `ontolex:LexicalSense` level. In those cases in which a register in the original dictionary did not match any individual in LexInfo, we instantiate a blank node with the label provided in the dictionary as description of the abbreviation used. This is shown in Example 5.4, which is the RDF rendering of the entry *goddess*, whose first sense is attested in the poetic register.

The entry for *dandies* mentioned above includes a temporal qualifier that specifies the term is an obsolete form (†). So is the case of the entry for *goddess*, too, as shown in Example 5.4. This information is captured by characterising the entry as a `lexinfo:obsoleteForm`, with the property `lexinfo:temporalQualifier`.

Connotations. In addition to register, the dictionary provides information on derogatory interpretations, e.g. with the word *busurmani* ‘infidel’, or the word *int’eligent’una* ‘wretched intellectual’. LexInfo does not contain an individual to encode this information, but the OLiA Reference Model (see Section 2.6) includes the class `http://purl.org/olia/olia.owl#PejorativeEvaluative`, which we have reused as object of the property `ontolex:usage`. This property serves to capture pragmatic conditions or meaning nuances, and the OLiA class is the closest match to the original annotation (*dg*, derogatory) available in the vocabularies that were analysed during this work.

¹⁵⁰<https://www.iana.org/assignments/language-subtag-registry/language-subtag-registry>

```

@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix lime: <http://www.w3.org/ns/lemon/lime#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#> .
@prefix lexinfo: <http://www.lexinfo.net/ontology/2.0/lexinfo#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .

:lexiconKAT/ღმერთა-ნ a ontolex:LexicalEntry .

:lexiconKAT/ღმერთა-ნ-form a ontolex:Form ;
  ontolex:writtenRep "ღმერთა"@kat .

:lexiconKAT/ღმერთა-ნ ontolex:lexicalForm :lexiconKAT/
  ღმერთა-ნ-form .

:lexiconKAT/ღმერთა-ნ-goddess-sense a ontolex:LexicalSense.

:lexiconKAT/ღმერთა-ნ-goddess-sense-lc a ontolex:LexicalConcept .

:lexiconKAT/ღმერთა-ნ-goddess-sense ontolex:isLexicalizedSenseOf
:lexiconKAT/ღმერთა-ნ-goddess-sense-lc ;
  lexinfo:register [rdfs:label "bookish, artificial, poetic register"@en ]
  .

:lexiconKAT/ღმერთა-ნ ontolex:sense :lexiconKAT/
  ღმერთა-ნ-goddess-sense ;
  lexinfo:partOfSpeech lexinfo:noun ;
  rdfs:label "ღმერთა -ახ()"@kat ;
  lexinfo:temporalQualifier lexinfo:obsoleteForm .

```

Example 5.4: Extract of the RDF Turtle rendering of the entry *goddess*

Multi-word expressions. Some entries embed a lexicalized multi-word expression that is composed of the lemma of the dictionary entry in which they are defined. For example, the entry *Adam* provides the fixed expression *of/since ancient times*, which, interestingly, is composed by the entry *Adam* (and *time*), as shown in Example 5.5. Figure 5.9 depicts the RDF representation of such a multi-word expression. Since these fixed expressions are lexicalized in the language, we treat them in the same fashion as a simple `ontolex:LexicalEntry` and associate to them a `ontolex:LexicalSense` (which in turn is linked to a `ontolex:LexicalConcept` encoding the actual meaning of the word) and an `ontolex:Form`, following *OntoLex*. However, we further mark that

these new `ontolex:LexicalEntry` elements, which were previously embedded in the dictionary entry, have as component the entry where they were originally defined. This statement is recorded through the property `decomp:subterm`, in this specific example, linking *since ancient times* as `ontolex:LexicalEntry` to the entry of *Adam*.

(8) ადამი <i>N</i> Adam; ადამის შვილი son of Adam, human being; ადამის უამისა of/since ancient times; ადამის ხნისა very old, ancient

Example 5.5: The entry *Adam*, which provides a fixed expression, *of/since ancient times*.

Figure 5.9: RDF representation of the entry *Adam* shown in Example 5.5

5.3.3 Modelling multilingual information in a nominal entry of Rayfield’s dictionary

Each entry of Rayfield’s dictionary provides either a single sense, or a series of senses separated by semi-colon. For each sense, in turn, there can be one or several translations separated by comma. A clear example of this is entry #5 listed in Example 5.2, given here below. For lack of numerical identifiers, we distinguish the two concepts the entry denotes as the *censurer*-concept and the *wet-blanket*-concept.

(5) აზროყლაპია (-ას) *<i>n</i>* censurer, caviller; wet-blanket

Example 5.10: An example of a polysemous noun with a sense with multiple translations

Translations are modelled as RDF using two strategies simultaneously: the *vartrans* approach, which makes the translation relation between the Georgian sense and the English sense explicit; and the cross-lingual synonymy approach, in which an *ontolex:LexicalConcept* is lexicalised by senses from entries in different languages. Figure 5.10 shows the rendering of Example 5.10 as RDF following these two approaches (the *vartrans:Translation* involving the sense of entry *cavillier*, as well as the RDF elements encoding the *wet-blanket*-concept, have been omitted for space reasons, but follow the same pattern as the one laid out for the first sense, *censurer*).

The table represented in Figure 5.11 shows the URI naming strategy followed for the conversion of this dictionary. Since it is based on Apertium’s URI naming strategy, the table only shows those elements that deviate from Apertium. Note that the URI pattern reflects whether the lemma included punctuation marks in the original resource, and that the different forms explained earlier in this section (cases 1-6) would be distinguished with different written representations in the URI. Multiword expressions are modelled as simple lexical entries and hence the only difference in their URI is the escape of the space character, and the lack of a part-of-speech, as this information is not given in the entry.

Figure 5.10: RDF representation of a portion of Example 5.10 as RDF

Element	Pattern	Example
LexicalEntry	<Lexicon>/<lemma>-<part of speech>	:lexiconKAT/აღლებადობა-n
LexicalEntry	<Lexicon>/<lemma with brackets or parentheses replaced by underscores>-<part of speech>	:lexiconKAT/ავცი_ბ_ობა-n
LexicalEntry	<Lexicon>/<multiword expression>	:lexiconKAT/აღამინ+შვილი
Form	<Lexicon>/<lemma>-<part of speech>-<possible writtenReps>-form	:lexiconKAT/აღაბხანა-n-აღაფხანა-form
LexicalSense	<Lexicon>/<lemma>-(<part of speech>-<concept identifier>-sense	:lexiconKAT/ავღარი-n-bad+weather-sense
LexicalConcept	<Lexicon>/<lemma>-<part of speech>-<concept identifier>-lc	:lexiconKAT/ავზორცობა-n-lust-sense-lc

Figure 5.11: URI naming strategy for Rayfield's Georgian-English Dictionary (showing only those elements that deviate from Apertium)

5.4 The modelling of Terminesp and Termcat: Terminoteca RDF

Chapter 4 provides an overview of the challenges in the modelling of two terminological resources, Terminesp and Termcat (specifically, three domain lexica from the latter one). These resources were deemed as representative because of their multilingual nature and, altogether, they gather many of the annotations commonly found in domain lexica, from scientific denominations to translations, definitions, terminological norms, semantic relations and abbreviations.

The conversion of Terminesp to RDF had already been addressed in the literature (Gracia et al., 2014b). In that effort, Terminesp data served as a validating example of a terminological resource to which the *lemon* translation module would apply. In the following section, we provide new modelling solutions based on that work to account for the features of this resource by reusing OntoLex-*lemon*. This work is presented in Bosque-Gil et al. (2015) and serves as validating example of the *vartrans* module that had recently been issued, together with the final model specification, in the same fashion that Apertium RDF served as a validating example of the *lemon* translation module. The conversion of Termcat was guided by the goal to create a prototype of an ecosystem of linked terminologies in Spain, named Terminoteca RDF (Bosque-Gil et al., 2016c). This ecosystem involved the linking of Termcat with Terminesp RDF, converted previously. With this goal in mind, our conversion of Terminesp was updated to follow the same strategy and configuration to create a set of homogeneous resources under the Terminoteca RDF namespace.

Both resources contained in Terminoteca RDF were modelled according to OntoLex-*lemon* and reuse SKOS (see Section 2.3.2). This decision was motivated by the following reasons:

1. The creation of an **ecosystem of linguistic resources sharing vocabularies** on which to test the hypotheses formulated in this thesis (see Section 3.3).
2. Interoperability with **other terminological language resources** published as linked data (Cimiano et al., 2015a) and with other lexicographic resources already available as *lemon* datasets (e.g. the Apertium RDF family) so as to enrich them with information coming from external sources in future steps.

3. The advantages of OntoLex-*lemon* in contrast to the sole use of SKOS. *lemon* subsumes most features of SKOS (e.g. preferred and alternative labels) and SKOS-XL reification of labels, but it also offers the advantages of the *vartrans* module to represent term variants and translations: the representation of translations as relations between lexical senses and not between labels, which we deem more accurate linguistically.

The following sections elaborate on the conversion of both resources, Terminesp and Termcat, to RDF, with special attention to their modelling and to their linking under the Terminoteca RDF namespace.

5.4.1 The representation of Terminesp as RDF

Terminesp contains a range of terms from specialised domains extracted from documents issued by the Spanish Association for Terminology. The entries contain definitions, notes to the definitions, norms and translations from Spanish to a variety of languages, along with scientific denominations in Latin. Some definitions and definition notes in these languages are included as well. In addition, the entries themselves range from simple nouns and adjectives to complex nominal, prepositional and adjectival phrases.

The entries of Terminesp raised modelling doubts so as to how exactly we could model definitions and definition notes in a way that reflects that the note string is related to the definition, but distinguishes it from the definition string itself, as well as how to record a terminological norm (norm codes) or standard and its entry in the most reusable way across terminologies. Furthermore, modelling scientific denominations posed problems as regards their treatment as terminological variants, translations or synonyms, for example.

We also renew the previous work with Terminesp (Gracia et al., 2014b) with the following goals in mind:

1. Detect errors and inconsistencies in the data before linking the dataset to other language resources (i.e. LexInfo).
2. Provide a validating example of the OntoLex-*lemon vartrans* module to account for variation across entries, with emphasis on scientific naming.
3. Extend the LD resource with definitions, norms, and part-of-speech categories.

4. Create a database of nominal, prepositional, and adjectival phrases lexicalized in a domain, and thus highly specialized, to be used by NLP applications.

The example provided in Table 4.2 in Chapter 4, for the record *admitancia cinética* ‘motional admittance’ has been mapped to its OntoLex-lemon version in Figure 5.12. The figure shows the RDF representation for the monolingual information contained in this record (for part-of speech-information, see the later content on this same section), while the translations, and multilingual information in general, are modelled in RDF in Figure 5.13.

Figure 5.12: RDF representation of the monolingual information of Terminesp entry *admitancia cinética* ‘motional admittance’ shown in Table 4.2

As in the previous version of Terminesp by Gracia et al. (2014b), Figure 5.12 shows that we instantiate a `skos:Concept` for any given Terminesp entry in order to ground the terms conceptually. These `skos:Concept` elements are created artificially by reusing the ID of the terminological record. A Terminesp entry, in turn, is modelled as an `ontolex:LexicalEntry` whose `ontolex:LexicalSense` points to the `skos:Concept` identified by the terminological record where the entry is defined. Definitions, notes, norm

Figure 5.13: RDF representation of the multilingual information of the Termesp entry *admitancia cinética* ‘motional admittance’

codes, norm titles and provenance were added as linked data as well. Specifically, definitions are attached to the `skos:Concept` to which the `ontolex:LexicalSense` refers. This is done through the property `skos:definition`. Moreover, some of these definitions include a note that provides additional information about the definition content, use cases, etc. In order to distinguish this from the definition itself, we use `rdfs:comment` to relate the `skos:Concept` to the string acting as note.

In *lemon*, the class `lemon:SenseDefinition` allowed to treat the definition of a term as an object whose property `lemon:value` pointed to the definition string itself. This was particularly well-suited for capturing elements that were not definitions but which were nonetheless related to them, as in the case of notes to definitions. The class `lemon:SenseDefinition` is not included in OntoLex. In fact, definitions are not encoded at the sense, but at the `ontolex:LexicalConcept` level with the property

skos:definition. The W3C community group of OntoLex recently agreed on the use of the class **ontolex:LexicalConcept** to provide the definitions given in dictionary entries as part of the development of the *lexicog* module (see Chapter 6).¹⁵¹ Since **ontolex:LexicalConcept** is a subclass of **skos:Concept**, our proposed modelling of definitions as strings attached to **skos:Concept** elements is compatible with the best practice decided within the community group.

Terminesp entries are all extracted from Spanish UNE documents and, in addition to definitions, the database provides norms and norm code information. The norm each entry comes from is thus captured by **dc:title**, and the norm code included with an instantiation of the **dc:source** property.

Figure 5.13 shows the representation as RDF of the multilingual data in Terminesp entries. Translations are included by instantiating a **vartrans:Translation** element with the properties **source** and **target** linking the two translation senses, one for each language. The senses are attached to the **vartrans:Translation** by an additional property too, **vartrans:relates**. Note that we have followed the same approach adopted in Apertium RDF, the previous *lemon* version of Terminesp, and the conversion of Rayfield’s Georgian-English dictionary as regards the extraction of translation targets and their representation as entries of different **lime:Lexicon** elements. These lexica are independent of each other and the relations across them occur at the sense level through translation relations (or terminological relations, in the case of Latin). The treatment of those translation targets follow the same pattern as in the case of the Spanish source term: the entries belong to a **lime:Lexicon** and are attached a lexical form with the information about the lemma and a sense. Since they are translations, we take as their denotation (or reference for the sense) the same concept of the Spanish term.

Section 5.4.2 dwells on the structure of the Terminoteca RDF in detail, the resource that gathers together Terminesp RDF and Termcat RDF. In Terminoteca RDF, lexica in the same language from both resources are merged together. However, the concepts are kept separate and do not belong to a specific lexicon. Figure 5.13 shows a SKOS concept in Terminesp, which serves as pivot, and translation relations that link senses in different languages belong to different lexica, as in Apertium RDF.

¹⁵¹See *Issue 11: Dictionary Senses as Lexical Senses or as Lexical Concepts* on the OntoLex Lexicography page: <https://www.w3.org/community/ontolex/wiki/Lexicography#Issues>.

In addition to translations, definitions, notes and terminological norm data, some Termesp entries also provide scientific denominations. Figure 5.14 shows the rendering of this information as RDF.

Figure 5.14: RDF representation of an extract of the entry *anacardo* ‘cashew’, with its scientific denomination in Latin, *Anacardium occidentale* *Linnaeus*

The concerns that arise when **modelling scientific denominations** are mainly four:

1. Whether to treat them as **translations** sticking to the fact that we are connecting terms in two different languages.
2. Whether to treat them as **terminological variants** on the grounds that they are international scientific denomination. In this sense, scientific names are considered a specific type of variants subject to domain and register rather than to any other factor.

3. Whether to consider the relation between scientific and common terms as a **special case of synonymy** (Soergel et al., 2006).
4. Whether these relations could be captured by **sameAs properties** (which might be refined into `hasScientificTaxonomicName` and `isScientificTaxonomicNameOf` (Sini et al., 2009)).

In our approach we opted for option 2, inspired by previous work on this domain, particularly on the LIR (Linguistic Information Repository) model (Montiel-Ponsoda et al., 2011), we have decided to identify the Latin entry as the international scientific denomination. In this sense, scientific names are considered a specific type of terminological variants subject to domain and register (*diaphasic*). Also, they are internationally accepted over scientific communities and can appear in texts written in any language, provided that the register and the domain are adequate for their use. Taking this into account, Latin terms are modelled in Figure 5.14 as terminological variants and, more specifically, as `lexinfo:InternationalScientificTerm(-s)`. Such terminological relations are established from the Spanish lexicon to the Latin one, and not from other languages. The Latin denominations are only present for the Spanish terms in the source dataset.

It is worth mentioning that *Terminesp* is not conceived as a taxonomic terminological database and in that respect its data lack the richness we find in AGROVOC (Soergel et al., 2006). A proposal for restructuring scientific names in AGROVOC to allow for a better identification of an organism through its various denominations has been suggested (Sini et al., 2009). The authors distinguish between the different scientific names of a term and the most common accepted one of them, and also address relations between a term a taxa concept (order, family) to allow for the modeling of the taxonomy via properties such as `hasSpecies`. In contrast, our proposal deems appropriate for terminological databases that contain Latin denominations but are not aimed at representing taxonomic knowledge. As such, we focus on the intuition that a Latin term is just another lexicalization of a concept and turn to the fact that scientific denominations are used as terminological variants in different registers and domains than common terms.

The URI naming strategy in *Terminesp*, shown in Table 5.2, follows the pattern decided for Apertium RDF for lexica and lexical entries. However, *Terminesp* source

data included for each entry a numeric identifier, which served as a unique concept identifier in the terminology and is propagated throughout the RDF elements from the `skos:Concept` to the `LexicalSense`. In fact, `Termine`sp URIs were constructed with identifier preservation in mind¹⁵², e.g.: `lexiconEN/36995en-sense`. For this reason, `Termine`sp URIs inherit the original ID of the `Termine`sp record for all elements that only bear semantic information (not lexical).

Element	Pattern	Example
Lexicon	:lexicon<LANGUAGE TAG>	:lexiconES
Lexical Entry	<Lexicon>/<lemma>-<part of speech> -<language tag>	:lexiconES/lexiconES/ admitancia%2Bcin%C3%A9tica-n-es
Form	<Lexicon>/<lemma>-<part of speech> -<language tag>-form	:lexiconES/lexiconES/admitancia%2B cin%C3%A9tica-n-es-form
Lexical Sense (source)	<Lexicon>/<ID of source entry><source language tag>-sense	:lexiconES/55272es-sense
Concept	<resource from which the concept is extracted (termine	:termine
Resource representing a terminological norm	:norms/<ID of the norm from which the source term is extracted>	:norms/UNE_21302-801
Lexical Sense (target)	<Lexicon>/<ID of source entry><target language tag>-sense	:lexiconDE/55272de-sense
Translation Set	:transSet<SOURCE LANGUAGE-TARGET LANGUAGE>	:tranSetES-DE
Translation	<TranslationSet>/<resource from which the concept is extracted (termine	:tranSetES-DE/termine

Table 5.2: `Termine`sp RDF URI naming strategy

This URI naming strategy reuses the lemma of the terms in the identifiers of lexical entries. This leads to a merge of homonymous independent entries in `Termine`sp, originally split by concept identifier, into polysemous lexical entries which are linked to a range of senses, referring to different concepts. This fits into the conversion of a concept-oriented terminology into a lexicon to query semasiologically, as `Termine`sp’s search function by lemma, but keeps open the onomasiological view by allowing queries with the `skos:Concept` as starting point (more on this on Chapter 7). The details on this merge are elaborated upon in section 5.2.5, which concerns the linking of `Termine`sp RDF and `Termcat` RDF in the single graph of `Terminoteca` RDF.

This URI naming strategy also allows to link this resource as RDF to other resources at the `ontolex:LexicalEntry` level. However, the designed URI naming strategy requires the use of the part of speech-of-terms in the URI, and, in fact, the figures pre-

¹⁵²As recommended at <https://joinup.ec.europa.eu/community/semic/document/10-rules-persistent-uris>

sented in this chapter show part-of-speech information which was not originally provided in the resource.

In order to link each entry to its corresponding syntactic category through `lexinfo:partOfSpeech` at the `ontolex:LexicalEntry` level, TreeTagger¹⁵³ was used. The initial idea was to tag the Spanish data in order to obtain part-of-speech information that holds for translations to different languages as well. In other words, given that the terms are highly specialised, a mechanical term in Spanish that is a noun, for example, is likely to have a corresponding translation in English or German that is a noun as well. However, the nature of Terminesp entries is mixed: the data set is made up of adjectives, verbs and nouns, along with complex noun phrases (NP), prepositional phrases (PP) and adjective phrases (AP). In order to represent this, we linked multi-word Terminesp entries to `lexinfo:NounPhrase`, `lexinfo:AdjectivePhrase` and `lexinfo:PrepositionPhrase` accordingly. In this way we are encoding the syntax of a prepositional phrase and also stating that it may function as an adjective or as an adverb (via `lexinfo:partOfSpeech`), for instance. Table 5.3 shows the distribution of the different part of speech tags, and which of them involve a complex phrase structure (NPs, PPs, APs).

<code>lexinfo:PartOfSpeech</code>	<code>Synt.Phrase</code>	Entries
noun	simple	13777
	NP	18552
verb	simple	69
adjective	simple	31
	PP	56
	AP	4
adverb	PP	2
Total		32491

Table 5.3: Distribution of part-of-speech categories and syntactic phrases in Terminesp Spanish entries. The tag *simple* is included here for comparative purposes: it refers to those entries that do not involve a complex constituent structure.

Interestingly, the PPs are regarded here as independent entries and there is not any information pointing to the head of the NP where they usually occur, even though

¹⁵³<http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>

there are complex NPs in the data that are formed by a noun and a PP that occurs as lexical entry, too. Thus, we find a PP such as *en reposo* ‘idle’ as a *TermineSp* entry and other entries (NPs) with that same PP as constituent: *masa en reposo*, *tinta en reposo*, *pasador en reposo*. However, the PP and the NP entries are not related in the data. Not only does this contrast with conventional dictionaries, where the preposition is usually accessed through its NP complement (**reposo**, en –) or the whole PP is accessed through the head of the full NP (**masa**, – en reposo), but it also prevents us from using OntoLex’s syntax and semantics module to encode syntactic behaviour: since we cannot access the head of the NP where the PP functions as complement, we lack knowledge on the kind of syntactic frame whose slot is being filled. Moreover, some of these PPs can also accompany a verb (*estar en reposo*, ‘to be idle’; *funcionar en reposo*, ‘to work idle’), so that the type of syntactic frame we are dealing with is not always inferable from the PP alone.

Most entries in *TermineSp* are actually NPs (see Table 5.3): e.g. *potencia isotrópica radiada equivalente* ‘equivalent isotropically radiated power’. With an initial random test set of 500 entries, TreeTagger achieved 0.997 precision and 0.995 recall. The reason behind this is the high number of nouns. In cases in which the entry was a complex phrase, TreeTagger tagged every element in it, which allowed us to identify prepositional phrases. The remaining multi-word entries were initially tagged as nouns. An analysis of the errors revealed that deverbal adjectives were used as nouns throughout the data, and that some multi-word entries included their tag (*capacitivo*, *adj.*, ‘capacitive’) or even a disambiguation note: *funcionar (para los relés elementales)*, ‘function (for elemental relays)’. The tags for the scarce adjectives and verbs were checked and corrected manually, and PPs were tagged as adjectives or adverbs according to their definition and sample uses, if the latter were available.

It is worth mentioning, however, that PPs pose modelling problems still. There seem to be different degrees of lexicalization among them: some of them are fixed both in the specialised domain and in the general language (*en reposo*); others, e.g. *a circuito abierto* ‘open-circuited’, may admit a certain degree of variation and are not even regarded as a set phrase outside the specialized domain. Being the meaning of these entries compositional (to a certain degree), they could not be considered *idioms* according to the definition of `lexinfo:idiom`, nor are they collocations in the sense of

`olia:Collocation`¹⁵⁴, and they do not correspond either to the category of prepositional constructions as, for instance, composite prepositions (*in front of*) or prepositional adverbs (*outside*), which are accounted for in linguistic terminology repositories. Furthermore, translations from the Spanish entry (a PP) into other language may be in the form of PPs as well, adjectives, or adverbs, depending on the target language. The part-of-speech that we assigned to the Spanish entry itself is subject to change given that we do not have the syntactic context in which the entry occurs: some of them are eligible for both adjectival and adverbial use. Capturing these nuances, however, was outside the scope of this thesis but will be considered for future work on terminological data.

5.4.2 The representation of Termcat as RDF

The open terminologies that we chose from Termcat comprised the domains *Internet*, *Telecommunications*, and *Electronics*. These domain lexica include translations from Catalan to Spanish, English, and, if available, to French, as well as synonyms, abbreviated forms, initialisms, and obsolete or dismissed denominations. As Termesp entries, Termcat records are concept-oriented, i.e, a polysemous term will be recorded in two different entries. The terminologies selected from Termcat share with Termesp the fact that they include translations, but, in contrast to Termesp, they lack definitions and include synonyms. Also, while Termesp lacks morphosyntactic information, Termcat includes part-of-speech and gender values, initialisms, and abbreviations for each term. A prototypical record in Termcat source XML files is shown in Example 5.11.

¹⁵⁴A *collocation* in OLiA (<http://purl.org/olia/olia.owl#Collocation>) is defined as «any habitually linked group of words - a kind of lexical partnership, e.g. ‘fish and chips’, ‘salt and pepper’, ‘don’t mention it’, ‘it’s nothing...’, ‘Oh well!’, ‘bangers and mash’... and so on. »

```

<fitxa num="32">
  [...]
  <denominacio llengua="ca" tipus="principal" jerarquia="terme pral."
    categoria="n m, f">administrador -a de xarxes</denominacio>
  <denominacio llengua="ca" tipus="remissio" jerarquia="terme pral." categoria
    ="n m, f">gestor -a de xarxes</denominacio>
  [...]
  <denominacio llengua="en" tipus="equivalent" jerarquia="terme pral."
    categoria="">network administrator</denominacio>
  [...]
</fitxa>

```

Example 5.11: XML of the entry ‘administrador -a de xarxes’

Figure 5.15: RDF representation of the entry ‘administrador -a de xarxes’

Following the modelling we proposed for Terminesp RDF, a Termcat entry is modelled as an **ontolex:LexicalEntry** extracted from the specific domain dictionary (provenance encoded by **dc:source**) and the lemma is given as the written representation of the lexical form. Being concept-oriented, Termcat lexica are structured around terminological records with an ID, as Terminesp entries. Following the approach adopted

in Termesp RDF, we reuse this ID and create a `skos:Concept` as reference for the lexical sense of the term. As noted in Section 4.2.2, Example 5.11 shows two terms in Catalan, with the first one (*administrador -a de xarxes*) being of type *principal* and the second one (*gestor -a de xarxes*) *remissio*, ‘cross-reference’. Since this second term element is included in the same conceptual record as the first one, and the types of cross-reference always refer to either other denominations of the same concept (obsolete or current) or synonyms, we have mapped the `ontolex:LexicalSense` elements of *gestor -a de xarxes* and *administrador -a de xarxes* to the same concept as reference, uniquely identified by the ID of the terminological record and the domain of the lexicon from which the terms were extracted. It is at the level of the conceptual reference where we state the domain of the lexicon we are converting, in this case, Information Society.

If this record provided a synonym marked as such, the modelling would resemble a translation relation. Synonyms have been modelled with the `vartrans` module and they constitute their own lexical entries in the lexicon. For any given entry that includes a synonym, the lexical sense of that synonym is defined as `vartrans:target` of a `vartrans:SenseRelation` with the sense of the first entry as `vartrans:source`. Both senses point to the same `skos:Concept`, which is domain-dependent. The `vartrans:category` property pointing to the LexInfo property `lexinfo:synonym` states the nature of the semantic relation. An example of the RDF of the entry in Catalan *error* ‘error’ in Termcat with its synonyms *brossa* ‘error, bug’ is presented in Example 5.12:

```

:lexiconCA/error-n-ca-Internet-sense a ontolex:LexicalSense .
:lexiconCA/error-n-ca ontolex:sense :lexiconCA/error-n-ca-Internet-sense
.

:termcat/613-T0Internet-concept a skos:Concept ;
    dc:subject "Societat de la informació"@ca .
:lexiconCA/error-n-ca-Internet-sense ontolex:reference :termcat/613-
    T0Internet-concept .
:lexiconCA/brossa-n-ca-Internet-sense a ontolex:LexicalSense ;
    ontolex:reference :termcat/613-T0Internet-concept .
:lexiconCA/brossa-n-ca ontolex:sense :lexiconCA/brossa-n-ca-Internet-
    sense .

:lexiconCA/error-n-ca-Internet-sense-brossa-n-ca-Internet-sense-relation
    a vartrans:SenseRelation ;
    vartrans:source :lexiconCA/error-n-ca-Internet-sense ;
    vartrans:target :lexiconCA/brossa-n-ca-Internet-sense ;
    vartrans:category lexinfo:synonym .

```

Example 5.12: Example of the representation of the entry *brossa* ‘error bug’ as a synonym of *error* in Catalan in Termcat RDF

Both the part-of-speech as well as the gender information are encoded by using LexInfo properties (`lexinfo:partOfSpeech` and `lexinfo:gender`, respectively). In addition to synonyms and translations from Catalan to other languages, Termcat data also includes abbreviations, initialisms, and dismissed terms. All translations, be they among full forms, initialisms or other abbreviated forms, were modelled with the `vartrans` module (see Figure 5.16, discussed below in relation to translations of full terms).

Example 5.13 shows the modelling of an entry in Termcat which has an initialism. In this example, the term *Electronic Programming Guide* has two different forms, one being the full form and the other one the initialism *EPG*. Its lexical entry is linked to its full form by an `ontolex:canonicalForm` property and by its initialism via `ontolex:otherForm`. The initialism form is itself the lexical form of an independent lexical entry *EPG*, which has as term type `lexinfo:initialism` and which is linked to the lexical entry *Electronic Programming Guide* through `lexinfo:initialismFor`. This entry has a lexical sense which is mapped to the same `skos:Concept` as the full form of the term. Other abbreviated forms that are *not* initialisms are considered `lexinfo:Abbre-`

viatedForm elements.

```
:lexiconEN/electronic+programming+guide-n-en a ontolex:LexicalEntry .
:lexiconEN/electronic+programming+guide-n-en-form a ontolex:Form ;
    ontolex:writtenRep "electronic programming guide"@en .
:lexiconEN/electronic+programming+guide-n-en ontolex:canonicalForm :
    lexiconEN/electronic+programming+guide-n-en-form .
:lexiconEN/electronic+programming+guide-n-en-Internet-sense a ontolex:
    LexicalSense;
ontolex:reference :termcat/730-T0Internet-concept .
:lexiconEN/electronic+programming+guide-n-en ontolex:sense :lexiconEN/
    electronic+programming+guide-n-en-Internet-sense .

:lexiconEN/EPG-n-en a ontolex:LexicalEntry .
:lexiconEN/EPG-n-en-form a ontolex:Form ;
    ontolex:writtenRep "EPG"@en .
:lexiconEN/EPG-n-en ontolex:lexicalForm :lexiconEN/EPG-n-en-form.
:lexiconEN/EPG-n-en-Internet-sense a ontolex:LexicalSense ;
    ontolex:reference :termcat/730-T0Internet-concept.

:lexiconEN/electronic+programming+guide-n-en ontolex:otherForm :
    lexiconEN/EPG-n-en-form .
:lexiconEN/EPG-n-en lexinfo:termType lexinfo:initialism ;
    lexinfo:initialismFor :lexiconEN/electronic+programming+guide-n-
        en .
```

Example 5.13: The relation between an initialism *EPG* and its full form *Electronic Programming Guide* in Termcat RDF

Lastly, multilingual data in Termcat was modelled following the same approach adopted for Terminesp RDF, based on the *vatrans* module. Figure 5.16 shows the RDF representation of the Catalan term presented in Example 5.13 and its translation to Spanish.

In Termcat we are dealing with different dictionaries and not with a single terminology file, as was the case of Terminesp. The information regarding each term in a domain lexicon is considered to describe a specific sense of that entry in that domain. This results in a configuration similar to that of Terminesp: a single entry with a URI

Figure 5.16: The representation of translations in Termcat RDF

formed by the written representation, the part-of-speech and the language code merges the different entries extracted from all domain lexica in which the term occurs as record. Table 5.4 summarises the URI naming strategy in those cases which deviates from the one followed in Terminesp RDF. However, for most elements, Table 5.2 for Terminesp RDF also holds for Termcat RDF.

The following section, on the Terminoteca RDF, elaborates on the linking of the two resources, and how one enriches the other.

5.4.3 Terminoteca RDF

The resource Terminoteca RDF aims to be a collection of interlinked multilingual terminologies in Spain, currently including Terminesp RDF and the three domain lexica of Termcat RDF described above. As a result of their integration, they are both connected in a common graph and exposed as linked data on the Web. The data is accessible

Element	Pattern	Example
Lexical Sense (source)	<Lexicon>/<URI of source target LexicalEntry>-<Domain>-sense	:lexiconCA/administrador%2B-a%2B de%2Bxarxes-n-ca-Internet-sense
Concept	<resource from which the concept is extracted (termesp termcat)> /<ID of terminological record in domain lexicon> -<domain>-concept	:termcat/32-TOInternet-concept
Translation	<TranslationSet>/<resource from which the concept is extracted (termesp termcat)>/<ID of terminological record in domain lexicon> -<domain>-<source language tag>-<target language tag>-sense-trans	:tranSetCA-ES/termcat/1-TOInternet-ca -es-sense-trans

Table 5.4: Termcat RDF URI naming strategy

through an SPARQL endpoint and also via a web interface.¹⁵⁵ In a first stage, we would like Terminoteca RDF to become a platform of reference for those languages that are official in Spain (Spanish, Catalan, Basque and Galician), but it already contains term descriptions in many other languages, coming from the term translations provided in each resource.

Following the guidelines for linked data generation (Vila-Suero et al., 2014a; Villazón-Terrazas et al., 2012), we analysed both resources, adopted a URI naming that allowing us to link both of them at the lexical entry level, and modelled them following the same model and decisions on how to apply it to these resources. Terminoteca RDF is currently in its version 3.

Terminoteca RDF combines two different URI naming strategies, as Tables 5.4 and 5.2 show, which converge on the URIs for `ontolex:LexicalEntry` and `ontolex:Forms` elements. Both terminologies differ in the way their original data were identified and they were converted independently of one another, which led to adhere to similar but not entirely equal strategies. While Termesp URIs for elements bearing semantic information were constructed with identifier preservation in mind, Termcat URIs follow the strategy for Apertium, where lexical entries, forms, and senses include in their URI the lemmas of the words they describe along with their part-of-speech and language code (e.g.:`lexiconEN/accelerator-n-en-Internet-sense`). In the case of senses, for the lack of a unique identifier at the resource level, the domain is included to avoid collisions. Concepts extracted from both terminologies also consist of a path element

¹⁵⁵<http://linguistic.linkeddata.es/terminoteca>

/terminep/ or /termcat/ as part of their URI in the namespace of Terminoteca RDF. This also holds for the translations between senses available in both sets of data.

By relying on the pattern “**lexicon<Language>/written representation - part of speech-language code**” for lexical entries, we are propagating the main advantages identified for Apertium RDF to Terminoteca RDF: independent (homonymous) terminological records in both Terminep and Termcat are grouped together under the same polysemous entry, serving as entry point to access all senses and translations of the word, or its relations to other lexical entries (e.g. initialisms); and lexica in different languages grow dynamically as more resources are converted.

Figure 5.17 illustrates the linking of Terminep RDF and Termcat RDF data at the `ontolex:LexicalEntry` level with the example in Spanish *red* ‘network (Internet), net (textile)’. The word *red* ‘network’ in Spanish occurred three times as a lexical entry in Terminep, one per conceptual record, one thus one per each concept it denotes: (concept IDs) 38756, 54593, and 63841 (Bosque-Gil et al., 2015). By transforming the URIs of the lexical entries according to the naming strategy devised for Termcat, these entries are then merged into one single lexical entry of the form `:lexiconES/red-n-es`, even though their senses still keep their identifiers as part of the URI and are mapped to different concepts. This results in a configuration similar to that of Terminep: a single entry with a URI formed by the written representation, the part-of-speech and the language code merges the different entries extracted from all domain lexica in which the term occurs as record.

Figure 5.18 depicts the initial configuration of Terminep and Termcat lexica before converting them to RDF. Each note represent a lexicon in a different language. In Terminep, the Spanish lexicon constitutes the core of the resource, with translations to German, Swedish, Italian, and French. In Termcat, the Catalan lexicon serves this same function, so that the Spanish, French and English lexica are extracted by gathering translations from Catalan. The current state of Terminoteca RDF is displayed at the bottom of the figure, with emphasis on the lexica that were affected by this merge. Note that the overlap between these lexica is not full, as not all entries in each language are present in both resources.

Table 5.5 provides an overview of the number of terms that occur separately in Terminep and Termcat Spanish, English and French lexica (for the selected domains) and the number of entries merged in Terminoteca RDF.

Figure 5.17: The entry *red* ‘network (Internet), net (textile)’ in Terminoteca RDF. `ontolex:LexicalSense` elements from Terminesp RDF and Termcat RDF are grouped together under the same lexical entry.

	LexiconES	LexiconEN	LexiconFR
Exclusive from Terminesp	26941	14211	13970
Exclusive from TERMCAT	5583	6006	1028
Present in both (merged)	1087	878	297
Total	33611	21095	15295

Table 5.5: Distribution of entries in the Terminoteca

By adopting the URI naming strategy introduced above, the entries that were previously divided by concept are now merged into a single one if they share the written representation and part-of-speech. For instance, the entry `:lexiconES/ensayo-n-es` ‘test’ is now linked to eleven lexical senses (e.g. `:lexiconES/39026es-sense`, `:lexiconES/ensayo-n-es-IndustriaElectronica-sense`, etc.), ten of them extracted from Terminesp and with different definitions, and one of them gathered from Termcat *Electronics* domain lexicon. Some of the senses are linked to the lexical senses of entries in other languages through a translation relation: `:lexiconCA/assaig-n-ca` (from Termcat), and `:lexiconFR/essai-n-fr` (from Terminesp). Likewise, one

Figure 5.18: Terminesp and Termcat before linking (above) and the Terminoteca RDF in its current state

of the lexical senses of the entry `:lexiconES/cámara+de+televisión-n-es`¹⁵⁶ receives a synonym from Termcat (`:lexiconES/telecámara-n-es`), as well as a translation to Catalan `:lexiconCA/càmera+de+televisió-n-ca`. The English translation `:lexiconEN/television+camera-n-en` comes from both Terminesp and Termcat, and the links to the French `:lexiconFR/caméra+de+télévision-n-fr` and German entries `:lexiconDE/Fernsehkamera-n-de` are provided by Terminesp.

LD technologies allow that terms formerly described in an isolated fashion, though belonging even to the same terminology, are now reusable, linked to one another, and enriched with the information provided by complementary terminological resources. The integration with other terminologies in official languages in Spain (Basque, Galician), as well as the linking to other available external resources such esDBpedia¹⁵⁷, the Apertium dictionaries or IATE RDF (Cimiano et al., 2015b) are planned as continuation steps for the Terminoteca RDF.

¹⁵⁶As percentage-encoded URI, this IRI actually reads: `:lexiconES/c%C3%A1mara+de+televisi%C3%B3n-n-es`. Accents are included here for the sake of readability.

¹⁵⁷<http://es.dbpedia.org/>

5.5 The modelling of K Dictionaries Multilingual Global Series Spanish dictionary as RDF

K Dictionaries (KD) Spanish dictionary from the multilingual Global series serves in this thesis as a representative example of the modelling of a multilingual dictionary aimed for both human and machine consumption. It shows a more fine-grained lexical description in comparison to Apertium, Rayfield’s Georgian-English dictionaries, Termisp or Termcat (see Table 4.4 at the end of Chapter 4). In 2016, we developed a prototype to model the Spanish dictionary as LLD which would also be applicable to the English and German datasets, although not exhaustive for these two languages. As a condition for this work, the proposed prototype had to allow for a reconstruction of the original XML from the RDF representation – a round tripping condition. Following the methodology for LLD generation (Vila-Suero et al., 2014a), after the analysis of the source data provided in Chapter 4 (Section 4.3), we proceeded with the definition of the model and the URI naming strategy for each of its elements. The results of this work are presented in Bosque-Gil et al. (2016b).

The difficulties in representing KD’s data as RDF concern the mapping from their proprietary DTD to OntoLex-*lemon* and LexInfo, which were the two selected vocabularies after Step 1 of our model definition process. The difficulties can be broadly grouped in three types, listed below:

1. Structural aspects: Due to the round tripping condition set by data providers for this effort, the model needed to account for information related to the arrangement of entries in the dictionary. This data is necessary in order to recreate the XML from the RDF counterpart in a future step. However, the selected vocabularies do not include elements aimed at encoding this structural information.
2. Dictionary annotations: the selected vocabularies are not exhaustive for the representation of information related to the specific use of headwords or senses. This information is frequently provided through abbreviations or notes in dictionary entries.
3. Incompatibilities between the selected vocabularies and the resource data model: concepts with the same label differ in granularity or are defined at different level of abstraction, etc.

The modelling effort for KD’s implemented in 2016 was revised and updated according to the practices discussed during 2017-2018 in the W3C OntoLex Community group for the development of the *lemon* lexicography module – *lexicog* (see Chapter 6). Such a module addresses problems (1) and (2).

In the following sections we will adhere to the incremental ordering we have followed in the previous pages, from the representation of a simple monolingual entry to the modelling of multilingual entries with examples of use, translations and further annotations, as the latter cases build on the former ones.

5.5.1 The representation of the micro-structure of an entry from KD multilingual Global series

Core of the model. Example 5.14, shown below, is an example of the monolingual lexical information in the source XML that the RDF needed to capture: the headword, its morphological and phonetic information (at the lemma level, in this case, *arriesgdo* ‘risky’), its list of senses, their definitions and the examples associated with each sense. Due to the round tripping condition set on this prototype, entry and dictionary entry IDs, as well as display information (needed for the user interface in the original XML) were modelled as RDF elements as well.

Figure 5.19 shows the RDF representation of the monolingual information provided in Example 5.14. Table 5.6 presents a simplified table of mappings from the DTD elements shown in Example 5.14 to the implemented RDF model. This example resembles the examples we have shown in previous sections as regards the representation of the morphosyntactic information of the lexical entry (part-of-speech, forms), the lemma itself, its pronunciation and its link to senses (in the picture, to only one sense, SE00007336). Most of the elements depicted in Figure 5.19 correspond to the OntoLex-*lemon* core. This part of the application profile of OntoLex-*lemon* and LexInfo for KD is shared with the Apertium RDF and the Terminoteca RDF representation, so that only those distinctive elements in KD have been included in Table 5.6 mappings for comparison purposes.

```

<Entry HomNum="" hw="arriesgado" identifier="EN00004777" pos="adjective">
  <DictionaryEntry identifier="DE00005461" version="1">
    <HeadwordBlock>
      <HeadwordCtn>
        <Headword>arriesgado</Headword>
        <PronunciationCtn>
          <Pronunciation>arjes'yaðo</Pronunciation>
        </PronunciationCtn>
      </HeadwordCtn>
      <HeadwordCtn>
        <Headword>arriesgada</Headword>
        <Display>-da</Display>
        <PronunciationCtn>
          <Pronunciation>arjes'yaða</Pronunciation>
          <Display>-ða</Display>
        </PronunciationCtn>
      </HeadwordCtn>
      <PartOfSpeech value="adjective" />
    </HeadwordBlock>
    <SenseBlock>
      <SenseGrp identifier="SE00007336" version="1">
        <Synonym>azaroso</Synonym>
        <Definition>que es peligroso o riesgoso</Definition>
        <TranslationCluster identifier="TC00017071" text="que es peligroso o
          riesgoso" type="def">
          [...]
        </TranslationCluster>
        <ExampleCtn type="sid" version="1">
          <Example>una misión arriesgada</Example>
          <TranslationCluster identifier="TC00017072" text="una misión
            arriesgada" type="exmp">
            [...]
          </TranslationCluster>
        </ExampleCtn>
        [...]
      </SenseGrp>
      <SenseGrp identifier="SE00007337" version="1">
        [...]
      </SenseGrp>
    </SenseBlock>
  </DictionaryEntry>
</Entry>

```

Example 5.14: Prototypical monolingual data in a dictionary entry in XML from KD's multilingual Global series, in this case, *arriesgado* 'risky'

Figure 5.19: The representation of the monolingual information in the entry *arriesgado* 'risky'

DTD Element	RDF modelling
Dictionary	kd:DictionaryEntry ->kd:dictionaryEntry ->ontolex:LexicalEntry
Entry Identifier	ontolex:LexicalEntry ->kd:entryId ->string
DictionaryEntry	ontolex:LexicalEntry
DictionaryEntry Identifier	ontolex:LexicalEntry ->kd:dictionaryEntryId ->string
HeadwordBlock	Different ontolex:Forms
Display	ontolex:Form ->kd:display ->kd:Display
ExampleCtn	ontolex:LexicalSense ->skos:example ->kd:UsageExample kd:UsageExample ->rdf:value ->string@source language.

Table 5.6: DTD elements and their RDF counterparts and modelling

In relation to general morphosyntactic information discussed in previous sections, however, note that the original source data in KD presents a block of headwords marked with a `HeadwordBlock` tag, which is interpreted as a set of `ontolex:Form` elements. This differs from the use of punctuation marks in Rayfield’s dictionary to signal various forms for the same entry. Senses are encoded through `ontolex:LexicalSense` and instances of `skos:Concept` representing the denotation, along with a definition string introduced with `skos:definition`, as in *Termine*sp RDF.

Synonyms and antonyms. Synonyms are treated as independent lexical entries related to the original dictionary entry through their senses taking part in a `var-trans:SenseRelation`, the same approach followed in the modelling of *Termcat* RDF. As shown in Example 5.14, synonyms in KD are provided as strings related to a sense of the source entry, without pointers to the dictionary entry of the synonym (if available). In *Termcat* RDF, whose source was concept-oriented, synonyms were associated a `ontolex:LexicalSense` identified with the name of the domain, and they denoted the concept described in the terminological record of the main entry. Since we lack this onomasiological perspective in KD and also concept identifiers, we create ‘artificial senses’ for the lexical senses of the synonyms by inheriting the source sense URI. These ‘artificial senses’ point to the same `skos:Concept` denoted by the source entry. A future step would involve to map those artificially created senses to real senses provided in the dictionary entry of the synonym. In the case of antonyms, artificial instances of `skos:Concept` needed to be created as well.

Usage examples. In addition to sense definitions, captured by the *OntoLex-lemon* core, each sense is further described with usage examples. While the *lemon* model provided a class `lemon:UsageExample` and a property `lemon:example`, these are no longer included in *OntoLex*. In order not to instantiate both *lemon* and its revisited in the same model application profile, we proposed the *ad hoc* class `kd:UsageExample` linked to the sense through `skos:example` to encode usage examples. Chapter 6 addresses this gap by proposing a new class `lexicog:UsageExample` as part of the *lemon* lexicography module.

Dictionary structure and display. Due to the conditions on information preservation set on this prototype, it was crucial to maintain in the RDF representation all lexicographical information that was not necessarily lexical in nature and rather concerned either the internal organisation of the entries or the user interface of the dictionary. This involved representing IDs in the RDF rendering (through the properties `kd:entryId` and `kd:dictionaryEntryId`) and display information for the interface (through `kd:Display` and `kd:display`), as shown in Table 5.6.

The distinction between lexical entries belonging to the original resource as headwords (e.g. as *arrigado* in Example 5.14) and lexical entries generated on the fly, which do belong to a `lime:Lexicon` but *may or may not* occur later on as headwords in the dictionary, was also an aspect for the RDF representation to reflect. Despite the fact that synonyms and antonyms should also be dictionary entries (i.e. they belong to the headword list), according to the editors' guidelines, there is no method ensuring this scenario. Likewise, compound terms composed of the headword are defined inside a `SenseGroup` but do not occur as headwords in the dictionary (although they belong to the lexicon of the language as lexicalized items). This approach of treating synonyms (and any other entry described inside a sense group of another entry, e.g. compounds, antonyms, etc.) as `ontolex:LexicalEntry` instances allows for the automatic growth of the source lexicon, but required a distinction between original dictionary entries and `ontolex:LexicalEntry` ones. This distinction assumes that a `lime:Lexicon` may be different from the original lexicographic resource that is converted to RDF. On these grounds, the class `kd:Dictionary` (in opposition to `lime:Lexicon`) was proposed. Chapter 6 will elaborate on the differences between these elements and

the classes `lexicog:LexicographicResource` and `lexicog:Entry` as complementary classes of OntoLex elements.

Monolingual information in KD goes beyond morphosyntactic data, examples, headword blocks or synonym and antonyms. Geographical usage, register, sense disambiguators, selectional restrictions or domain information are also provided in these entries.

Geographical usage and contextual data. Example 5.15 shows the entry *alcachofa* in Spanish, meaning ‘artichoke’ in its first sense (omitted here) and also ‘shower head, sprinkler’ in its second sense. The **GeographicalUsage** tag in the second sense indicates that this use, *alcachofa* as ‘sprinkler’, is only attested in Spain: (*Esp*). Note that this does not indicate that the spelling or the word itself is only attested in Spain (as opposed to spelling variants in British and American English), but that a single sense is recorded in a specific place. This sense also has a sense disambiguator or indicator *dispositivo* ‘device’, specifying the kind of entity the headword denotes in this sense. The data provided by the tag **SenseIndicator**, in general, according to the editor guidelines, either refers to the hypernym of the entry or provides contextual aspects by a prepositional phrase that gives usage clues ([used] *in...*).

Figure 5.20 shows this information as RDF, with the implemented mappings listed in Table 5.7.¹⁵⁸ Because sense indicators in KD data provide either contextual information as a prepositional phrase or as an hyperonym, we interpreted the hyperonym also as pragmatic information on the usage of the word (with this sense), and hence reuse the `ontolex:usage` property to represent this data.

¹⁵⁸The elements explained in Table 5.6 which are not required to illustrate the mappings of **GeographicalUsage** and **SenseIndicators** are not included in the figure.

```

<DictionaryEntry identifier="DE00002870" version="2">
  <HeadwordCtn>
    <Headword>alcachofa</Headword>
    [...]
  </HeadwordCtn>
  <SenseBlock>
    <SenseGrp identifier="SE00003650" version="2">
      [...]
    </SenseGrp>
    <SenseGrp identifier="SE00003651" version="2">
      <GeographicalUsage value="Esp" />
      <SidCtn identifier="SI00002742" version="2">
        <SenseIndicator>dispositivo</SenseIndicator>
      </SidCtn>
      <Definition>pieza con agujeros pequeños para esparcir
      agua</Definition>
      [...]
    </SenseGrp>
  </SenseBlock>
</DictionaryEntry>
</Entry>

```

Example 5.15: An extract of the Spanish entry *alcachofa* ‘artichoke, shower head, sprinkler’ with a `GeographicalUsage` tag and a `SenseIndicator`.

The data on the geographical usage of dictionary entries comes at different levels in KD, as explained above. In the example, it belongs to a sense group. *OntoLex-lemon* or *LexInfo* do not provide an element to include geographical information beyond variants (i.e. by establishing a relation between two entries), written form features (e.g. for the description of spelling variants), or the `ontolex:usage` property, which is already mapped to a `SenseIndicator` tag in our application profile for KD. The *Lexvo* ontology does provide a property `lexvo:usedIn` that has been used in the literature to indicate the geographical area in which an entry is used (Moussallem et al., 2018), but, according to the documentation, the property is defined to be used to refer to a language or to a writing system. We have opted for the creation of a `kd:geographicalUsage` property to point to individuals from the KD custom vocabulary corresponding to the values that the `GeographicalUsage` tag can take. A future step would be to replace those values with their matching DBpedia instances of geographical regions.

Since this tag, `GeographicalUsage`, occurs at different levels of the DTD, the subject

DTD Element	RDF modelling
SenseIndicator	ontolex:LexicalSense ->ontolex:usage ->[rdf:value->string]
SenseGrp/GeographicalUsage	ontolex:LexicalSense ->kd:geographicalUsage ->individual from the KD Ontology

Table 5.7: DTD elements `SenseIndicator` and `GeographicalUsage` and their RDF counterparts and suggested modelling

of `kd:geographicalUsage` may be an `ontolex:LexicalSense` (cases in which the tag occurs at the sense group level), `kd:InflectionBlock` (the tag is directly under an `InflectionBlock` element in the XML, which groups different word forms together – see the modelling of *inflection containers* later in this section) or `ontolex:Form` (e.g. the tag is inside an `InflectionCtn` element).

Figure 5.20: The RDF representation of the data concerning geographical usage and contextual information provided in the entry *alcochofa* ‘artichoke, shower head, sprinkler’

In cases in which the tag `GeographicalUsage` occurs inside an XML element describing the pronunciation of the lemma, this information is preserved as part of the language tag of the phonetic representation of the headword: e.g. `@en-CA-fonipa` for Canadian pronunciation.

Selectional restrictions. The monolingual description of the senses of the entries also includes selectional restrictions, i.e. restrictions on the semantic types of the syn-

tactic arguments of the headword. This information is encoded with the tag **RangeOfApplication** in KD’s DTD. While this can be interpreted also as contextual information, as that provided by **SenseIndicator** elements, after a careful analysis of the data and the guidelines, we interpreted this tag to be solely restricting the semantics of the allowed arguments, and thus we aimed to reflect this distinction on the RDF model as well.

Example 5.16 shows a portion of the entry *arriscado* in Spanish, ‘formed/composed of cliffs’, commonly used to describe a geographical landscape. When used in its second sense, it is predicated of a person who goes through landscapes with cliffs, for instance, a climber. Note that the **RangeOfApplication** is *persona* ‘person’, which is different from the hypernym data given as **SenseIndicator** for Example 5.15 (*shower head* is an hyponym of *device*). Figure 5.21 shows the RDF representation of **RangeOfApplication** as **synsem:condition**, thus capturing our interpretation that being a person is a semantic feature required by the argument of *arriscado*. This is also illustrated in Table 5.8. Interestingly, this scenario is similar to the lexicalization of properties that *OntoLex-lemon* allows with the *synsem* module, with elements such as **synsem:OntoMap** and **ontoCorrespondence**. These elements aim at relating subjects and objects of the property in the ontology to the arguments in the syntactic construction. Since in this case we are not lexicalizing ontology properties, we cannot use this workaround to establish a mapping between the syntactic frame of the word lexicalizing the ontology property and the domain and range of the ontology property itself.

```

<Entry HomNum="" hw="arriscado" identifier="EN00004786" pos="adjective">
  <DictionaryEntry identifier="DE00005471" version="3">
    <HeadwordBlock>
      <HeadwordCtn>
        <Headword>arriscado</Headword>
        <PronunciationCtn>
          <Pronunciation>aris'kaʃo</Pronunciation>
        </PronunciationCtn>
      </HeadwordCtn>
    </HeadwordBlock>
    [...]
    <SenseBlock>
      <SenseGrp identifier="SE00007352" version="3">
        [...]
      </SenseGrp>
      <SenseGrp identifier="SE00007353" version="3">
        <Register value="formal" />
        <RangeOfApplication>persona</RangeOfApplication>
        [...]
      </SenseGrp>
      <Definition>que va por lugares peligrosos o difíciles de
        escalar</Definition>
      <TranslationCluster identifier="TC00017112" text="que va por lugares
        peligrosos o difíciles de escalar" type="def">
        [...]
      </TranslationCluster>
    </ExampleCtn>
  </SenseBlock>
</DictionaryEntry>
</Entry>

```

Example 5.16: An extract of the Spanish entry *arriscado* ‘formed/composed of cliffs, climber’ with a *RangeOfApplication* tag and a *Register* value *formal*

Register. Example 5.16 also contains a *Register* tag in the sense annotated with a *GeographicalUsage* tag. *Register* occurs also at the lemma level throughout the data but, after some discussion with the data owners, it was agreed to represent it as a property of the sense. This was decided on the grounds that register information is considered to pertain to the semantics and pragmatics of the entry and its use, and

Figure 5.21: The representation of selectional restrictions and register information on the Spanish entry *arriscado* ‘formed/composed of cliffs, climber’.

DTD Element	RDF modelling
SenseGrp/RangeOfApplication	ontolex:LexicalSense ->synsem:condition ->[rdf:value string@language-tag]
SenseGrp/Register HeadwordBlock/Register HeadwordCtn/Register	ontolex:LexicalSense ->lexinfo:register ->Indivfual from lexinfo:Register or from the KD Ontology

Table 5.8: Mappings from the DTD elements **RangeOfApplication** and **Register** capturing selectional restrictions amd register information, and their counterpart in *OntoLex-lemma* and *LexInfo*.

it does not relate to morphological information or to other properties inherent to the entry (these are types of information recorded at the lexical entry or word form levels). This decision also allows us to reuse the property **lexinfo:register** and not to create an *ad hoc* property under the KD namespace, as illustrated in Figure 5.21. Table 5.8 summarises the mappings from the DTD to the selected vocabularies.

Compounds. Inside a sense group, in numerous cases dictionary entries include a compound entry or a related term that is composed of the headword. This information is grouped in KD’s data inside the **CompositionalPhrase** tag, and refers to compounds, figures of speech, idioms, phrasal verbs, etc. Figure 5.22 represents in RDF the mod-

Figure 5.22: The representation of *CompositionalPhrases* as RDF, with the example *earn a living*, defined in the entry *earn*.

elling of a portion of the entry *earn*. In the original data, in one of its senses the entry provides a full definition of the idiom *earn a living*. This embedding is interpreted to encode that the verb *earn* in *earn a living* bears (at least partially) the same sense that is described in that *SenseGrp*.

In order to capture this, we extract the compositional phrase and treat it as lexical entry itself, following the same approach adopted for synonyms and antonyms. While the original dictionary entry belongs to the *kd:Dictionary*, the extracted lexical entry does not, even though both lexical entries belong to the same *lime:Lexicon*. The fact that the sense of *earn a living* is related to the second sense of *earn* is recorded with the property *vartrans:senseRel*. A further representation mechanism would be to use the *decomp:subterm* property and specify that the main lexical entry is a constituent of the compound. However, the *decomp* module is used for decomposition of compounds, and, after discussion with the data owners, not all *CompositionalPhrases* in KD involve in fact a morphological compositional process that results in a lexicalised entry. Table 5.9 lists this mapping. The newly created lexical entry of the *CompositionalPhrase* be-

DTD Element	RDF modelling
CompositionalPhrase	<code>ontolex:LexicalSense</code> of main entry- <code>>vartrans:senseRel</code> <code>->ontolex:LexicalSense</code> of <code>CompositionalPhrase</code> <code>lime:Lexicon ->lime:entry ->LexicalEntry</code> of the compositional phrase

Table 5.9: Mappings from the DTD element `CompositionalPhrase` and its counterpart in *OntoLex-lemon*

has as a regular `ontolex:LexicalEntry`, and its description follows the mappings presented in previous sections.

Sense hierarchy. In a dictionary entry we also find hierarchical arrangements. Sub-senses are sense blocks embedded inside another sense block, and they aim at representing senses of a word that are related to another sense of the same word, but with nuances distinguishing the two. In Example 5.17, the preposition *about* in KD’s English dictionary has several senses. The second one of them has as ID `SE000000031`, and lacks a definition. Inside it, there is another sense group, with two members, the first one related to the sense of the compositional phrase *what about...?* (used to suggest something), and the second one related the sense of *about* used in interrogative sentences such as *how about Norah?* (examples not shown in the figure). Both the sense of *how about* for suggesting and the sense of *about* to obtain information share some semantic features, according to the editor, and are thus represented as sub-senses of a sense that is not defined. This latter sense functions abstractly almost as container of the above two. For this specific scenario, we suggest the use of `skos:narrow` to relate senses to their children, but, as it has been pointed out in the literature (Khan et al., 2016c), the concepts that are the reference of those parent-child senses are not necessarily in a taxonomic relation.

```

<Entry identifier="EN00000017" hw="about" HomNum="2" pos="preposition">
  <DictionaryEntry identifier="DE00000021" version="1">
    <HeadwordCtn>
      <Headword>about</Headword>
      <HomographNumber value="2" />
      [...]
    </HeadwordCtn>
    <SenseBlock>
      [...]
      <SenseGrp identifier="SE00000031" version="1">
        <SenseBlock>
          <SenseGrp identifier="SE00000032" version="1">
            <CompositionalPhraseCtn version="1">
              <CompositionalPhrase>how/what about ... ?</CompositionalPhrase>
              <Definition>indicates you are suggesting sth</Definition>
              [...]
            </CompositionalPhraseCtn>
          </SenseGrp>
          <SenseGrp identifier="SE00000033" version="1">
            [...]
            <Definition>indicates that you want information</Definition>
            [...]
          </SenseGrp>
        </SenseBlock>
      </SenseGrp>
    </SenseBlock>
  </DictionaryEntry>
</Entry>

```

Example 5.17: An extract of the English entry *about*, with a **SenseBlock** tag inside a **SenseGrp**

Nested Entries and Inflection Containers. There are other elements in KD that serve a structural function. Nest entries and inflection blocks serve as containers that group together different dictionary entries (nested) or inflections that all share a series of grammatical properties (inflection blocks). Example 5.18 in XML shows the entry of the verb *besuchen* ‘visit’ in German. The element **NestEntry** groups together three different dictionary entries: *besuchen* (‘v. visit’), *Besuch* (‘n. visit’) and *Besucher* (‘n. guest, visitor’) that are related, although the nature of relation is not explicitly stated. After

the analysis of the data, we realised that **NestEntry** groups together derivations or, in some cases, verbs that share lemma but not subcategorization and are not represented as homonyms – and the editors decided not to treat this case as a polysemous entry with senses differing in subcategorization. Figure 5.23 is an extract of the RDF modelling of Example 5.18 that reflects this encoding: the `ontolex:LexicalEntry` that names the whole **Entry** tag in the XML is related to the other two dictionary entries by `vartrans:lexicalRel`, which solely represents a lexical relation.

```
<Entry HomNum="" hw="besuchen" identifier="EN00002666" pos="verb">
  <NestEntry>
    <DictionaryEntry identifier="DE00003297" version="1">
      <HeadwordCtn>
        <Headword>besuchen</Headword>
        [...]
      </HeadwordCtn>
      [...]
    </DictionaryEntry>
    <DictionaryEntry identifier="DE00003298" version="1">
      <HeadwordCtn>
        <Headword>Besuch</Headword>
        [...]
        <GrammaticalGender value="masculine" />
        <InflectionBlock>
          <InflectionCtn>
            <Inflection>Besuch(e)s</Inflection>
            <Display>-(e)s</Display>
          </InflectionCtn>
          <InflectionCtn>
            <Inflection>Besuche</Inflection>
            <Display>-e</Display>
          </InflectionCtn>
        </InflectionBlock>
      </HeadwordCtn>
      [...]
    </DictionaryEntry>
    <DictionaryEntry identifier="DE00003299" version="1">
      <HeadwordBlock>
        <HeadwordCtn>
          <Headword>Besucher</Headword>
```

```

[...]  

<GrammaticalGender value="masculine" />  

<InflectionBlock>  

  <InflectionCtn>  

    <Inflection>Besuchers</Inflection>  

    <Display>-s</Display>  

  </InflectionCtn>  

  <InflectionCtn>  

    <Inflection>Besucher</Inflection>  

    <Display>-</Display>  

  </InflectionCtn>  

</InflectionBlock>  

</HeadwordCtn>  

<HeadwordCtn>  

  <Headword>Besucherin</Headword>  

  [...]  

  <GrammaticalGender value="feminine" />  

  <InflectionBlock>  

    <InflectionCtn>  

      <Inflection>Besucherin</Inflection>  

      <Display>-</Display>  

    </InflectionCtn>  

    <InflectionCtn>  

      <Inflection>Besucherinnen</Inflection>  

      <Display>-nen</Display>  

    </InflectionCtn>  

  </InflectionBlock>  

</HeadwordCtn>  

<PartOfSpeech value="noun" />  

</HeadwordBlock>  

[...]  

</DictionaryEntry>  

</NestEntry>  

</Entry>

```

Example 5.18: An extract of the German entry *besuchen* ‘v. visit’, with a `NestEntry` container that groups the dictionary entries *besuchen*, *Besuch* ‘n. visit’, and *Besucher* ‘guest, visitor’

Figure 5.23 also shows word forms. In German, nouns inflect for case and number. In addition, *Besucher* ‘guest’ inflects for gender, which leads to two different headwords,

Figure 5.23: An extract of the RDF representation of the dictionary entry *besuchen*, which groups inside a *NestEntry* tag the dictionary entries *besuchen*, *Besuch* and *Besucher*.

in a similar scenario as we saw with the Spanish entry *arriesgado* ‘risky’ (Example 5.14), but not reflected in this extract of the figure. Case and number inflection are provided in *InflectionBlock* elements for each of these headwords. Even though the case is not described, lexicographic tradition usually dictates the order of the cases in which inflections are presented, being the lemma the nominative form (e.g. *Besuch*), followed by the genitive singular (*Besuch(e)s*), and by the nominative plural (*Besuche*). Since the condition for the conversion of this resources was to enable a round trip step in the future by recording all information presented in the original format, the different word forms of the headword were represented as *ontolex:Form* instances but also grouped according to the XML presentation. For this reason, the *kd:InflectionBlock* element and its properties *kd:inflectionHeadword* and *kd:inflectedForm* enable to reconstruct the original file and accurately represent the original structural information from the resource. Table 5.10 summarises the mappings from the DTD elements recording

DTD Element and path	RDF modelling
NestEntry	ontolex:LexicalEntry -> vartrans:lexicalRel -> ontolex:LexicalEntry
InflectionCtn	ontolex:Form
InflectionBlock	kd:InflectionBlock -> kd:headwordInflection -> ontolex:Form kd:InflectionBlock -> kd:inflectedForm -> ontolex:Form

Table 5.10: DTD elements recording inflections (**InflectionCtn** and **InflectionBlock**), and dictionary entry hierarchy (**NestEntry**), and their counterparts in *OntoLex-lemon* and in the custom KD vocabulary

inflection and dictionary entry hierarchy to their *OntoLex-lemon* counterparts.

Each entry in KD Global series contains monolingual and multilingual information. The multilingual information is divided into three different types of data: translations of headwords (dependant on the sense), translations of examples of use of a sense, and information regarding these translations (e.g. their morphological description in the case of translated headwords, data about their context of usage, domain, etc., or transliterations in the case of examples in alphabets other than Latin).

Translations of headwords. Example 5.19 shows an excerpt of the entry *abadía*, ‘abbey’ in Spanish, with a description of its first sense. Inside the **SenseGrp** encapsulating the information of the first sense, there is an element **TranslationCluster** with **Locale** groups that in turn embed containers with headword translations. Note that in Brazilian Portuguese, in addition to the translation, the grammatical gender is provided. Figure 5.24 shows how these translations are modelled in the RDF representation, namely, as instances of `ontolex:LexicalEntry` that belong to a `lime:Lexicon` of the target language (but they will not be linked to the `kd:Dictionary` resource, as they are not dictionary entries, but new, extracted `ontolex:LexicalEntry` elements from the **TranslationClusters**). The information regarding each translation (grammatical features, senses, examples, etc.) is represented following the mappings discussed on previous pages as regards the modelling of monolingual data. However, note that target lexical senses, in contrast to the source sense, do not have a sense ID in the source data. For the URI naming strategy, we inherit this information from the target sense, creating thus artificial senses that resemble those of synonyms. As Figure 5.24 depicts, these translations are represented by means of the *vartrans* module of *OntoLex-lemon*, as in Apertium RDF and Terminoteca RDF.


```

<Entry HomNum="" hw="abadía" identifier="EN00000019" pos="noun">
  <DictionaryEntry identifier="DE00000020" version="1">
    <HeadwordCtn>
      <Headword>abadía</Headword>
      [...]
    </HeadwordCtn>
    <SenseBlock>
      <SenseGrp identifier="SE00000039" version="1">
        [...]
        <Definition>monasterio con territorio propio</Definition>
        <TranslationCluster identifier="TC00000097" text="monasterio con
          territorio propio" type="def">
          <Locale lang="nl">
            <TranslationBlock>
              <TranslationCtn>
                <Translation>abdijs</Translation>
                [...]
              </TranslationCtn>
            </TranslationBlock>
          </Locale>
          [...]
          <Locale lang="br">
            <TranslationBlock>
              <TranslationCtn>
                <Translation>abadia<GrammaticalGender
                  value="feminine"/></Translation>
              </TranslationCtn>
            </TranslationBlock>
          </Locale>
          [...]
        </TranslationCluster>
        <ExampleCtn type="sid" version="1">
          <Example>la Abadía de Westminster</Example>
          <TranslationCluster identifier="TC00000098" text="la Abadía de
            Westminster" type="exmp">
            <Locale lang="nl">
              <TranslationBlock>
                <TranslationCtn>
                  <Translation>de Abdijkerk van Westminster</Translation>
                </TranslationCtn>
              </TranslationBlock>
            </Locale>
          </TranslationCluster>
        </ExampleCtn>
      </SenseGrp>
    </SenseBlock>
  </DictionaryEntry>
</Entry>

```

```

        </Locale>
        [...]
        <Locale lang="br">
            <TranslationBlock>
                <TranslationCtn>
                    <Translation>a Abadia de Westminster</Translation>
                </TranslationCtn>
            </TranslationBlock>
        </Locale>
        [...]
    </TranslationCluster>
</ExampleCtn>
<ExampleCtn type="def" version="1">
    <Example>El conde Guillermo de Tolosa se retiró a una abadía
        benedictina.</Example>
    [...]
</ExampleCtn>
</SenseGrp>
[...]
</SenseBlock>
</DictionaryEntry>
</Entry>

```

Example 5.19: An extract of the Spanish entry *abadía* ‘abbey’, with translations into Dutch and Brazilian Portuguese, examples, and translations of the examples into the target languages

Examples of translations and translations of examples. Examples are a challenging aspect to model: while stating that a sense (of translation) has a usage example was accounted for by using `skos:example` and a custom class `kd:UsageExample`, the tag `TranslationCluster` inside example containers groups translations of these usage examples into the target languages. However, as examples are not lexical entries (they are not units of analysis of the lexicon, but compositionally built structures that are not lexicalised in the language), we cannot represent them as such, create lexical senses for them or link these senses through a `vartrans:Translation`. For this reason we propose a `kd:TranslationExampleCluster` to group `kd:UsageExamples` that are translations of each other. An alternative modelling solution to this problem would be to use the property `lexinfo:translation`, but this would not capture that two exam-

Figure 5.24: An extract of the RDF representation of the Spanish entry *abadía* ‘abbey’, which illustrates the encoding of translations with the *vartrans* module.

DTD Element	RDF modelling
SenseGrp/TranslationCluster/.. Locale/TranslationBlock/TranslationCtn	(target) ontolex:LexicalEntry -> ontolex:sense -> (target) ontolex:LexicalSense -> ontolex:reference -> skos:Concept (of source sense) vartrans:Translation -> vartrans:source -> source ontolex:LexicalSense vartrans:Translation -> vartrans:target -> target ontolex:LexicalSense
ExampleCtn/TranslationCluster/... Locale/TranslationBlock/TranslationCtn	(target) ontolex:LexicalSense -> skos:example -> kd:UsageExample kd:TranslationExampleCluster -> vartrans:relates -> kd:UsageExample

Table 5.11: Mappings from the DTD element **TraslationCtn** capturing translations at two different levels, inside and outside **ExampleCtn**, and its XML paths

ples in different target languages are translations of each other unless we instantiate a `lexinfo:translation` triple for each combination. For a solution to this problem as part of the discussion on the lexicography module under the umbrella of OntoLex community group, see Chapter 6. Table 5.11 provides the mappings from the **TranslationCtn** tag at the two different levels in occurs in the DTD, inside **ExampleCtn** (for translations of examples) and outside example containers but inside sense groups (for headword translations).

Alternative scripting. Some of these examples and translations are in languages with alphabets other than Latin, as illustrated in Example 5.20. For transliterations and other alternative forms of headword translations, new `ontolex:Form` elements are created and the language tag of the lemma (or written representation, in OntoLex terms) will reflect the specific alphabet used, e.g. ‘`yasuku, renka de`’@ja-Latn.¹⁵⁹ The same approach is followed for examples in Japanese: a single `kd:UsageExample` element is characterised by more than one `rdf:value`, each with a different language tag: @ja and @ja-Latn.

¹⁵⁹Note that translations options are not separated here into several `ontolex:LexicalEntry` elements, as they are embedded in the same **Translation** tag in the source data. Future version of KD’s data will see to the separation of alternative translations to be reflected in both the source and the RDF representations.

```

<DictionaryEntry identifier="DE00007110" version="1">
  <HeadwordCtn>
    <Headword>barato</Headword>[...]
    <PartOfSpeech value="adverb" />
  </HeadwordCtn>
  <SenseGrp identifier="SE00009753" version="1">
    <SubjectField value="commerce" />
    <Definition>por poco precio</Definition>
    <TranslationCluster identifier="TC00022903" text="por poco precio" type="
      def">
      [...]
      <Locale lang="ja">
        <TranslationBlock>
          <TranslationCtn>
            <Translation>安（やす）く、廉（れんか）で</Translation>
            <AlternativeScripting type="kanji"></AlternativeScripting>
            <AlternativeScripting type="romaji"> yasuku, renka
              de</AlternativeScripting>
          </TranslationCtn>
        </TranslationBlock>
      </Locale>
    </TranslationCluster>
    <ExampleCtn type="sid" version="1">
      <Example>vender barato</Example>
      <TranslationCluster identifier="TC00022904" text="vender barato" type="
        exmp">
        [...]
        <Locale lang="ja">
          <TranslationBlock>
            <TranslationCtn>
              <Translation>安くる、安りをする</Translation>
              <AlternativeScripting type="kanji"></AlternativeScripting>
              <AlternativeScripting type="romaji">yasuku uru, yasu uri o
                suru</AlternativeScripting>
            </TranslationCtn>
          </TranslationBlock>
        </Locale>
      </TranslationCluster>
    </ExampleCtn>
    [...]
  </SenseGrp>
</DictionaryEntry>

```

Example 5.20: An extract of the Spanish entry *barato* ‘cheap, inexpensive’, with translations into Japanese

5.5.2 The K Dictionaries custom vocabulary

As we have shown, some elements and structures in the original XML do not have a counterpart in *OntoLex-lemon* or *LexInfo* allowing for their representation in RDF: usage examples, translation relations among examples, identifiers of entries, nested dictionary entries, aspects related to the display of the data in the interface or in (hierarchical) groupings, etc., and other distinctive elements of the KD data. This is due to the fact that *OntoLex-lemon* was not developed to model a dictionary with all its lexicographical annotations, but to lexicalise an ontology. Furthermore, these dictionaries have been compiled with the human as final user as well as for their use by NLP applications. Besides the above-mentioned elements, the KD vocabulary proposed as extension includes classes, individuals and properties that could not be directly mapped to *LexInfo* or *OntoLex-lemon*, primarily for two reasons: (1) mismatches between the DTD values of a tag and *LexInfo* classes (e.g. predefined values of a tag in the DTD are individuals from a class in *LexInfo* that is not compatible with the tag in KD), and (2) different level of granularity in the predefined values in the DTD and the individuals in *LexInfo* (e.g. `kd:possessive` vs `lexinfo:possessiveAdjective`, `lexinfo:possesiveDeterminer`, etc.). Given that in these cases a one-to-one mapping from KD into *LexInfo* was not viable, new elements had to be created under the KD namespace (private, for internal use at KD). Example of these KD tailored elements are `kd:GeographicalUsage`, `kd:neuter-masculineGender`, and `kd:TranslationExampleCluster`.

The complete URI naming strategy adopted for the application profile of *OntoLex-lemon* to KD's data is summarised in Table 5.12.

Entity	URI pattern	Examples
base uri (proposal)	<code>http://kdictionaries.com/id/</code>	
lime:Lexicon	<code>"lexicon" + caps(language code)</code>	<code>:lexiconEN</code> , <code>:lexiconES</code> , etc.
kd:Dictionary	<code><KD Dictionary Series> + "_" + caps(language code) + <version> + "_" + <version></code> (this follows the naming strategy of the XML files)	<code>:mlds-ES3-10</code>
ontolex:LexicalEntry	<code><Lexicon> + "/" + headword + "_" + part of speech</code>	<code>:lexiconEN/table-n</code> , <code>:lexiconES/bueno-adj</code> , <code>:lexiconDE/mit-prep</code>
ontolex:LexicalEntry with homograph number and same part of speech as other of its homographs	<code><Lexicon> + "/" + headword + "_" + part of speech + value of HomNum</code> or <code>HomographNumber</code>	<code>:lexiconEN/bow-n-3</code>
Artificial ontolex:Lexical Entry in cases of homographs	<code><Lexicon> + "/" + headword + "_" + part of speech</code>	<code>:lexiconEN/bow-n</code>
ontolex:Form	<code><LexicalEntry> + "_" + form</code>	<code>:lexiconEN/table-n-form</code> , <code>:lexicoDE/mit-prep-form</code>
ontolex:Form>1	<code><LexicalEntry> + "_" + [gender] + "_" + [number] + "-" + form</code> . If no specific value defined, use numbers: <code><LexicalEntry> + "-" + form-1"/"-form-2"</code> , etc.	<code>:lexiconES/bueno-adj-form-1</code> , <code>:lexiconES/bueno-adj-form-2</code> (because no gender information is provided) <code>:lexiconEN/child-n-sg-form</code> , <code>:lexiconEN/child-n-pl-form</code>
kd:Display	<code><Form> + "-" + display</code>	<code>:lexiconES/bueno-adj-form-1-display</code>
ontolex:LexicalSense (identifier-available)	<code><LexicalEntry> + "_" + SenseGroupIdentifier + "-" + sense</code>	<code>:lexiconES/abadia-n-SE00000039-sense</code>

Table 5.12: URI naming strategy for KD Multilingual global series RDF

Entity	URI pattern	Examples
ontolex:LexicalSense (created artificially from other sense)	:Lexicon>+ "/" + {synonym antonym translated word compositionalphrase} + "_" + part of speech + "-" + <LexicalSense>of source sense	:lexiconNL/abdiij-n-abadia-n-SE00000039-sense
skos:Concept	SenseGroupIdentifier + "-concept"	:SE00000039-concept
vartrans:TranslationSet	"transSet" + caps(source lang code) + caps(target lang code)	:tranSetEN-NL
vartrans:Translation	<Translation Set>+ "/" + <LexicalSense>(source) + "_" + <Lexical Sense>(target) + TranslationClusterIdentifier + "-trans"	:tranSetEN-NL/abadia-n-SE00000039-sense-abdiij-n-abadia-n-SE00000039-sense-TC00000097-trans
vartrans:SenseRelation (but not a Translation)	<LexicalSense>(source) + "_" + <LexicalSense>(target) + "-sr"	:lexiconES/arriesgado-adj-SE00007336-sense-azaroso-adj-arriesgado-adj-SE00007336-sense-sr
kd:UsageExample	<LexicalSense>+ "_" + TranslationClusterIdentifier + "-ex"	:lexiconES/abdiij-n-abadia-SE00000039-sense-TC00000098-ex
kd:TranslationExampleCluster	TranslationClusterIdentifier + "-trans-ex-cl"	:TC00000098-trans-ex-cl
kd:HomographGroup	<LexicalEntry>+ "-hg-" + HomNum	:lexiconEN/bow-n-hg-3
kd:InflectionBlock	<LexicalEntry>+ "-inf-block". If there is more than one inflection block, use numbers: "<LexicalEntry>+ "-inf-block-1".	:lexiconEN/Besucher-n-inf-block-1
"artificial" skos:Concepts created for CompositionalPhrases, antonyms, etc.	<LexicalSense>but replace "sense" ending with "-concept"	:earn/%20a%20living-earn-v-SE000006116-concept

URI naming strategy for KD Multilingual global series RDF (continuation)

5.5.3 RDF Generation and Linking

Provided a table of mappings that collects modelling solutions as the ones proposed throughout these sections, Semantic Web Company was in charge of the development of the pipeline from the XML to RDF serialisation by means of the software Pool Party¹⁶⁰, in particular, UnifiedViews.¹⁶¹ The process is detailed in Turdean and Joshi (2017). As the authors explain, the process of converting the files from their original XML format to RDF followed a pipeline in UnifiedViews that involves the transformation of the files with a XSLT created on the basis of the proposed model. The resulting RDF is then uploaded to Virtuoso Universal Server. The evaluation performed for these steps is detailed in Chapter 7. After the conversion performed by Semantic Web Company, the task of generating the RDF was later overtaken by KD in order to transform the data into JSON-LD. The evaluation of this conversion was performed at KD with a JSON-Schema (described in Bosque-Gil et. al 2019 (to appear)) and with SPARQL queries as the ones described in Chapter 7.

5.6 The lexicalisation of Zhishi.me into Zhishi.lemon

The lexicalisation of Zhishi.me, the encyclopaedic resource gathering RDF triples from the three largest Chinese encyclopedic Web sites (Baidu Baike, Hudong Baike and Chinese Wikipedia) (see Section 4.4) was a challenging task because it involved the application of *lemon* to different dialectal variants of Chinese, as well as the representation of translations, transliterations (pin-yin data), and links to disambiguation pages that were included in the original resources.

Example 5.21 shows the entry *divorce*, where a sample input label query for “divorce” (离婚) allows to merge the content from the different sites into one page that gathers all the data that the representation model needs to address. For the lexicalization, the abstract is not modelled as belonging to the lexicon, since it is encyclopaedic information beyond purely lexical content. The focus was drawn to the lemma of the entry in Madarin Chinese, alternative forms in simplified Chinese, its pin-yin transliteration, translations to English, Spanish, a series of definition or explanations, and a list of internal links and metadata.

¹⁶⁰<https://www.poolparty.biz/>

¹⁶¹<https://unifiedviews.eu/>

Figure 5.24: RDF representation of the monolingual information provided on the entry 离婚

Following the monolingual modelling of previous resources explained in this chapter, Figure 5.24 shows the application profile of *lemon* to this data. The entry is linked to as many senses as disambiguation options, with two senses being linked to the entry, each with a different reference coming from Baidu. Note that there is an element that represents the disambiguation page and links directly through `zhishi:pageDisambiguates`. With regard to the morphological description, the lemma of the entry is captured with `lemon:LexicalForm` distinguishable as *canonical*, and with a language tag indicating Mandarin Chinese. The figure shows a different form in Latin alphabet, which stands in a romanisation relation with the Mandarin one. In addition, simplified Chinese is also accounted for by instantiating a different form which differs from the canonical. The language tags follow the pattern: language-extlang-script-region-variant-

`extension-privateuse` as defined by the Best Current Practice 47.

Because Zhishi.lemon was linked at the sense level with BabelNet to obtain translations, the sense from BabelNet is also represented as a sense of the entry, but it stands in an equality relation with other senses coming from Baidu. Translations, represented in Figure 5.25, are modelled with the *lemon* translation module following the same approach adopted in Apertium RDF (Section 5.2).

Figure 5.25: Translations for the entry “divorce” (离婚) represented as RDF with the *lemon* translation module

5.7 Summary

This chapter reported on the development of application profiles of *lemon* and *OntoLex-lemon* for six language resources whose features were analysed in Chapter 4: the Apertium family of dictionaries, Rayfield’s Comprehensive Georgian-English Dictionary (only nominal entries), the multilingual terminologies Termcat and Terminesp

(three domain lexica), the Spanish, English and German datasets from KD's multilingual Global series, and Zhishi.me. We detected two types of cases: (1) cases in which OntoLex-*lemon* meets representation needs of the lexicographic resources but there are no guidelines on its use to cover those specific needs, and (2) cases in which the sole application of OntoLex-*lemon* falls short of covering structural aspects of dictionary entries and common annotations specifying the use of senses and usage examples. For the first case, we proposed application profiles of the model and LexInfo, while for the second case we suggested the creation of *ad hoc* elements to preserve the information in the original resource. Throughout this process, we defined a URI naming scheme for each resource, on the basis of the naming strategy adopted for Apertium RDF 5.1. Chapter 6 takes the experience reported in this chapter as a foundation for the analysis of further lexicographic sources and the subsequent motivation of a module for lexicography under the umbrella of the W3C Ontology-Lexica Community Group.

zhwiki: 离婚

hudong: 离婚

baidu: 离婚 (脱离夫妻婚姻关系)

Index

- **zhishi:abstract**
- **infobox**
- **dcterms:subject**
- **zhishi:relatedImage**
- **zhishi:internalLink**
- **zhishi:externalLink**
- **zhishi:relatedPage**

zhishi:abstract

离婚是指夫妻双方通过协议或诉讼的方式解除婚姻关系,终止夫妻间权利和义务的法律行为。按照我国《婚姻法》的规定,如感情确已破裂,调解无效,应准予离婚。夫妻“感情确已破裂”是判决离婚的法定条件。解除婚姻关系,后专指通过法律程序解除夫妻关系。《晋书愍怀太子遹传》:“初,太子之废也。妃父王衍表请离婚。”南朝宋刘义庆《世说新语贤媛》:“贾充前妇是李丰女,丰被诛,离婚徙边。”宋陆游《老学庵笔记》卷三:“郭知运以离婚为逐客。”曹禺《北京人》第二幕:“他,他疯了,他要离婚。”印度法院准许一对夫妻用网络电话办理离婚,他们因而成了全球第一对以网络电话离婚的夫妻。

>>TOP

infobox

外文名	divorce;get a divorce;break a marriage
拼音	li hun
解释	解除婚姻关系
法定条件:	感情确已破裂
解释:	解除婚姻关系
中文名	离婚
法定条件	感情确已破裂
外文名:	divorce

Example 5.21: The merged results for the query 离婚 ‘divorce’ with content extracted from Baidu Baike, Hudong Baike and Chinese Wikipedia.

Chapter 6

lexicog: A Lexicography Module for OntoLex-*lemon*

Lexicographic resources are increasingly being transformed into LLD relying on the *lemon* and OntoLex-*lemon* models, but this conversion is not always straightforward and numerous efforts resort to the creation of *ad hoc* classes to cover a range of information in the RDF representation of the resource (see Chapter 2). Chapters 4 and 5 showed that the activities of representing **structural aspects** from lexicographic resources, as well as common **lexicographic annotations** and data provided in dictionary entries reflect a **lack of both guidelines** on how to develop application profiles of the *de facto* standards and **elements** to encode these data.

Chapter 5 proposes modelling practices to cover the representation requirements of a series of representative resources as LLD. In this chapter, such proposals serve as basis to generalise over the resources addressed in previous chapters, and to motivate and lead the creation of a module for lexicography under the umbrella of the W3C OntoLex Community group. A document issued by the community group gathering guidelines for the instantiation of the model with lexicographic resources is motivated as well in this section, with its development expected to start during the second half of 2019.

This chapter is divided in three sections. Section 6.1 motivates the development of a module for lexicography by relying on the experiences and analysis reported in Chapters 4 and 5. We provide examples of gaps and mismatches between additional lexicographic resources, besides those already discussed, and the OntoLex-*lemon* model. Section 6.2 presents the *lemon* lexicography module *lexicog* as developed by the On-

toLex community and also as a result of the work on analysing and modelling resources attained in this thesis. The last section of this chapter, Section 6.3, presents the application and validation of the *lexicog* module with the KD’s Global series in order to demonstrate how the modelling difficulties presented in Chapter 4 and solved *ad hoc* in Chapter 5 are resolved by the standardised mechanisms provided by the new module.

6.1 Motivation for a module for lexicography and specific guidelines to represent lexicographic data as LLD

Chapter 2 showed that there have been several reports in the literature on the conversion of dictionaries to LLD, most of them relying on *lemon* or *OntoLex-lemon*. The added value of using LD technologies in lexicography and its implications for the micro- and macrostructure of dictionaries has been explored as well (details on this are provided in Chapter 7, based on Bosque-Gil et al. (2016a)). However, proprietary formats, such as that of KD, often have XML tags used in their annotation schemes that refer to linguistic categories or features absent in available repositories of linguistic categories, or which lack a compatible definition that prevents us from reusing external ontology entities.

In addition, structures typically found in dictionaries, such as the sense and sub-sense hierarchy or groupings in an entry, are not trivial to model. This leads to the creation of *ad hoc* vocabularies to migrate specific content and structures (Abromeit et al., 2016b; Bosque-Gil et al., 2016b; Declerck and Wandl-Vogt, 2015; Khalfi et al., 2016; Khan et al., 2016b; Klimek and Brümmer, 2015), as well as to the development of alternatives to *OntoLex-lemon* to represent dictionary content, as the Oxford Global Languages Ontology (OGL) (Section 2.3.2).

These works approached issues which affect, for example, the relation between a lexical sense and the lexicalized phrases and idioms in which it occurs, regional restrictions, lexical and semantic selection (in general) of lexical entries, groups of homographs, examples and translations of examples, as well as alternative forms of those translations, tone and register indications, inflection groups, context of use, frequency modifiers to register, temporal information, etymological data, etc.

We take as reference the problems reported in the literature, and those detected in Chapter 4 and addressed in Chapter 5, as basis for the motivation of a module for lexi-

cography for *OntoLex-lemon* as *de facto* standard for the representation of LLD. To abstract from the specific resources converted in this thesis, we analysed further dictionary entries in e-dictionaries of English (Oxford Dictionaries (2017), Merriam-Webster Dictionary On-line (2017), American Heritage Dictionary Online (2017), Collins Cobuild Advanced Learner’s English Dictionary (2003)), German (Duden (2017), PONS (2017)), and Spanish (Real Academia Española - DLE (2014), SM Dicionarios - CLAVE (2017)).

In this process, we do not target the representation of resource-specific features of particular dictionaries, but of common features found across them. We present the main issues we gathered, which **either pose problems for the modelling with *OntoLex-lemon***, and thus call for the definition of a new module to account for them, or **reflect the lack of guidelines** to represent information common in lexicographic records **with *OntoLex-lemon*** elements.

We ground our proposal for a lexicography module on the following four points:

1. The use of *OntoLex-lemon* by the majority of the community to convert linguistic resources to LLD instead of to lexicalize ontologies,
2. the nature of *lemon* being descriptive but not prescriptive and the respect towards different lexicographic views,
3. the coming together of the lexicography and the Semantic Web communities and potential benefits that LLD may bring about to lexicography, assuming it minimises lexical information loss, and
4. the reuse of already available mechanisms in *OntoLex-lemon*.

6.1.1 Issues

In the following we report on the main issues we have encountered after our experiences in converting lexicographic resources to LLD, explained in Chapters 4 and 5, and after our analysis of additional dictionary entries in the English, German and Spanish dictionaries mentioned above. These issues can be broadly divided in three groups:

1. Detected incompatibilities of the model if applied to (certain) lexicographic resources (i.e., aspects that cannot be easily modelled with the current elements) (**Type 1**, T1). For example, these are cases in which applying *OntoLex* imposes a different view on the data than the one provided in the original resource.

2. Missing entities in *OntoLex-lemon* or existent catalogues (e.g., LexInfo) to account for specific lexicographic information (**Type 2**, T2)
3. Best practises when applying *OntoLex-lemon* and existent modules (i.e.: best ways of modelling certain data with the existent ontology elements) (**Type 3**, T3)

These issues were raised and discussed as part of the Ontology-Lexica Community Group in its bi-weekly telcos¹⁶², and are reported in Bosque-Gil et al. (2017). They aim at motivating the need for a module in *OntoLex-lemon* to address specific gaps of the current model, as well as to justify the creation of a series of guidelines on how to use the current elements of the model to account for lexicographic information.

Issue 1 (T1). Headwords that can take different parts-of-speech

Both *lemon* and *OntoLex-lemon* specify a lexical entry as “*a word, a multiword expression or an affix with a single part-of-speech, morphological pattern, etymology and set of senses*”.¹⁶³ While the series of lexicographic resources addressed in the previous chapters of this thesis do not pose problems in the mapping from dictionary entries or terminological records to instances of `ontolex:LexicalEntry` in this respect, a headword in a lexicographic resource may occur with different parts-of-speech depending on context, and its senses are nonetheless defined in the same record if all of them derive from the same etymology. Examples of this are *poison*, *bread*, *water* (noun and verb), Sp. *lento* ‘slow, slowly’ (adjective and adverb), or Sp. *alto* ‘tall, loudly, height’ (adjective, adverb and noun). Applying *OntoLex-lemon* would lead to the generation of several `ontolex:LexicalEntry` elements, one per each part-of-speech with which the headword is attested. Splitting the dictionary entry into several lexical entries would cause loss or duplication of information (e.g. shared etymology, pronunciation, senses implicitly related) and would not keep track of the original representation in the source data.

¹⁶²<https://www.w3.org/community/ontolex/wiki/Lexicography>

¹⁶³<https://www.w3.org/2016/05/ontolex/#lexical-entries>

Issue 2 (T2). Lexical sense requiring a particular form

Some senses of a lexical entry require a particular form, e.g. in English a plural form or in Spanish a masculine or a feminine one. This is the case of *refreshment(s)*, and Sp. *cometa* (m.) ‘comet’, (f.) ‘kite’. In these examples, the dictionary entry can be a single one (e.g. *refreshment* in English or *cometa* in Spanish) but one of its senses indicates a preferred form: in the case of *refreshment*, the plural form is used if the intended meaning is snacks and beverages; with Spanish *cometa*, the feminine form is applicable when referring to a kite, the masculine when denoting a comet. Since the meaning in these cases is associated with the form and it may differ significantly from other senses, in contrast to the related senses mentioned in the previous issue, splitting the dictionary entry into different lexical entries would be an option. However, this splitting causes the situation discussed in Issue 1 with respect to either duplication of information or loss of it, if it is not propagated to both lexical entries. Further examples are *good(s)*, *manner(s)*; Sp. *frente* (m.) ‘front’, (f.) ‘forehead’.

Issue 3 (T2). Usage examples and their translations

Chapters 4 and 5 showed that usage examples of a word or multi-word expression are often provided for each of its definitions. LexInfo includes a property `lexinfo:senseExample` to describe an example of a sense (as a sub-property of `lemon:definition`) and which is linked to the example data category in ISOCat, nowadays DatCatInfo. Nonetheless, due to it being a datatype property, it does not enable including further information on the example or to establish translation relations among examples, which is common practice in bilingual and multilingual dictionaries. The *lemon* model included a `lemon:UsageExample` class and a property `lemon:example` to link to it, but OntoLex-*lemon* does not cover this aspect yet. Examples: Sp. *Preocuparse* ‘worry’; Sp. *no hay por qué preocuparse* ‘there is nothing to worry about’ (Collins English-Spanish Dictionary).¹⁶⁴

Issue 4 (T2). Sense order

The order of senses in a dictionary entry may be based on frequency of use, date of origin, concreteness (from the most concrete to most abstract sense, etc.). Their

¹⁶⁴“worry”. Intransitive Verb. Def. 1, Example 5. Collins English-Spanish Dictionary. <http://www.collinsdictionary.com/dictionary/english-spanish/worry>. Last accessed 13/06/2019.

order should be searchable and retrievable as to recover the information provided in the original resource. Examples: *Boa*: noun. (1) any of a family (Boidae) of large snakes that kill by constriction and that includes the boa constrictor, anaconda, and python (2) a long fluffy scarf (Merriam Webster Dictionary).¹⁶⁵ However, different descriptions of the same `ontolex:LexicalEntry`, e.g. *biscuit*, can show different order of senses depending on the target audience (e.g. American or British) and hence it is debatable whether sense order should be encoded at the sense level itself (as an inherent property of the sense) or at a structural layer.

Issue 5 (T2). Geographical information.

One of the lexical senses of a lexical entry might be attested in a specific region. For instance, Sp. *manejar* means ‘to drive’ in Venezuela and ‘to handle’ in Spain; while *braces* in the sense of “suspenders” is only attested in British English (in relation the plural form, see Issue 2). Lexvo Ontology includes the property `lexvo:usedIn`, with range `lexvo:GeographicRegion`, and described as “*The property of a language or writing system being used somewhat extensively in a particular geographical region at some point in time*”.¹⁶⁶ LexInfo and OntoLex-lemon offer elements to represent geographical variants, but there is not any guideline on how to record geographical information with respect to a sense in the model independently of other entries or forms.

Issue 6 (T3). Dictionary Senses as Lexical Senses or as Lexical Concepts.¹⁶⁷

As *lemon* is intended to be used to lexicalize an ontology, the semantic layer is assumed to be provided by the ontology. However, in the conversion of lexicographic resources, we lack the ontological layer and the description of the meaning is described in the resource through sense groups, containers, or sections. With the increasing amount of lexicographic resources in LLD, there still lacks a widespread practice addressing whether dictionary senses should be modelled as lexical concepts or lexical senses: there

¹⁶⁵“boa”. n. Merriam Webster Dictionary. <http://www.merriam-webster.com/dictionary/boa>. Last accessed 13/06/2019. Example of logical order of senses inspired by Diccionario de la Lengua Española, Guía de Consulta, <http://dle.rae.es/>.

¹⁶⁶<http://lexvo.org/ontology>

¹⁶⁷This issue was detected on the basis of an example provided by Francesca Frontini for the modelling of the dictionary Petite Larousse Illustré, 1905 edition.

is nothing in the dictionary leading to assume an external ontology entity to which dictionary senses refer. If translations or lexico-semantic restrictions are to be represented, `ontolex:LexicalSense` is the class used in those cases, but it requires an ontological reference that we lack in a lexicographic resource.

Issue 7 (T3). Representation of complex forms, idioms, etc. and their relation to the lexical entry and/or lexical sense.¹⁶⁸

Most dictionaries include a set of complex forms related to the entry, usually compounds, idioms, collocations, etc. These sometimes are defined in relation to a particular sense (see Chapter 5), sometimes are encoded as sub-senses of a sense (OED, s. *fool* – *fool around*¹⁶⁹), but very frequently they are defined under the same lexicographic record, following the description of senses. While there is a practice on how to model morphological decomposition (*decomp*), there is no documentation on how to represent this very frequent data with the OntoLex-*lemon* model in the context of a dictionary record.

Issue 8 (T3). Semantic selection

Some lexicographic resources indicate the semantic features of the lexical items that an entry (in one of its senses) selects or even the exact lexical items with which it collocates. This is usually indicated either with a specific tag (e.g. KD’s tag `RangeOfApplication`, see section 4.3), or in-between parentheses at the beginning of a definition. Some examples are the dictionary entry for the German verb *dämmen*, which in its sense ‘to insulate, absorb, mute’ selects arguments that denote warmth or sound (German *Wärme*, *Schall*, etc.) (KD)¹⁷⁰; the adjective *cozy*, meaning *beneficial to all those involved and possibly somewhat corrupt* if predicated from a transaction or an arrangement (Google Dictionary)¹⁷¹; or the collocational measure words of *luck*: *stroke*, *piece of* (Oxford Collocations Dictionary).¹⁷² The `synsem:OntoMap` class allows to map a syntactic frame to an ontology entity, so that the frame and its arguments are linked to

¹⁶⁸This issue was detected on the basis of an example provided by Francesca Frontini for the modelling of the dictionary *Petite Larousse Illustré*, 1905 edition.

¹⁶⁹*fool*. LEXICO provided by Oxford Dictionaries. <https://www.lexico.com/en/definition/fool>. Las accessed 04.09.2019

¹⁷⁰“dämmen”. Sense 2. K Dictionaries Global Series Multilingual German Dictionary

¹⁷¹“cosy”. Def. 1., sub-sense 3. Google Dictionary. Last Accessed 13/07/2017.

¹⁷²“luck”. QUANT. Deuter, M. (Ed.). (2008). Oxford Collocations Dictionary for students of English. Oxford University Press, p. 473.

the ontology elements that they lexicalize. Even though dictionaries commonly include information on subcategorization (transitive/intransitive/reflexive etc. annotations for verbs, for instance), details on the syntactic frame are not always provided beyond those annotations. Since in dictionary conversion we often lack a given ontology and detailed syntactic information is not provided, the mapping between syntactic arguments and ontology entities seems difficult to establish automatically via `synsem:OntoMap`: if the morphosyntactic information provided in a dictionary is limited to the fact that *cozy* is an adjective and *luck* a noun, representing that *cozy* has a meaning only applied to *transaction* or *agreement*, and that the measure words that collocate with *luck* are *stroke* and *piece* is not trivial. Furthermore, `synsem:condition` (in its turn subsuming `synsem:propertyRange` and `synsem:propertyDomain`) enables to state constraints on the arguments of a predicate in a given ontology.¹⁷³ The possibility of reusing it to state the constraints on syntactic arguments even in cases in which we lack a given ontology, and therefore are not mapping to given ontology properties, deserves thus further consideration in the OntoLex community.

The OntoLex Lexicography Wiki page lists more issues detected by the community with regards the representation of lexicographical data from lexicographical resources, e.g. thesauri, which lead to the proposal of *lemon-Tree* (Stolk, 2019), a vocabulary to address onomasiological orderings, among other aspects specific to thesauri which are out of the scope of this thesis.

In this section we have drawn attention to a series of issues raised in the literature on LLD related to the conversion of lexicographic resources to LD already mentioned in Chapter 2, and to eight of the ones we encountered during the analysis and modelling of the resources presented in previous chapters, along with further dictionaries of English, German and Spanish. We argue that the OntoLex-*lemon* model **should enable the preservation** of the content and the structure of the original resource, even if the LLD expert opts for a different representation more suited to the exploitation of the data by external applications or more in line with their view on the lexicon-ontology interface.

¹⁷³<https://www.w3.org/2016/05/ontolex/#conditions>

6.2 *lexicog*: a module for lexicography

The content of this section is extracted and adapted from the report that describes the lexicography module of the Lexicon Model for Ontologies (*lemon*) as a result of the work of the Ontology Lexicon community group (OntoLex).¹⁷⁴ The module is targeted at the representation of lexicographic resources, and addresses structures and annotations commonly found in lexicography. This module operates in combination with the OntoLex-*lemon* core module, OntoLex.

After analysing the literature (see Chapter 2) and the challenges we mentioned and addressed in previous chapters, we perceived a strong need for reaching some agreement that allows for a better and more inter-operable migration of existing dictionaries into LD. Interoperability is a key issue in LD technologies, thus building a common space in which these concepts can be agreed upon and commonly defined is a logical next step. The OntoLex community is the natural forum to accomplish this for several reasons:

1. The extended use of *lemon* or OntoLex-*lemon* to convert language resources (including lexicographic resources) into LD, their nature being descriptive but not prescriptive (which facilitates neutrality towards different lexicographic views),
2. the coming together of the lexicography and the Semantic Web communities, and
3. the possibility of reusing already available mechanisms in OntoLex-*lemon*, preventing researchers from “re-inventing the wheel.”

6.2.1 Aim and scope

The main goal of this module is to complement the OntoLex-*lemon* core module, OntoLex, and to overcome its limitations when modelling lexicographic information as LD in a way that is agnostic to the underlying lexicographic view and aims for information preservation.

The scope of the model is two-fold:

1. Modelling existing lexicographic resources as LD, to allow their later publication by following Semantic Web standards, and

¹⁷⁴Available at <http://jograccia.github.io/ontolex-lexicog/> at the time of writing, and expected to be published at <http://www.w3.org/ns/lemon/lexicog>.

2. providing a conceptual / abstract model of language and linguistic objects in lexicography that will allow for building LD-based lexicographic resources from scratch (more on LD-native resources in Chapter 8)

Modelling existing dictionaries as LD can be done by first extracting the features they have in common and modelling them in a way that is decoupled from specific linguistic views, which is the approach followed in Chapter 5. Although the primary goal might be to model existing resources, the module should remain highly re-usable across dictionaries and useful for new LD-based lexicographic works developed from scratch. Naturally, we assume that many specific ingredients from specific lexicographic resources and views will still need *ad hoc* constructs. Still, having a small set of common entities defined in the module will help to overcome the limitations that have been detected in the experiences of instantiating *OntoLex-lemon* with lexicographic data, and can greatly promote interoperability.

This module adds some complexity by providing additional description capabilities to the purely lexical description accounted by *OntoLex*. If this information is not needed for a specific conversion, i.e, if the lexicographic view is not essential, reusing *OntoLex* would allow to keep the representation simpler but yet sufficient.

Figure 6.1 depicts the *OntoLex* lexicography module (*lexicog*). Boxes represent classes of the model. Arrows with filled heads represent object properties, while arrows with empty heads represent the `rdfs:subClassOf` relation. Note that the module provides elements to account for structures common in lexicographic resources (on the upper part of the figure) as well as some to capture annotations and additional information describing lexical elements as found in dictionaries (on the bottom part).

Lexicographic Resource. The class `lime:Lexicon` has been used in the literature as a natural way of modelling a collection of lexicographical records. However, there are certain situations in which there is no one-to-one mapping between the entries in a dictionary, or lexicographic resource in general, and the lexical entries in a *lemon* lexicon (`lime:Lexicon`). For instance, some dictionary entries contain information about their translations and synonyms, though the latter might not have their own dictionary entry in the source dictionary. Still, such translations and synonyms can be treated as ‘first class citizens’ in the RDF graph by representing them as instances of `ontolex:LexicalEntry` ?? . In such situations, there is a disparity between the ensuing

Figure 6.1: The OntoLex lexicography module (*lexicog*). The upper part accounts for structures common in lexicographic resources and the bottom part describes the lexicon, i.e., the elements describing lexical information.

lemon lexicon and the original dictionary, as the LD representation would include entries not encoded in the original resource during the headword selection process.

The class `LexicographicResource` is intended to represent the original resource in order to keep track of the collection of lexicographic entries contained in it. This does not replace, but complements, the use of `lime:Lexicon`, which will include the lexical entries explicitly declared in the resource along with other ones (e.g., synonyms, translations, multiword expressions, etc.). Thus, a `LexicographicResource` represents a collection of lexicographic entries (`lexicog:Entry`) in accord with the lexicographic

criteria followed in the development of that resource.

Entry. The class **Entry** is introduced to represent the lexicographic record as it is encoded in the resource, e.g. a dictionary entry, thereby reflecting the arrangement decided upon by the lexicographer (supported by the occurrences of the word or phrase in the corpus used for dictionary compilation). This class is intended to be the counterpart of `ontolex:LexicalEntry` in the LD-representation of a dictionary, and fulfils a structural function. It is a structural element, sub-class of `LexicographicComponent` (see below) that represents a lexicographic article or record as it is arranged in a source lexicographic resource. As such, it supports the description of lexical entries or senses according to the lexicographic microstructure, decided upon during the compilation process of the dictionary,

entry. The property **entry** relates an instance of the class representing the lexicographic resource to the actual entry. This property is a counterpart of `lime:entry`, which relates instances of `lime:Lexicon` to `ontolex:LexicalEntry` elements. A simple example of an entry belonging to a lexicographic resource in English is given in Example 6.1. Here, a lexicographic record for the word *animal* groups (under the same dictionary entry) descriptions for the lexical entries *animal* (n.), and *animal* (adj.), which both share phonetic representation and etymology. The example has been extracted from the American Heritage Dictionary.¹⁷⁵

In this case, a dictionary provides one single dictionary entry for the word *animal*, comprising its noun and adjective definitions under one same structure. This is not compliant with the definition of `ontolex:LexicalEntry`, which requires a single part-of-speech. Therefore, a `lime:Lexicon` accounting for the lexical information in this dictionary will contain two lexical entries, one per part-of-speech. Example 6.2 shows that an instance of the `LexicographicResource` class will gather entities as they were presented in the dictionary, in order to keep track of the original representation. In this simplified case, *animal* will be the only entry of the lexicographic resource.

¹⁷⁵*animal*. <https://www.ahdictionary.com/word/search.html?q=animal>. Last accessed on 04.09.2019.

an·i·mal

n.

1. *Any of numerous multicellular eukaryotic organisms of the kingdom Metazoa (or Animalia) [...]*
2. *An animal organism other than a human, especially a mammal. [...]*

adj.

1. *Relating to, characteristic of, or derived from an animal or animals, especially when not human: animal cells; animal welfare.*
2. *Relating to the physical as distinct from the rational or spiritual nature of people: animal instincts and desires.*

Example 6.1: Extract of the dictionary entry *animal* from the American Heritage Dictionary

```
@prefix base: <http://example.org/> .
@prefix lime: <http://www.w3.org/ns/lemon/lime#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix lexicog: <http://www.w3.org/ns/lemon/lexicog#> .
@prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#> .
@prefix dc: <http://purl.org/dc/terms/> .

:myDictionary a lexicog:LexicographicResource ;
  dc:language "en" ;
  lexicog:entry :animal_entry .

:myLexicon a lime:Lexicon;
  lime:language "en" ;
  lime:entry :animal_n, :animal_adj .

:animal_entry a lexicog:Entry .
:animal_n a ontolex:LexicalEntry .
:animal_adj a ontolex:LexicalEntry .
```

Example 6.2: The distinction between a `lexicog:Entry` and a `ontolex:LexicalEntry` on the example of the lemma *animal* from the American Heritage Dictionary

Lexicographic Component. A `LexicographicComponent` is a structural element in the lexicographic resource that represents either a lexicographic record or any other structure to refer to senses, sense groups, or nested entries. Lexicographic components do not necessarily describe lexical data but can often fulfil a grouping function to gather other components that do. `Entry` is a particular subclass of `LexicographicComponent` used to represent the main ‘entry point’ in the dictionary, i.e., the headword or the root of the lexicographic record. These components can describe lexical senses (`ontolex:LexicalSense`) or lexical entries (`ontolex:LexicalEntry`), depending on the arrangement in the original resource. Components can in turn be arranged in a specific order and/or hierarchy or be declared to be part of an entry.

In Example 6.2, two different instances of `ontolex:LexicalEntry` are grouped together in the same lexicographic record (`Entry`). To capture this in our RDF representation, we have to start by stating that the single entry has two lexicographic components. Example 6.3 indicates that the components belong to the entry by simply using the RDF native mechanisms for containers, in particular the `rdfs:member` property.

describes. Since `LexicographicComponent` elements fulfil a structural function to reflect the arrangement of the descriptions provided in a lexicographic resource (order, hierarchy, grouping, etc.), they do not encode lexical content themselves. This is usually accounted for by `OntoLex` elements, which are *described* by the components. This is grounded on the idea that a lexicographic resource describes the lexicon of a language (set of languages, dialect, field of study, etc.) following different criteria, e.g. the intended use of the resource, target audience, lexicographic tradition, etc. (see Section 1.2.2), which can affect the description of the lexical content, leading to, for instance, two different resources describing the same lexicon differently.

```

@prefix base: <http://example.org/> .
@prefix lime: <http://www.w3.org/ns/lemon/lime#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix lexicog: <http://www.w3.org/ns/lemon/lexicog#> .
@prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#> .
@prefix dc: <http://purl.org/dc/terms/> .

:myDictionary a lexicog:LexicographicResource;
  dc:language "en" ;
  lexicog:entry :animal_entry .

:myLexicon a lime:Lexicon;
  lime:language "en" ;
  lime:entry :animal_n, :animal_adj .

:animal_entry a lexicog:Entry .
:animal_n a ontolex:LexicalEntry .
:animal_adj a ontolex:LexicalEntry .

:animal_entry rdfs:member :animal_n_comp, :animal_adj_comp .
:animal_n_comp a lexicog:LexicographicComponent .
:animal_adj_comp a lexicog:LexicographicComponent .

```

Example 6.3: Two different instances of `ontolex:LexicalEntry` grouped together in the same lexicographic record (Entry)

However, in theory, a component could describe any information provided in the resource, e.g. etymology, so the range of the property is not limited to current *OntoLex-lemon* elements. The property `describes` relates a `LexicographicComponent` to an element that represents the actual information provided by that component in the `LexicographicResource`. In most cases, this information will be lexical, and hence the object of the property will be an instance of `ontolex:LexicalEntry` or `ontolex:LexicalSense`.

Turning back to Example 6.1, Example 6.4 shows how the property `describes` can be used to relate the two components belonging to the lexicographic resource with the lexical entries that they describe.

```

@prefix base: <http://example.org/> .
@prefix lime: <http://www.w3.org/ns/lemon/lime#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix lexicog: <http://www.w3.org/ns/lemon/lexicog#> .
@prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#> .
@prefix dc: <http://purl.org/dc/terms/> .

:myDictionary a lexicog:LexicographicResource;
  dc:language "en" ;
  lexicog:entry :animal_entry .

:myLexicon a lime:Lexicon;
  lime:language "en" ;
  lime:entry :animal_n, :animal_adj .

:animal_entry a lexicog:Entry .
:animal_n a ontolex:LexicalEntry .
:animal_adj ontolex:LexicalEntry .

:animal_entry rdfs:member :animal_n_comp, :animal_adj_comp .

:animal_n_comp a lexicog:LexicographicComponent ;
  lexicog:describes :animal_n .
:animal_adj_comp a lexicog:LexicographicComponent ;
  lexicog:describes :animal_adj .

```

Example 6.4: The `describes` property to put in relation the two components belonging to the lexicographic resource with the lexical entries that they describe

In order to provide a complete example, we add the lexical information represented in OntoLex (thus, not specific of the *lexicog* module) in Example 6.5.

subComponent. There are cases in which `LexicographicComponent` instances are arranged in a particular hierarchy: a dictionary entry may include a list of senses (even with sub-senses) or sub-entries. This can be expressed by means of the RDF native mechanisms for containers. In the case that it is necessary to explicitly reflect that this is a hierarchical relation between `LexicographicComponent` elements, the `subComponent` property can be used instead of, or in addition to, such RDF container mechanisms.

Example 6.6, below Example 6.5 shows a revisited version of Example 6.3 to describe that the only entry in our lexical resource has two (embedded) components.

```
@prefix base: <http://example.org/> .
@prefix lime: <http://www.w3.org/ns/lemon/lime#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix lexicog: <http://www.w3.org/ns/lemon/lexicog#> .
@prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#> .
@prefix dc: <http://purl.org/dc/terms/> .

:myDictionary a lexicog:LexicographicResource;
  dc:language "en" ;
  lexicog:entry :animal_entry .

:myLexicon a lime:Lexicon;
  lime:language "en" ;
  lime:entry :animal_n, :animal_adj .

:animal_entry a lexicog:Entry .
:animal_n a ontolex:LexicalEntry .
:animal_adj a ontolex:LexicalEntry .

:animal_entry rdfs:member :animal_n_comp, :animal_adj_comp .

:animal_n_comp a lexicog:LexicographicComponent ;
  lexicog:describes :animal_n .
:animal_adj_comp a lexicog:LexicographicComponent ;
  lexicog:describes :animal_adj .

:animal_n_sense_1 a ontolex:LexicalSense .
:animal_n_sense_2 a ontolex:LexicalSense .
:animal_adj_sense_1 a ontolex:LexicalSense .
:animal_adj_sense_2 a ontolex:LexicalSense .

:animal_n_sense_1_comp a lexicog:LexicographicComponent ;
  lexicog:describes :animal_n_sense_1 .
:animal_n_sense_2_comp a lexicog:LexicographicComponent ;
  lexicog:describes :animal_n_sense_2 .
:animal_adj_sense_1_comp a lexicog:LexicographicComponent ;
```

```

    lexicog:describes :animal_adj_sense_1 .
:animal_adj_sense_2_comp a lexicog:LexicographicComponent ;
    lexicog:describes :animal_adj_sense_2 .

:animal_n_comp rdf:_1 :animal_n_sense_1_comp ;
    rdf:_2 :animal_n_sense_2_comp .

:animal_adj_comp rdf:_1 :animal_adj_sense_1_comp ;
    rdf:_2 :animal_adj_sense_2_comp .

:animal_form a ontolex:Form ;
    ontolex:writtenRep "animal"@en .

:animal_n ontolex:lexicalForm :animal_form ;
    lexinfo:partOfSpeech lexinfo:noun ;
    ontolex:sense :animal_n_sense_1 ;
    ontolex:sense :animal_n_sense_2 .

:animal_adj ontolex:lexicalForm :animal_form ;
    lexinfo:partOfSpeech lexinfo:adjective ;
    ontolex:sense :animal_adj_sense_1 ;
    ontolex:sense :animal_adj_sense_2 .

:animal_n_1_concept a ontolex:LexicalConcept ;
    skos:definition "Any of numerous multicellular eukaryotic organisms of the
        kingdom Metazoa (or Animalia)"@en .

:animal_n_2_concept a ontolex:LexicalConcept;
    skos:definition "An animal organism other than a human, especially a mammal
        ."@en .

:animal_adj_1_concept a ontolex:LexicalConcept ;
    skos:definition "Relating to, characteristic of, or derived from an animal
        or animals, especially when not human: animal cells; animal welfare."@en
        .

:animal_adj_2_concept a ontolex:LexicalConcept;
    skos:definition "Relating to the physical as distinct from the rational or
        spiritual nature of people: animal instincts and desires."@en .

:animal_n_sense_1 ontolex:isLexicalizedSenseOf :animal_n_1_concept .

```



```

:animal_n_sense_2 ontolex:isLexicalizedSenseOf :animal_n_2_concept .
:animal_adj_sense_1 ontolex:isLexicalizedSenseOf :animal_adj_1_concept .
:animal_adj_sense_2 ontolex:isLexicalizedSenseOf :animal_adj_2_concept .

```

Example 6.5: The complete RDF representation of Example 6.1 using *lexicog* and OntoLex

```

@prefix base: <http://example.org/> .
@prefix lime: <http://www.w3.org/ns/lemon/lime#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix lexicog: <http://www.w3.org/ns/lemon/lexicog#> .
@prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#> .
@prefix dc: <http://purl.org/dc/terms/> .

:myDictionary a lexicog:LexicographicResource ;
  dc:language "en" ;
  lexicog:entry :animal_entry .

:myLexicon a lime:Lexicon;
  lime:language "en" ;
  lime:entry :animal_n, :animal_adj .

:animal_entry a lexicog:Entry .
:animal_n a ontolex:LexicalEntry .
:animal_adj ontolex:LexicalEntry .

:animal_entry lexicog:subComponent :animal_n_comp ;
  lexicog:subComponent :animal_adj_comp .

:animal_n_comp a lexicog:LexicographicComponent ;
:animal_adj_comp a lexicog:LexicographicComponent .

```

Example 6.6: Representing hierarchy by means of the `subComponent` property linking instances of `LexicographicComponent`

FormRestriction There are numerous cases in which a lexicographic record specifies the grammatical features the lemma shows when used in some of its senses, as pointed out in Issue 2 in Section 6.1. For example, in English, the lemmas *good*, *manner*, *air* or

air

n. [...]

2. *An impression of a quality or manner given by someone or something.*

[...]

2.1 (*airs*) *An annoyingly affected and condescending manner.* [...]

Example 6.7: An extract of the entry *air* from the Oxford English Living Dictionaries Online which indicates that sub-sense 2.1 is usually attested with the entry in plural form *airs*¹⁷⁶

wood have senses which are usually (or only) attested when the word is used in plural, often indicated with the lemma in plural form, as shown in Example 6.7, or with a usage note in the dictionary entry.

While there are plural nouns which only occur in plural form (e.g. *amends*, *scissors*), the so-called *pluralia tantum*, and they are usually defined in their own dictionary entry, there is also some variance across lexicographic resources in cases in which the noun also occurs in singular form. For instance, *glasses* in Lexico (Oxford Dictionaries) is defined an entry separate from *glass*¹⁷⁷, as opposed to its treatment as a sub-sense of a sense of the lemma *glass* in Merriam Webster Dictionary, with the annotation ‘*glasses plural*’.¹⁷⁸

There are dictionaries which provide grammatical annotations indicating varying gender in the senses of the same dictionary entry as well. For example, in Spanish, the lemma *policía* can be used in feminine or masculine if denoting a police officer, but only in feminine when denoting the police force or administration. Depending on the criteria followed for headword selection during the compilation of the resource, e.g. etymology, in these cases we may find two independent entries or one single entry with senses having different gender annotations.

The **FormRestriction** class is intended to provide a way to specify, from an external

¹⁷⁶ *air*. LEXICO by Oxford Dictionaries. <https://www.lexico.com/en/definition/air>. Last Accessed 04.09.2019.

¹⁷⁷ *glasses*. LEXICO by Oxford Dictionaries. <https://www.lexico.com/en/definition/glasses>. Last accessed on 04.09.2019

¹⁷⁸ *glass*. Merriam-Webster Dictionary. <https://www.merriam-webster.com/dictionary/glass>. Last accessed 04.09.2019.

catalogue of linguistic data categories, the set of grammatical features of a lexical entry when used in a specific `ontolex:LexicalSense` in cases in which it does not allow for all of those reflected in the lexical forms provided. For illustration, in Spanish, *ratón* ‘mouse’, in its meaning of ‘animal’, has a masculine and a feminine form, for male and female mice. In its meaning of ‘computer device’, ‘small rock’ or ‘biceps’ (in Costa Rica), it only allows for a masculine form, and therefore these senses would receive a `FormRestriction`. Thus, the class `FormRestriction` represents (a set of) grammatical features of the instances of `ontolex:Form` attested with that sense.

restrictedTo. The property `restrictedTo` relates an `ontolex:LexicalSense` to a `FormRestriction` when a lexicographic resource provides information about the specific morphological features of the `ontolex:Form` with which that sense is attested.

Example 6.8 (see below) shows the RDF representation of the above mentioned sense of *air*. For the sake of simplicity, we do not account for sense hierarchy in this example, and hence we do not instantiate `LexicographicComponent` here.

UsageExample. There are occasions in which lexicographic records also include an example of use of an entry for each of its senses. While these examples can be captured as strings with the property `skos:example` and linked to instances of `ontolex:LexicalConcept`, there might be more information about the example beyond the string itself. The class `UsageExample` is intended to represent a textual example of the usage of a sense in a given lexicographic record. A usage example can group several string values, in which case they will encode the same meaning. Thus, if such values are expressed in different languages, they can be interpreted as translations. Usage examples can be linked to any other relevant information related to them, from provenance to usage notes.

```

@prefix base: <http://example.org/> .
@prefix lime: <http://www.w3.org/ns/lemon/lime#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix lexicog: <http://www.w3.org/ns/lemon/lexicog#> .
@prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#> .
@prefix dc: <http://purl.org/dc/terms/> .

:air_n a ontolex:LexicalEntry ;
  ontolex:sense :air_n_sense_2, :air_n_sense_2_1 ;
  ontolex:canonicalForm :air_n_form ;
  ontolex:otherForm :airs_n_form .

:air_n_form a ontolex:Form ;
  ontolex:writtenRep "air"@en ;
  lexinfo:number lexinfo:singular .

:airs_n_form a ontolex:Form ;
  ontolex:writtenRep "airs"@en ;
  lexinfo:number lexinfo:plural .

:air_n_sense_2 a ontolex:LexicalSense ;
  ontolex:isLexicalizedSenseOf :air_n_sense_2_lc .

:air_n_sense_2_1 a ontolex:LexicalSense ;
  ontolex:isLexicalizedSenseOf :air_n_sense_2_1_lc ;
  lexicog:restrictedTo :air_n_formRes .

:air_n_formRes a lexicog:FormRestriction ;
  lexinfo:number lexinfo:plural .

:air_n_sense_2_lc a ontolex:LexicalConcept ;
  skos:definition "An impression of a quality or manner given by someone or
    something"@en .

:air_n_sense_2_1_lc a ontolex:LexicalConcept ;
  skos:definition "An annoyingly affected and condescending manner"@en .

```

Example 6.8: The RDF representation of the sense of *air* shown in Example 6.7 using the class `FormRestriction` and the property `restrictedTo` to indicate that sense `:air_n_sense_2_1` is only attested with the headword in plural form

monastery *n* (monk's residence) [English] monasterio *nm* [Spanish]

We visited a Buddhist monastery deep in a jungle.

Visitamos un monasterio budista en medio de la selva.

Example 6.8: Usage examples and their translations in the entry *monastery* from the English-Spanish WordReference Bilingual Dictionary

usageExample. In those cases in which it is necessary to instantiate a `UsageExample` because there is further information about an example beyond the text itself, the `usageExample` property allows us to link an `ontolex:LexicalSense` to a `UsageExample` element.

Example 6.8 shows that a bilingual dictionary such as the English-Spanish Word-Reference Dictionary¹⁷⁹ can include, for a sense, a usage example and a translation of the usage example in the target language. This information is presented as RDF in Example 6.9.

```
@prefix base: <http://example.org/> .
@prefix lime: <http://www.w3.org/ns/lemon/lime#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix lexicog: <http://www.w3.org/ns/lemon/lexicog#> .
@prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#> .
@prefix dc: <http://purl.org/dc/terms/> .

# Entries
:monastery_n a ontolex:LexicalEntry ;
    ontolex:sense :monastery_sense .
:monasterio_n a ontolex:LexicalEntry ;
    ontolex:sense :monasterio_monastery-sense .

# Senses
:monastery_n_sense a ontolex:LexicalSense ;
    ontolex:isLexicalizedSenseOf :monastery_n_sense_concept ;
    lexicog:usageExample :monastery_n_sense_ex .
```

¹⁷⁹*monastery*. English-Spanish Word Reference Dictionary. Last accessed 04.09.19. <http://www.wordreference.com/es/translation.asp?tranword=monastery>

```

:monasterio_monastery-sense a ontolex:LexicalSense ;
    ontolex:isLexicalizedSenseOf :monastery_n_sense_concept .

# Concepts
:monastery_n_sense_concept a ontolex:LexicalConcept ;
    skos:definition "monk's residence"@en .

# Translations
:monastery_n_sense-monasterio_monastery-sense-tr a vartrans:Translation ;
    vartrans:source :monastery_n_sense ;
    vartrans:target :monasterio_monastery-sense .

# Examples
:monastery_n_sense_ex a lexicog:UsageExample;
    rdf:value "We visited a Buddhist monastery deep in a jungle."@en ;
    rdf:value "Visitamos un monasterio budista en medio de la selva."@es .

```

Example 6.9: RDF Turtle representation of the entry provided in Figure 6.8 instantiating `UsageExample` and the property `usageExample`

With these ten minimal elements inspired by the micro-structure of lexicographic records, *lexicog* aims at providing mechanisms general enough to allow for flexibility in the representation of lexicographic content as RDF while at the same time reflecting the output of those lexicographic decisions taken during the compilation of the original resource.

6.3 Validating the *lexicog* module K Dictionaries' Global series (2019)

In this section we revisit previous work on the conversion of KD's multilingual Global series and report on our experience of validating the *lexicog* module by applying it to KD's data, and thus assessing whether the limitations of *OntoLex-lemon* detected in Chapter 4 and addressed *ad hoc* in Chapter 5 are now resolved in a standardised manner. The analysis and results provided here are extracted from Bosque-Gil et al. (2019).

6.3.1 Overview

In the first work in the conversion of KD’s data to RDF (Klimek and Brümmer, 2015), the authors identified some gaps in the *lemon* model to represent KD’s data, for instance, the way to link a compound phrase defined inside of a sense group to that same sense. The lack of an ontology to provide ontological references for `lemon:LexicalSense` instances was also highlighted. This point is strongly connected to the original aim of the *lemon* model to serve to lexicalize ontologies, not to represent lexicographic resources in the Web of Data. In addition, authors identified some gaps in the LexInfo catalog of grammar categories and created their own custom vocabulary to capture the values of KD’s DTD attributes. In the conversion of the multilingual series to *OntoLex-lemon* described in Chapter 5, some problems previously identified in the initial conversion effort in *lemon* did not hold any longer, as both the model and its modules had evolved to cover more cases (for instance, the *vartrans* module allows now to represent lexical relations).

It is worthy to note that in the first two attempts the conversion was carried out under the strict principal of round-tripping, i.e. aiming to obtain full and complete 1:1 data transformation from XML to RDF and from RDF back to XML – so the RDF structure had to convey each and every detail of the complex features of the original XML structure. In contrast, for the last conversion of KD’s data (2019) this restriction was lifted.

Despite that this restriction does not longer hold, Chapter 5 showed that *OntoLex-lemon* proved not exhaustive to cover the representation requirements of the resource. Four kinds of challenging situations were detected in the modelling of KD’s multilingual data in 2016:

1. Cases in which solely applying the *OntoLex-lemon* model would lead to a loss of structural information reflecting lexical distinctions. For example, entries *not* originally conceived as dictionary entries in KD data are treated equally as original entries in the RDF representation (i.e. as `ontolex:LexicalEntry`). This highlighted a lack of elements for representing the components of a lexicographic entry in cases in which there is no 1:1 mapping with *OntoLex-lemon* classes and properties.

In KD data, we encounter several examples of this type of situation: compounds,

synonyms, antonyms, and translations. Compounds are defined inside the dictionary entry as one of its components and do not occur as lemmas (i.e. in their own dictionary entry). Synonyms and antonyms, even though they are usually independent lemmas in that same resource, are embedded in dictionary entries and they do not necessarily have their corresponding record in that resource (but could occur as dictionary entries in another KD dictionary). In addition, a translation of a headword into another language is treated as an `ontolex:LexicalEntry` (see Chapter 5), too, but, as a synonym, the source data in its current state does not guarantee for the word to be a lemma in the dictionary of the target language. This fact required a distinction between an original dictionary entry and the `ontolex:LexicalEntry` newly created in the process, thus recording the outcome of the headword selection step in the compilation of the dictionary. In lexicographic resources other than KD, the same gap would surface in those cases in which a dictionary entry needs to be split into more than one `ontolex:LexicalEntry`, each with a different part of speech, in order to be OntoLex-compliant.

2. Cases in which OntoLex-*lemon* or LexInfo falls short of covering the representation needs that KD's dictionary entries give rise to. This concerned the representation of examples and translations of examples.
3. Cases in which OntoLex-*lemon* does contain elements to cover a particular type of information, but there are no specific guidelines on how to use them in the process of conversion of lexicographic data to RDF (without involving ontology lexicalisation). For example, the representation of lexicographic definitions with the OntoLex core (e.g. with `skos:Concept` or `ontolex:LexicalConcept`), the encoding of geographical usage restrictions on senses, or the modelling of selectional restrictions for predicate arguments.
4. Mismatches between LexInfo elements and KD's DTD tags and values.

Since situations of type (1) and (2) were also generalizable to other lexicographic resources, *lexicog* was proposed as an extension of OntoLex-*lemon*. For cases of type (3), the OntoLex Community, in its bi-weekly telcos on Lexicography, discussed a series of practices for the use of OntoLex elements in the conversion of lexicographic data to RDF. Cases of type (4) were addressed in Chapter 5 by creating a custom vocabulary

for KD to account for specific DTD tags and values, which now is reduced to those specific cases in which nor OntoLex and its modules (including *lexicog*) nor LexInfo provide elements to capture the semantics of KD's DTD.

In this section we present the application of *lexicog* for the LD-based representation of the Global series of KD. Our main contribution for this section is two-fold:

1. This experience serves to validate the recently created model with an actual use case as well as to introduce some recommendations for future applications.
2. By focusing on KD's data, we examine how the limitations of the *lemon* model already reported in the previous chapters (in the literature, in Bosque-Gil et al. (2016b); Klimek and Brümmer (2015)) are successfully addressed by the *lexicog* module.

The *lexicog* modules draws a distinction between the lexical layer, captured mainly by OntoLex, and the structural elements that describe the lexicon and can be arranged as desired in a particular lexicographic work. We will adhere to this distinction also in this section and will first present those problematic cases of KD of type (1), concerning the distinction between a dictionary entry and an `ontolex:LexicalEntry`, and the grouping of dictionary entries, and will follow with the representation of examples and their translation as LLD.

6.3.2 `lexicog:Entry` and `ontolex:LexicalEntry`

One of the shortcomings of OntoLex-*lemon* concerned the lack of a way to capture what was originally a dictionary entry in the resource and differentiate it from `ontolex:LexicalEntry` instances created on the fly during the conversion process, which may or may not have their corresponding dictionary entry in the resource (or in a work of the same series, i.e the Global series from KD). In addition, a `lime:Lexicon` gathers a collection of `ontolex:LexicalEntry` instances, which, in turn, can share the language and come from different lexicographic resources from the same series. The `lime:Lexicon` class is thus not intended to represent uniquely the lexicographic resource as it was conceived originally, but a collection of lexical entries belonging to the lexicon of a language.

In KD’s data, compounds, synonyms, antonyms and translations are defined or described inside a dictionary entry of another lemma (in the case of compounds, inside the dictionary entry of one of their components). In order to represent their definition, form, inflection or pronunciation according to the OntoLex core, they need to be treated as `ontolex:LexicalEntry` instances, which causes this distinction between original dictionary entries and embedded lexical entries to be lost.

Example 6.10 shows an extract of the dictionary entry *arte* ‘art’ in Spanish, with its translation into Dutch and the definition of the compound *artes plásticas* ‘visual, plastic arts’. This example, in addition to a description of the headword (shortened due to space reasons) provides a synonym of the headword in its first sense (*inspiración* ‘inspiration’). Below the section devoted to translations, the compound *artes plásticas* is defined.

By applying *lexicog* to Example 6.10, we instantiate different elements to represent lexical entries and dictionary entries respectively. Example 6.11 shows an extract of the RDF Turtle representation of Example 6.10. While the Spanish and Dutch lexica (instances of `lime:Lexicon`) gather any unit of the lexicon that is described in the original dictionary (as dictionary entry or as embedded entry), represented as `ontolex:LexicalEntry`, a `lexicog:LexicographicResource` is intended to group only dictionary entries, i.e. instantiations of `lexicog:Entry`. In this way, the RDF reflects that *artes plásticas* is a unit of the lexicon but it is not a headword in this dictionary.

```

<DictionaryEntry identifier="DE00005536" version="1">
  <HeadwordCtn>
    <Headword>arte</Headword> [...]
  </HeadwordCtn>
  <SenseBlock>
    <SenseGrp identifier="SE00007455" version="1">
      <Synonym>inspiración</Synonym> [...]
      <TranslationCluster identifier="TC00017354" text="manifestación humana
        con intención estética" type="def">
        <Locale lang="nl">
          <TranslationBlock>
            <TranslationCtn>
              <Translation>kunst</Translation> [...]
            </TranslationCtn>
          </TranslationBlock>
        </Locale> [...]
      </TranslationCluster>
      <CompositionalPhraseCtn version="1"> [...]
      <CompositionalPhrase>artes plásticas</CompositionalPhrase>
      <Definition>artes que utilizan el dibujo o el volumen: la pintura,
        la escultura y la arquitectura</Definition> [...]
      </CompositionalPhraseCtn> [...]
    </SenseGrp> [...]
  </SenseBlock>
</DictionaryEntry>

```

Example 6.10: An extract of the dictionary entry *arte* ‘art’ in Spanish from KD’s Global series with its translation into Dutch and the definition of the compound *artes plásticas* ‘visual plastic arts’.

`lexicog:Entry` serves a structural function to only capture the structure of the resource as result of the lexicographical selection process, and it does not bear lexical information. To close this gap, the property `lexicog:describes` links dictionary entries (as structures) to the lexical units in the lexicon. If the RDF representation were also to reflect that *artes plásticas* or *inspiración* are lexical entries *defined* inside the dictionary entry of *arte*, the *lexicog* module provides elements to establish this structural connection. In this case, however, reflecting the whole microstructure of the entry was not a requisite for the expected output: we limit ourselves to capture the semantic relations between these different lexical entries (translation, synonymy) through the

elements of the OntoLex-*lemon* model, following the approach proposed in Chapter 5, based on the *vartrans* module.

```
@prefix base: <http://lexicala.com/id/global/> .
@prefix lime: <http://www.w3.org/ns/lemon/lime#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix lexicog: <http://www.w3.org/ns/lemon/lexicog#> .
@prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#> .

:mlds-ES3 a lexicog:LexicographicResource;
  dc:language "es" ;
  lexicog:entry :ES_DE00005536 .
:ES_DE00005536 a lexicog:Entry ;
  lexicog:describes :lexiconES/arte-n .

:lexiconES/arte-n a ontolex:LexicalEntry .
:lexiconES/artes-plásticas-n a ontolex:LexicalEntry .
:lexiconES/inspiración-n a ontolex:LexicalEntry .
:lexiconNL/kunst-n a ontolex:LexicalEntry .

:lexiconES a lime:Lexicon;
  lime:language "es" ;
  lime:entry :lexiconES/arte-n, :lexiconES/artes-plásticas-n, :lexiconES/
    inspiración-n .

:lexiconNL a lime:Lexicon;
  lime:language "nl";
  lime:entry :lexiconNL/kunst-n .
```

Example 6.11: RDF Turtle representation of Example 6.10 (definitions not included)

6.3.3 Nested Entries

There are other types of information in KD's Global series that do require the RDF version of the dictionary to reflect structural aspects. In KD's DTD, the element **NestEntry** works as container grouping together several dictionary entries. Example 6.12 in XML shows the entry of the verb *besuchen* 'visit' in German, presented in Example 5.18 in Chapter 5. The element **NestEntry** groups together three different dictionary entries: *besuchen* (v. 'visit'), *Besuch* (n. 'visit') and *Besucher* (n. 'guest, visitor') that are related, although the nature of relation is not *explicitly* stated. These containers

group together derivations or, in some cases, verbs that share lemma but not subcategorization value and are not homonyms (and the editors decided not to treat this case as a polysemous entry with various senses).

```
<Entry HomNum="" hw="besuchen" identifier="EN00002666" pos="verb">
  <NestEntry>
    <DictionaryEntry identifier="DE00003297" version="1">
      <HeadwordCtn>
        <Headword>besuchen</Headword> [...]
      </HeadwordCtn> [...]
    </DictionaryEntry>
    <DictionaryEntry identifier="DE00003298" version="1">
      <HeadwordCtn>
        <Headword>Besuch</Headword> [...]
      </HeadwordCtn>
      [...]
    </DictionaryEntry>
    <DictionaryEntry identifier="DE00003299" version="1">
      <HeadwordBlock>
        <HeadwordCtn>
          <Headword>Besucher</Headword> [...]
        </HeadwordCtn>
        <HeadwordCtn>
          <Headword>Besucherin</Headword> [...]
        </HeadwordCtn> [...]
      </HeadwordBlock> [...]
    </DictionaryEntry>
  </NestEntry>
</Entry>
```

Example 6.12: An extract of the German entry *besuchen* ‘visit’ with a *NestEntry* container that groups the dictionary entries *besuchen* *Besuch* ‘n. visit’ and *Besucher* ‘guest visitor’.

Example 6.13 shows the RDF rendering of example 6.12 in Turtle serialisation. In *lexicog*, grouping is reflected by creating a `lexicog:LexicographicComponent` and indicating that other components, namely, the three dictionary entries *besuchen*, *Besuch* and *Besucher* (instances of `lexicog:Entry`) are contained in that component. This is captured by the property `rdfs:member`.

```

@prefix base: <http://lexicala.com/id/global/> .
@prefix lime: <http://www.w3.org/ns/lemon/lime#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix lexicog: <http://www.w3.org/ns/lemon/lexicog#> .
@prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#> .

:lexiconDE/besuchen-v a ontolex:LexicalEntry .
:lexiconDE/Besuch-n a ontolex:LexicalEntry .
:lexiconDE/Besucher-n a ontolex:LexicalEntry .

:lexiconDE a lime:Lexicon;
  lime:entry :lexiconDE/besuchen-v, :lexiconDE/Besuch-n, :lexiconDE/Besucher-n
  .

:mlds-ES3 lexicog:entry :DE_DE00003297, :DE_DE00003298, :DE_DE00003299 .

:DE_DE00003297 a lexicog:Entry;
  lexicog:describes :lexiconDE/besuchen-v .

:DE_DE00003298 a lexicog:Entry ;
  lexicog:describes :lexiconDE/Besuch-n.

:DE_DE00003299 a lexicog:Entry ;
  lexicog:describes :lexiconDE/Besucher-n .

:DE_EN00002666 a lexicog:LexicographicComponent ;
  rdfs:member :DE_DE00003297, :DE_DE00003298, :DE_DE00003299 .

```

Example 6.13: RDF rendering of the `NestEntry` structure presented in Example 6.12 in Turtle serialisation

6.3.4 Usage Examples

The data from KD's Global series provides, for each sense of a headword, a usage example in the source language and the translations of the headword in the target language. Each example, in turn, is also translated into the target language and it often serve as examples of usage for the translations. Example 6.14 shows another excerpt of

the entry *arte* in Spanish. Inside the **SenseGrp** encapsulating the information of the first sense, there is an element **TranslationCluster** with **Locale** groups that include the headword translations for *arte* in its first sense: *kunst* (Dutch) and *kunst* (Norwegian). Below the translations follows a **ExampleCtn** with the example of usage of *arte* in that sense, «La música, la danza y la pintura son formas de arte» ‘Music, dance and painting are art forms’. This example is in turn translated to Dutch and Norwegian.

```
<SenseGrp identifier="SE00007455" version="1">
  [...]
  <TranslationCluster identifier="TC00017354" text="manifestación humana con
    intención estética" type="def">
    <Locale lang="nl">
      <TranslationBlock>
        <TranslationCtn>
          <Translation>kunst</Translation> [...]
        </TranslationCtn>
      </TranslationBlock>
    </Locale>
    <Locale lang="no">
      <TranslationBlock>
        <TranslationCtn>
          <Translation>kunst</Translation> [...]
        </TranslationCtn>
      </TranslationBlock>
    </Locale> [...]
  </TranslationCluster>
  <ExampleCtn type="sid" version="1">
    <Example>La música, la danza y la pintura son formas de arte.</Example>
    <TranslationCluster identifier="TC00017355" text="La música, la danza y la
      pintura son formas de arte." type="exmp">
      <Locale lang="nl">
        <TranslationBlock>
          <TranslationCtn>
            <Translation>Muziek, dans en schilderen zijn vormen
              vankunst.</Translation>
          </TranslationCtn>
        </TranslationBlock>
      </Locale>
      <Locale lang="no">
        <TranslationBlock>
          <TranslationCtn>
```

```

        <Translation>Musikk, dans og maling erkunsttyster.</Translation>
      </TranslationCtn>
    </TranslationBlock>
  </Locale> [...]
</TranslationCluster>
</ExampleCtn>
</SenseGrp>

```

Example 6.14: An extract of the Spanish entry *arte* ‘art’, with translations into Dutch and Brazilian Portuguese, examples, and translations of the examples

While the *lemon* model provided a class `lemon:UsageExample` and a property `lemon:example`, used previously in the literature to capture this information (Klimek and Brümmer, 2015), these are no longer included in the *OntoLex-lemon* model. The approach followed in Chapter 5 proposed a custom class `kd:UsageExample` in order not to instantiate both *lemon* and *OntoLex-lemon* in the same resource. If an example is to be linked to a sense, the property `skos:example` would suffice to include the example as a string at the sense level. For cases in which the example has additional information or there are other elements linkable to it, the *lexicog* module offers the class `lexicog:UsageExaple` to link an `ontolex:LexicalSense` to an element representing the example. A `lexicog:UsageExaple` can be further linked to other elements and described with data-type properties.

Example 6.15 shows the RDF Turtle representation of Example 6.14. As showed in Example 6.11, the headword and the translations belong to different lexica, one per language. The example is recorded through an instance of `lexicog:UsageExample` linked to the senses via `lexicog:usageExample`. Note that this instance has different values, each for the realisation of that example in a different language.

```

@prefix base: <http://lexicala.com/id/global/> .
@prefix lime: <http://www.w3.org/ns/lemon/lime#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix lexicog: <http://www.w3.org/ns/lemon/lexicog#> .
@prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#> .

(Continuation)

:lexiconES/arte-n a ontolex:LexicalEntry ;

```



```

    ontollex:sense :lexiconES/arte-n-SE00007455-sense .
:lexiconNL/kunst-n a ontollex:LexicalEntry;
    ontollex:sense :lexiconNL/kunst-n-arte-n-SE00007455-sense .
:lexiconNO/kunst-n a ontollex:LexicalSense;
    ontollex:sense :lexiconNO/kunst-n-arte-n-SE00007455-sense .

:lexiconNO a lime:Lexicon;
    lime:language "no";
    lime:entry :lexiconNO/kunst-n .

:lexiconES/arte-n-SE00007455-sense a ontollex:LexicalSense ;
    lexicog:usageExample :lexiconES/arte-n-SE00007455-sense-TC00017355-ex .
:lexiconNL/kunst-n-arte-n-SE00007455-sense a ontollex:LexicalSense ;
    lexicog:usageExample :lexiconES/arte-n-SE00007455-sense-TC00017355-ex .
:lexiconNO/kunst-n-arte-n-SE00007455-sense a ontollex:LexicalSense ;
    lexicog:usageExample :lexiconES/arte-n-SE00007455-sense-TC00017355-ex .

:tranSetES-NL/arte-n-SE00007455-sense-kunst-n-arte-n-SE00007455-sense-
    TC00017354-trans a vartrans:Translation ;
    vartrans:source :lexiconES/arte-n-SE00007455-sense;
    vartrans:target :lexiconNL/kunst-n-arte-n-SE00007455-sense;
    dc:source :mlds-ES3 .
:tranSetES-NO/arte-n-SE00007455-sense-kunst-n-arte-n-SE00007455-sense-
    TC00017354-trans a vartrans:Translation ;
    vartrans:source :lexiconES/arte-n-SE00007455-sense;
    vartrans:target :lexiconNO/kunst-n-arte-n-SE00007455-sense
    dc:source :mlds-ES3 .

:lexiconES/arte-n-SE00007455-sense-TC00017355-ex a lexicog:UsageExample ;
    rdf:value "La música, la danza y la pintura son formas de arte."@es ;
    rdf:value "Muziek, dans en schilderen zijn vormen van kunst."@nl ;
    rdf:value "Musikk, dans og maling er kunsttyper."@no .

```

Example 6.15: RDF Turtle representation of Example 6.14 showing usage examples and translations of the usage examples into the target language

6.4 Summary

In this chapter, we have drawn attention to eight of the issues we encountered in the analysis and conversion of six lexicographic resources (see Chapters 4 and 5) and in the

further analysis of dictionary entries in English, German and Spanish. We argue that the OntoLex-*lemon* model should enable the preservation of the content and the structure of the original resource, even if the LLD expert opts for a different representation more suited to the exploitation of the data by external applications or more in line with their view on the lexicon-ontology interface.

The new *lexicog* module of OntoLex-*lemon* emerges in response of the need to overcome the limitations of the *de facto* standard to model lexicographic information as LD in a way that is agnostic to the underlying lexicographic view and aims at information preservation. It complements the *lemon* core module, OntoLex, by providing an abstract model of language and linguistic objects in lexicography that will allow for both the conversion of electronic and digitalised print dictionaries to RDF and the compilation of LD-lexicographic resources from scratch (see Chapter 8).

In the last section of this chapter we report on the application of the module to KD’s multilingual data as a real use case scenario. We have shown that *lexicog* addresses the gaps identified in previous chapters and reported in the literature (Bosque-Gil et al., 2017, 2016b; Klimek and Brümmer, 2015) as regards the loss of structural and implicit lexical information in the original resource, and provides elements to capture data frequently found in lexicographic records, such as examples of use, their translations, or the morphosyntactic features of headwords and the relation of these features with senses.

The module is compatible with mechanisms suggested in the state-of-the art on dictionaries in LLD (See Chapter 2), as of the moment of writing, and also with other potential modules for the representation of specific lexical aspects which are under development at the time of writing (*morph* module for morphology and the *frac* module for corpus and attestations¹⁸⁰). *lexicog* is intended to be dictionary-agnostic in the sense that it should be applicable (and combined with other modules if necessary) to different kinds of dictionaries (e.g., general, collocations, learner’s, etymological, historical, etc.). This would bring LD closer to lexicography not only with the aim of leveraging already available dictionaries in LD for NLP tasks, but also of introducing LD in the work implemented in that discipline.

¹⁸⁰See <https://www.w3.org/community/ontolex/wiki/Morphology> and <https://acoli-repo.github.io/ontolex-frac/>.

Chapter 7

Evaluation

In this chapter we elaborate on the methodology we have followed to evaluate the contributions of this thesis and the results of such an evaluation. In the process, we will focus on how these results verify the hypotheses defined in Chapter 3, and how each contribution serves to attain the goals for the thesis established in that same chapter.

7.1 Evaluation methodology

The evaluation is performed in two different and complementary ways: as a query-based evaluation, and as theoretical analysis-based one. The whole evaluation consists of two ordered steps, namely:

1. Evaluation of the **conformance** of the RDF resources to the proposed application profiles. This constitutes the first step so as to ensure that the RDF generated from the work presented in Chapter 5 follows the suggested models.
2. Evaluation towards the **testing of hypotheses** with the results of the contributions, once the results of step 1 are satisfactory.

Step 1 is a query-based evaluation conducted in order to assure that the RDF resulting from the conversion does in fact comply to the proposed application profiles of the selected vocabularies (see Section 7.2). This step is performed iteratively until the results are satisfactory, and the data owners (if applicable) agree with the final representation. It is framed on Step 6 of the general methodology followed in this thesis (Section 3.6). This kind of evaluation involves the definition of two types of queries:

1. General queries (**Type A**): queries *common* to all resources that share the same application profile of *OntoLex-lemon* or *lemon*. E.g. queries to check that an `ontolex:LexicalSense` is not linked to *two* `skos:Concepts`, or queries to check that a `skos:definition` takes an `ontolex:LexicalConcept` or a `skos:Concept` as subject, etc.
2. Resource-specific queries (**Type B**): queries that consider the particularities of the application profile (e.g. whether the resource includes or not register information, terminological norms, definitions, terminological variants, etc.).

Note that queries of Type A are simple routine queries that can be reused or adapted to validate compliance to the *OntoLex-lemon* model and the best practices discussed in the framework of the Lexicography module with any other resource. Queries of Type B are resource-specific, but they can serve as guidance to monitor the instantiation of *OntoLex-lemon* or *LexInfo* in other resources that share features with those discussed in this thesis.

Step 2 is focused on assessing whether the contributions verify or not the hypotheses presented in Chapter 3, listed here below for reference. Depending on the hypothesis, this evaluation is query-based or guided by a theoretical analysis.

- H1. Current *de facto* standards for the representation of linguistic information as LLD and the available best practices for the generation of multilingual LLD provide mechanisms to capture lexical data contained in a range of NLP and human-oriented multilingual terminologies and NLP-oriented lexica, but fall short of preserving lexical data from lexicographic records (explicit, in the form of annotations or implicitly encoded through microstructural aspects) in mono- and multilingual general human-oriented trade dictionaries.
- H2. New modelling solutions and application profiles of the *de facto* standard and extensions to it are able to represent both the microstructure of the lexicographic record as well as annotations found in mono- and multilingual general human-oriented trade dictionaries. Such modelling solutions and applications are able to preserve explicit lexical data in lexicographic entries and implicit lexical information reflected on structural aspects during their conversion into RDF.
- H3. The LD paradigm brings benefits to the lexicographic work:

- H3.1. By virtue of being based on graphs and URIs, the LD paradigm resolves the limitations of tree-based formats for the linking of lexicographic data in a resource by enforcing the reuse of its elements and the instantiation of typed relations.
- H3.2. By virtue of being based on graphs and URIs, the LD paradigm allows for onomasiological and semasiological access on the same lexicographic resource.
- H3.3. It is possible to reuse and integrate the content of lexicographic resources, if represented as RDF, in NLP systems for further lexicographic enrichment.
- H4. Lexicographic resources represented as LLD allow linguists and lexicographers to formulate single queries that combine and filter lexical data from different sources to obtain data for their research, which would not be possible if the resources were not linked and did not share a common representation format.

7.1.1 Query-based evaluation

The query-based evaluation is conducted for Hypotheses H2, H3.2 and H4 over the RDF generated as result of Step 4 of our methodology (Section 3.6).¹⁸¹ This leads to the definition of three types of queries, one per hypothesis:

1. Queries to compare the number of entities of different types (e.g. of type <DictionaryEntry>) in KD's original resource format to their counterpart in the RDF, focusing on the disparities, their causes and implications. This evaluates the **information preservation** requirement defined for this conversion in particular, and aimed to with the definition of the model application profile and the *lexicog* extension for future similar cases. These queries serve thus for the evaluation of Hypothesis H2.
2. Sample queries to showcase the **various lexicographic access forms**, onomasiological and semasiological, on the same graph. These queries are defined in order to validate Hypothesis H3.2 and prove a benefit from the application of LLD to a single resource due to the **features of an RDF graph** (independently of its linking to external lexicographic resources).

¹⁸¹For the sake of the evaluation of this thesis, the RDF data is temporarily accessible at the time of writing at <http://jbosque.oeg-upm.net/sparql> with a password provided to panel members due to some datasets being private.

3. Sample queries across different lexicographic resources loaded to the same graph. These queries **combine complementary information on previously isolated resources**, and aim at providing prototypes of the kind of queries that would allow lexicographers, linguists and any other user that consults this kind of resources to obtain the complementary data they need. These queries are defined and performed so as to test Hypothesis H4.

7.1.2 Theoretical analysis

The theoretical analysis-based evaluation is applied for Hypotheses H1, H2, and H3.1. Two analyses are performed:

1. An analysis of the current *de facto* models and the proposed application profiles and extensions in terms of coverage of the representation needs of lexicographic resources detected throughout the thesis. These needs are inspired by the selected cases studied in Chapter 4 and the extension phase of the methodology presented in Chapter 6 as motivation for the *lexiog* module. This analysis focuses, as queries of type 1 introduced in the previous section, on **information preservation** from the original resource to its LLD counterpart. This analysis serves to test Hypotheses H1 and H2.
2. An analysis of the implications of moving from a hierarchical structure to a graph-based representation for lexicographic objects. This analysis complements queries of type 2 related with the different lexicographic access options in order to test Hypothesis H3.2, and also serves to validate Hypothesis H3.1, concerning the benefits of LLD for the representation of the macro and micro-structure of lexicographic resources.

Lastly, an analysis on the projects and uses of the generated resources by third parties for specific tasks in order to validate Hypothesis H3.3 is also provided.

7.2 Conformance of the RDF resources to the proposed application profile

Three examples of *common* queries (Type A) are presented here for illustration purposes: Examples 7.1, 7.2 and 7.3. Example 7.1 checks whether there are on-

`tolex:LexicalEntry` elements that have any `ontolex:Form` with a lemma tagged with a language that does not match the language of the lexicon. This query rests on the fact that a `lime:Lexicon` is described with a `lime:language` and should gather lexical entries in that language. If there were entries which deviate from this (not necessarily errors if borrowings are tagged with the source language tag), these query would help in detecting them. Example 7.2 identifies whether the *semantics by reference* approach has been implemented in the resource by checking whether entries are linked to an `ontolex:LexicalSense` with a reference or whether the sense is linked to an `ontolex:LexicalConcept`, following the core (OntoLex). Lastly, Example 7.3 yields results that show how gender has been modelled in the resource, provided that it should be encoded either at the lexical entry or the form level (e.g. some adjective in Romance languages). This query helps the LD expert in monitoring whether these distinction (e.g. between adjectives and nouns in Spanish and the gender encoding) has been respected in the modelling.

Examples 7.4 and 7.5 provide queries of type B, i.e. *resource-specific*. The first query extracts Spanish entries from *TermineSp* and their Latin denomination, provided that international scientific terms have been correctly modelled according to the proposed application profile of *OntoLex-lemon*, namely, as entries whose sense is related to the Spanish one through a `vartrans:TerminologicalRelation`. Example 7.5 is a simple query to inspect the lexical entries from Rayfield’s Georgian-English dictionary so as to assess whether lemmas that included commas or optional characters in the source data have been correctly mapped to different `ontolex:writtenRepresentation` elements according to the suggested application profile.

```

prefix lime: <http://www.w3.org/ns/lemon/lime#>
prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#>
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT DISTINCT ?entry ?writtenRep
{
  ?entry ontolex:lexicalForm ?lexForm;
    a ontolex:LexicalEntry.
  ?lexForm a ontolex:Form;
    ontolex:writtenRep ?writtenRep .
  BIND (datatype(?writtenRep) as ?word) .
  BIND (lang(?word) as ?wordLang ) .
  ?dict a lime:Lexicon;
    lime:entry ?entry;
    lime:language ?dictLang .
  FILTER (?wordLang != ?dictLang)
} LIMIT 200

```

Example 7.1: (Type A) SPARQL query to detect whether there are lexical entries with a written representation in a language that does not match the language of the lexicon that gathers them

```

prefix lime: <http://www.w3.org/ns/lemon/lime#>
prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#>
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT DISTINCT ?entry ?concept
{
  {?entry a ontolex:LexicalEntry ;
    ontolex:sense ?sense .
    ?sense ontolex:reference ?concept
  }UNION{?sense ontolex:isLexicalizedSenseOf ?concept .}
} LIMIT 200

```

Example 7.2: (Type A) SPARQL query to check that entries do have a sense with a reference or a lexical concept to which the sense is linked


```

prefix lime: <http://www.w3.org/ns/lemon/lime#>
prefix ontollex: <http://www.w3.org/ns/lemon/ontollex#>
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix lexinfo: <http://www.lexinfo.net/ontology/2.0/lexinfo#>

SELECT DISTINCT ?entry ?word ?gender ?pos
{
  {?lexform ontollex:writtenRep ?word ;
   lexinfo:gender ?gender .
  ?entry ontollex:lexicalForm ?lexform ;
   lexinfo:partOfSpeech ?pos . }
  UNION{
    ?entry ontollex:lexicalForm ?lexform ;
     lexinfo:partOfSpeech ?pos ;
     lexinfo:gender ?gender . }
  }
LIMIT 200

```

Example 7.3: (Type A) SPARQL query to check that gender is encoded at the form level (e.g. some adjective in Romance languages) or at the entry level

```

prefix lime: <http://www.w3.org/ns/lemon/lime#>
prefix ontalex: <http://www.w3.org/ns/lemon/ontalex#>
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix lexinfo: <http://www.lexinfo.net/ontology/2.0/lexinfo#>
prefix vartrans: <http://www.w3.org/ns/lemon/vartrans#>

SELECT DISTINCT ?esWord ?laWord
{
  ?dict lime:language "la";
    lime:entry ?scTerm .
  ?scTerm ontalex:sense ?laSense ;
    ontalex:lexicalForm [ontalex:writtenRep ?laWord] .
  ?trans a vartrans:TerminologicalRelation;
    vartrans:source ?esSense ;
    vartrans:target ?laSense ;
    vartrans:category lexinfo:internationalScientificTerm .
  ?entry ontalex:sense ?esSense ;
    ontalex:lexicalForm [ontalex:writtenRep ?esWord]
}

```

Example 7.4: (Type B) SPARQL query to check that scientific denominations in Termesp RDF are encoded according to the proposed application profile

```

prefix lime: <http://www.w3.org/ns/lemon/lime#>
prefix ontollex: <http://www.w3.org/ns/lemon/ontollex#>
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix lexinfo: <http://www.lexinfo.net/ontology/2.0/lexinfo#>
prefix vartrans: <http://www.w3.org/ns/lemon/vartrans#>

SELECT DISTINCT ?lemma ?wrRep
{
  ?entry rdfs:label ?lemma;
    ontollex:lexicalForm ?form .
  ?form ontollex:writtenRep ?wrRep .
  FILTER regex (?lemma, "\\,",")
  FILTER regex (?lemma, "\\(")
}

```

Example 7.5: SPARQL query to monitor that lemmas in Rayfield’s dictionary with a comma and optional characters are represented as lexical entries with different forms with correct written representations

In addition to queries Type A and B queries, in the case of the conversion of the KD’s Multilingual global series in 2019, a JSON-LD schema was developed by KD as a further verification mechanism to see to RDF conformance. Details on this schema are provided on Bosque-Gil et al. (2019).

7.3 Information preservation

7.3.1 Coverage

Hypotheses H1 and H2 concern the evaluation of the *de facto* standards to represent LLD models in terms of their **coverage** of the information encoded in source data models. Coverage is said to describe the set of concepts and relations in an ontology (Ouyang et al., 2011). We perform a manual analysis that hinges on a set of representation needs that arise in the conversion of lexicographic resources. In our evaluation, we state that a specific representation need from a lexicographic resource is met by a model is such information can be encoded by using such a model. Thus, if a tag in a dictionary does not have 1:1 mapping to an ontology element in *OntoLex-lemon*, but the model does provide other mechanisms to encode such information, we deem that it meets the representation need.

Table 7.1 lists these representation needs (first column), the coverage of OntoLex-*lemon* (second column), LexInfo (third column), and the coverage provided by the contributions of these thesis (fourth column). In this last column, the following abbreviation apply, and aim to specify in which way our contributions are meeting the given representation needs:

AP. An application profile from the modelling suggestions provided in Chapter 5 for those features for which one of the two conditions hold:

- (a) they are not already covered by the models presented in columns 1 and 2 and documented in their ontology definition, specification or in the BPMLOD outcomes.
- (b) Their coverage is not addressed specifically with respect to lexicographic resources.

For example, if one of our application profiles of OntoLex for KD covers gender, this will not be marked in the fourth column as *AP*: gender is addressed in the OntoLex-*lemon* Final Model Specification and clear examples are given so as how to model it for cases such as Spanish. Thus, as (a) is not applicable, this information does not make our suggested application profile distinctive. Latin denominations would be a different case: since encoding Latin denominations of terms is *not* addressed in OntoLex Final Model Specification or any of the above mentioned documentation sources in the context of lexicographic resources (conditions (a) and (b)), but it is addressed in our suggested application profile of OntoLex-*lemon* for Terminesp RDF, the feature *Latin denominations* in the first column will receive the value *AP* for the fourth column, highlighting the contributions of this thesis.

LX. The *lexicog* module.

In this way, the features that are specifically addressed by our contributions (and which makes them distinctive) are clearly highlighted either by the abbreviations *AP* or by *LX*. In addition, if a model in column 1-2 covers a feature, the fifth column (source) specifies where that coverage has been reported. For this column, the following abbreviations apply:

- S. The model final **specification**, the ontology itself, or the outcomes of the BPM-LOD group provide comments, definitions, examples, or modelling suggestions so as to how to cover the information given in column 1 with the models given in columns 1 and/or 2.
- L. There are efforts in the **literature** that reuse the models in columns 1 and 2 to cover the feature provided in column 1 (*custom extensions not taken into consideration*). However, these application profiles of the models are not explicitly described in their specification or suggested as part of the BPMLOD community group outcomes. For example, for the feature *Multiword expressions involving the lemma of an entry or one of its senses* there are no explicit guidelines on how to model the relation between a MWE and the lemma in a dictionary converted to OntoLex-*lemon*: the *decomp* module allows to specify which elements compose a given entry, but it does not focus on how to proceed in the case of embedded MWEs in a given dictionary entry that may not fit the definition of `ontolex:LexicalEntry` or in cases in which we want to link the `ontolex:LexicalSense` of the lemma to that of the MWE. In this case, there are AP profiles in the literature that address this issue within the model, and hence column 5 receives a value of L.

Table 7.1 shows that, despite that there are still representation needs which are not widely addressed in the community (e.g. dialectal information beyond the language tag level - if there is no available language tag for the dialect under study; or geographical usage encoded at the sense level) the application profiles of OntoLex-*lemon* and LexInfo suggested in this dissertation, along with the newly developed *lexicog* module, allow for a wider range of representation needs to be met, documented and covered in a standardised manner.

7.3.2 Information preservation in K Dictionaries

Evaluating information preservation not only concerns ensuring that there are mechanisms to encode the data provided in the original resource (coverage), but also analysing whether the conversion allows to retrieve the same entities present in the original resource, in the same number. This concerns the performance of queries of type 1 from the list of queries presented above.

#	Feature	OntoLex-lemon	LexInfo	This thesis	Source of the reported coverage
1	Lemma	✓	×	×	S
2	Pronunciation	✓	✓	×	S
3	Varios lemmas and alternative forms	✓	×	AP	S
4	Inflection	✓	✓	×	S
5	Transliteration	✓	✓	AP	S
6	Gender	✓	✓	×	S
7	Definitions	✓	✓	AP, LX	For synsets (S) For dictionary senses (L)
8	Notes to definitions	×	✓	AP	S for general notes not specific for definitions
9	Translations across senses	✓	✓	×	S
10	Initialisms and abbreviations	✓	✓	AP	S
11	Synonyms and antonyms (as sense relations)	✓	✓	AP	S (sense relations in general)
12	Register	×	✓	AP	S (general context condition)
13	Temporal qualifiers	×	✓	AP	S (only for chronological variants)
14	Domain	×	×	AP	L (using external vocabularies, e.g. DC)

Table 7.1: Coverage of various information types in *lemon*, *OntoLex-lemon* and *LexInfo*.

#	Feature	OntoLex-lemon	LexInfo	This thesis	Source of the reported coverage
15	Dialect	✓	×	×	S (at the language tag)
16	Geographical usage	×	×	×	
17	Connotations	✓	×	AP	S (only usage in general)
18	MWE expressions involving the entry lemma (A) or one of its senses (B)	✓	×	AP (for case B)	L (for case A)
19	Latin denominations	×	✓	AP	S
20	Selectional restrictions	✓	×	AP	S (for ontology lexicalization)
21	Sense examples	×	×	LX	
22	Translations of examples	×	×	LX	
23	Entries with several POS tags	×	×	LX	
24	Form restrictions on senses	×	×	LX	
25	Sense hierarchy	×	×	LX	
26	Inflection groups	×	×	LX	
27	Nested entries	×	×	LX	
28	Sense order	×	×	LX	

(continuation) Coverage of various information types in *lemon*, *OntoLex-lemon* and *Lex-Info*.

Element/Format	KD source	RDF	Element in RDF
DictionaryEntry	1500	1500	lexicog:Entry
DictionaryEntry	1500	1903	ontolex:LexicalEntry
CompositionlPhrase	66	66	ontolex:LexicalEntry
SenseGrp	1891	2373	ontolex:LexicalSense

Table 7.2: Comparison of the # of main components of the microstructure for the first 1500 entries of KD Spanish dictionary

For a sample dataset of 1500 dictionary entries (the first 1500 of the Spanish dictionary), K Dictionaries provided us with a list of numbers for a small set of elements.

Given that the application profile for KD represents headwords as `ontolex:LexicalEntry` elements, as well as synonyms, antonyms, and compositional phrases, we expected the number of lexical entries to be equal or higher than the number of dictionary entries (depending on the presence of embedded lexical entries in the analysed excerpt of data). Table 7.2 verifies that, while there is a one-to-one mapping between KD’s DTD tag `<DictionaryEntry>` and `lexicog:Entry`, there is a disparity in the number of lexical entries with respect to the number of dictionary entries, with a difference of 403 lexical entries in the Spanish lexicon that do not occur in this extract as headwords. This is due to the number of synonyms and antonyms, which are modelled as `ontolex:LexicalEntry` elements and described by `lexicog:Entry` ones only if they also belong to the headword list of the dictionary. Since the extract analysed covers only 1500 entries, some synonyms or antonyms might be headwords in the rest of the dictionary but do not occur as such in this extract. In addition, 66 of those 403 lexical entries are compositional phrases which **are never** headwords in KD, as they are embedded inside sense groups. This difference between dictionary entries and embedded lexical entries that receive ‘artificial senses’ (see Section 5.5) is also causing an increase of 482 `ontolex:LexicalSenses` with respect to the `<SenseGrp>` tag for headword senses. These numbers indicate that the main components of the microstructure of the entry are preserved, but also that extracting the embedded pieces allows for the `lime:Lexicon` to seamlessly grow and gather elements that were not linkable to other ones as ‘first class citizens’ independent of the main entry, which now is possible in the RDF version.

In this way, the analysis presented in Section 7.3.1 and the previous results based on KD's elements serve to validate Hypotheses H1, H2, focused on information preservation.

7.4 From trees to graphs: benefits of LLD for lexicography

Hypotheses H3.1 and H3.2 concern the benefits that LD brings to lexicography due to the representation of lexical data as a graph. Based on the experiences presented in Chapter 5, in this section we focus on the qualitative analysis of the advantages that LD offer in comparison to tree-based formats for the representation of lexical data, showcasing the two different kinds of access to lexical content, onomasiological and semasiological, that LLD lexical resources allow. This analysis is presented in Bosque-Gil et al. (2016a) and Gracia et al. (2017).

In modern electronic dictionaries, entries are typically represented as a tree (usually encoded in XML), following a hierarchical data structure where every element has at most one parent. As discussed by Měchura (2016), this choice of data structure makes some aspects of the lexicographer's work unnecessarily difficult, from deciding where to place multiword items to reversing an entire bilingual dictionary. This is a consequence of the fact that dictionary writing, although assisted by computing methods, still tends to replicate what lexicographers would be doing on paper or with a word processor. This raises a number of issues.

First, we can mention the problem of headword selection for multiword phrasemes (Měchura, 2016), e.g., under which entry to place *bow and scrape* (meaning *to be overly polite*), *bow* or *scrape*? Ideally, it should be placed under both entries. However, in a tree-like representation, this obliges the lexicographer to copy the same information in both places, which makes the data more difficult to be maintained or updated (changes in one place need to be propagated into other places). Of course, clever search mechanisms can be built to work around this problem, as modern digital dictionaries do, in which a lemma is provided just once and the system is able to search it wherever it appears. However, that does not solve the problem at source, and the search mechanism is not able to infer the particular sense or homograph of the parent entries that should be associated to the phraseme.

Another example of an issue caused by tree-based view of the dictionary information is that cross-references typically depend on the order of appearance of lexical entries or senses, being usually indicated by a superscript in numeric form in printed or electronic format, e.g., *bow*², meaning, for instance, the second homograph of the entry *bow*. The problem of this approach is that the introduction of new elements in the middle of the sequence obliges to review and redefine all the involved cross-references across the dictionary, making this modelling technique very sensitive to any change in the ordering criteria. Techniques such as the latter are prone to errors and might result in the collision of identifiers. Again, mechanisms have been implemented that reduce such a problem, although they do not solve it at source.

In the following we provide an analysis on the role that LD plays to address issues such as these. This analysis has been performed on two layers, the micro and the macro-structure of lexicographic resources.

7.4.1 Benefits of LD for the macro-structure of a lexicographic resource: internal reusability

The most evident advantage for the application of LLD technologies to lexicography is that LD enables the integration with other external resources thanks to the semantic and syntactic interoperability achieved by the use of RDF and linguistic ontologies. In addition and focusing on the lexicographic resource on its own, we have identified the following benefits in the course of our work towards the migration of language resources to LLD, some of them also highlighted in the literature.

Firstly, the entries of a dictionary become **internally reusable** (Klimek and Brümmer, 2015) thanks to their URIs. This does not seem novel given that entries might already have numeric identifiers to point to each other, but the choice of transparent URIs, i.e, human-readable, which reflect the semantic content, and a suitable URI naming strategy (see Chapter 5), the editor of a dictionary entry will be able to refer to another entry without the need to know its identifier in advance. Following this, the entry `:lexiconEN/risk-n` can be linked to `:lexiconEN/risky-adj` through a relation of morphological derivation without the need of an identifier. If, later on, the noun *risk* occurs as an entry in another dictionary of the same or a different family of dictionaries, the information can be integrated in a straightforward manner without relying on dictionary-dependant numeric IDs.

This in turn relates to a second advantage: we **no longer depend on the order of appearance of lexical entries or senses in cross-references**, which is usually indicated by a superscript in numeric form in printed or electronic format, e.g: *bow*², meaning, for instance, the second homograph of the word *bow*. There are ways of keeping track of the order and the lexical entry to which that position refers, but a change in the original order of entries or the integration with other dictionaries in which the order differs would then require the update of all cross-references to any of the ordered entries. Since entries and senses are now identifiable throughout the data and graphs are not actually ordered, cross-references can be direct pointers to the entry or sense to which they refer.

The third advantage is intrinsically related with the first one, too: **we can represent an “abstract” lexicon that gathers all the entries in a specific language**. In other words, have a “pool” of lexical entries extracted from different dictionaries of the same or different type, monolingual or multilingual, without losing provenance information about which data comes from which dictionary. Thanks to an appropriate URI naming strategy, this pool of entries will grow dynamically (Gracia et al., 2018) with each dictionary converted into RDF that has any information about an entry in that specific language.

If the approach mentioned above is applied in the conversion of multilingual dictionaries, for instance, Spanish-French and French-English, linking the French entries from the ES-FR dictionary with their corresponding entries in the FR-EN dictionary will bring us a fourth advantage: **translation relations** can be established through a language acting as a pivot (Villegas et al. 2016).

The fifth benefit concerns the **onomasiological (as well as the semasiological) view that LD naturally enables**, which brings us to queries of type 2, from the list of queries for the verification of hypotheses mentioned in the previous sections. The source dictionary has probably been compiled from a semasiological perspective, by putting the word as the centre of attention and listing its different senses. Given the semantics by reference in LLD mentioned above, the synonym of a word and the word itself will point to the same concept, which is modelled as a node in the graph and has therefore a URI. Accessing that node will allow us to see which words lexicalize it, i.e. putting the concept as our focus and traversing the graph from it to the lexical elements related

to it. This way of thinking is well illustrated in the case of multilingual dictionaries in LLD, where we can see how a concept is verbalized in different languages.

To illustrate this, the query in Example 7.6 takes the concept as entry point, e.g. the concept “*hacer que una persona sienta cobardía*” “to cause fear in someone” (concept `kd-base:SE00001149-concept`). By accessing through the concept, we easily get to the different entries in the lexicon that verbalise that concept, in this case, two in Spanish, one in Portuguese, one in Danish and two in English. Table 7.3 shows the results of Example 7.6.

URI of Lexical Entry	URI of Lexical Concept
http://kdictionaries.com/id/LexiconES/achicar-vb	http://kdictionaries.com/id/SE00001149-concept
http://kdictionaries.com/id/LexiconES/acobardar-vb	http://kdictionaries.com/id/SE00001149-concept
http://kdictionaries.com/id/LexiconBR/acovardar-vb	http://kdictionaries.com/id/SE00001149-concept
http://kdictionaries.com/id/LexiconDK/goe,re_bange-vb	http://kdictionaries.com/id/SE00001149-concept
http://kdictionaries.com/id/LexiconEN/to_daunt-vb	http://kdictionaries.com/id/SE00001149-concept
http://kdictionaries.com/id/LexiconEN/to_intimidate-vb	http://kdictionaries.com/id/SE00001149-concept

Table 7.3: Results of the query in Example 7.6

While this would be the onomasiological access, the same graph would allow to enter through the lexical entry and access all the information regarding its meaning and form (semasiological view), which is exemplified in query 7.7. One of the entries that denotes the concept *to cause fear in someone* is the Spanish verb *achicar*. Query 7.7 gives as result the all the different senses of this entry, and note that one of them is the one we have taken earlier as entry point, *hacer que una persona sienta cobardía*.

The potential is however no less interesting in monolingual dictionaries. In Chapters 5 and 6 we showed that definitions have been encoded at the level of the concept. Even though definitions can be fine-grained, SPARQL queries over them are feasible. For instance, we can search for entries that verbalize concepts in whose definition the word *fear* or *terror* occurs, which will yield the series of entries such as *scare* or *intimidate* that denote related concepts. The query in Example 7.8 gives the results illustrated in Table 7.5.

```

prefix dc: <http://purl.org/dc/elements/1.1/>
prefix kd: <http://www.kdictionaries/def/ontolexKD#>
prefix kd-base: <http://kdictionaries.com/id/>
prefix lemon: <http://www.lemon-model.net/lemon#>
prefix lexicog: <http://www.w3.org/ns/lemon/lexicog#>
prefix lexinfo: <http://www.lexinfo.net/ontology/2.0/lexinfo#>
prefix lime: <http://www.w3.org/ns/lemon/lime#>
prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#>
prefix owl: <http://www.w3.org/2002/07/owl#>
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix skos: <http://www.w3.org/2004/02/skos/core#>
prefix synsem: <http://www.w3.org/ns/ontolex-synsem#>
prefix vartrans: <http://www.w3.org/ns/lemon/vartrans#>
prefix void: <http://rdfs.org/ns/void#>
prefix xml: <http://www.w3.org/XML/1998/namespace>
prefix xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT DISTINCT ?entryA ?concept
{
  ?concept a ontolex:LexicalConcept;
    skos:definition "hacer que una persona sienta cobardía"@es.
  ?senseA ontolex:isLexicalizedSenseOf ?concept .
  ?senseB ontolex:isLexicalizedSenseOf ?concept .
  ?entryA ontolex:sense ?senseA .
  ?entryB ontolex:sense ?senseB .
  FILTER (?senseA!=?senseB)
} LIMIT 20

```

Example 7.6: Query to access those lexical entries that verbalize the concept of ‘to cause fear in someone’

URI of Lexical Entry	Definition of Lexical Concept
http://kdictionaries.com/id/LexiconES/achicar-vb	disminuir el tamaño de una cosa
http://kdictionaries.com/id/LexiconES/achicar-vb	sacar agua de un lugar usando cubos
http://kdictionaries.com/id/LexiconES/achicar-vb	hacer que una persona sienta cobardía

Table 7.4: Results of query in Example 7.7

```

SELECT DISTINCT ?entry ?def
{
  ?entry a ontolex:LexicalEntry;
    ontolex:lexicalForm [ontolex:writtenRep "achicar"@es];
    ontolex:sense ?sense .
  ?sense ontolex:isLexicalizedSenseOf ?concept .
  ?concept skos:definition ?def .
} LIMIT 20

```

Example 7.7: Query to access all definitions (concept definitions) of the entry *achicar* ‘shrink, scoop out (water), intimidate’

```

SELECT DISTINCT ?entryA
{
  ?concept a ontolex:LexicalConcept;
    skos:definition ?def .
  ?senseA ontolex:isLexicalizedSenseOf ?concept .
  ?senseB ontolex:isLexicalizedSenseOf ?concept .
  ?entryA ontolex:sense ?senseA .
  ?entryB ontolex:sense ?senseB.
  FILTER (?senseA!=?senseB)
  FILTER regex (?def, "miedo | terror | pavor | pánico ")
} LIMIT 50

```

Example 7.8: Query to access entries that verbalize concepts in whose definition the word *fear* or *terror* occurs

Thus, by taking the concept as entry point we can get a set of concepts that are related but are not necessarily equivalent.

The sixth advantage is related to **cross-references** in the sense of any reference to another entry that might occur inside the lexicographic article: orthographic variants, synonyms, antonyms, genus terms, semantic types, etc. Not only are the entries

URI of Lexical Entry
kd-base:LexiconBR/intimidar-se-vb
kd-base:LexiconES/achantarse-vb
kd-base:LexiconBR/acovardar-se-vb
kd-base:LexiconDK/at _ være/blive _ bange/rædselsslagen-vb
kd-base:LexiconEN/to _ get _ scared-vb
kd-base:LexiconES/acobardarse-vb
kd-base:LexiconBR/acovardar-vb
kd-base:LexiconDK/skræmme-vb
kd-base:LexiconEN/to _ frighten-vb
kd-base:LexiconES/acobardar-vb
kd-base:LexiconES/acogotar-vb
kd-base:LexiconES/domeñar-vb
kd-base:LexiconBR/amedrontar-vb
kd-base:LexiconBR/intimidar-vb
kd-base:LexiconDK/undertrykke-vb
kd-base:LexiconEN/to _ dominate-vb

Table 7.5: Results of the query shown in Example 7.8

reusable throughout the data (first advantage), but the **pointers to them are now typed** (Klimek and Brümmer 2015, McCracken 2015). This might not seem like an evident benefit to the user of online dictionaries, for whom the label antonym or a typographical mark may suffice, but typed properties allow users to perform queries not dependant on the (proprietary) format of the data and LD-aware systems to find any needed information. At the same time, by virtue of being defined in a public external vocabulary, e.g. LexInfo, the same properties can be reused in the conversion of other lexica of the same series into LLD, thus gaining interoperability. This responds to the need of standardization among the high number of heterogeneous annotation schemas, tagsets, and proprietary DTDs that are being used to create language resources. Furthermore, given that these vocabularies are extensible, new properties and individuals or classes can be added. If the hierarchy defined in a linguistic ontology is not compatible with the view other domain experts might have, new vocabularies can be created and aligned to the ones already available.

As the last paragraphs suggest, the concept of reusability lies at the heart of LD. If the enterprise of compiling a dictionary is seen through the looking glass of LD from the very beginning, it will affect the whole process. Chapter 8 will delve into the decisions such as, for example, keeping independent lexical entries for an entry and its homographs will have to be considered from the point of view of lexicography (two words that share form but are not related etymologically could thus be regarded as independent entries) and LD. It will not be a matter of converting lexical data to LD, but of creating them from scratch in a reusable, interoperable and linguistically accurate way.

7.4.2 Benefits of LD for the micro-structure of a lexicographic resource

The previous section elaborated on the benefits of representing a lexicon as LD but it did not deepen into the modelling of information present in a single lexicographic article, whose information (definitions, grammatical data, semantic relations, etc.) is itself represented as a graph, as we saw in Chapters 5 and 6. Ideally, everything in the lexicographic entry can be modelled as a node (McCracken 2015) but, in general, and on the basis of *OntoLex-lemon* the representation revolves around lexical entries (`ontolex:LexicalEntry`), concepts, the relation between entries and concepts

reified as lexical senses (`ontolex:LexicalSense`), word forms (`ontolex:Form`), definitions, phonetic representations, register, syntactic frames, etc. Relations between nodes have a well-defined domain and range, and, with actual data, every node will be an instance of a class defined in an ontology. Following the *OntoLex-lemon* model, the English entry *cloud*, with the sample URI `:lexiconEN/cloud-n` will have `rdf:type ontolex:LexicalEntry`, will denote as many `skos:Concepts` as senses or meanings it has, and the relation from the word to the concept will be encoded as `ontolex:LexicalSense`. Word forms (*cloud*, *clouds*) will be recorded at the `ontolex:Form` level, together with grammatical number information and phonetic transcription. Definitions, usage examples, etc. are likewise linked to the entry through edges and intermediate nodes.

On the one hand, one of the consequences of this configuration is that elements previously embedded in the lexicographic article become entry points in the graph and are no longer subsumed under any entry, since the **hierarchy is lost**. This implies that an idiom or collocation, for instance, will not be encapsulated under the container of the entry in which it was originally defined, but will be related to it with the suitable property. Since the idiom now becomes a node, we are able to link it to any other node from any other entry in the lexicon: *like a cat on a hot tin roof* could then be linked, for example, to the appropriate sense of *cat*, of *hot* and of *roof*, if desired, which will allow to access the idiom from any of those entries. Also, in the case of idioms and frequent collocations, we are creating new lexical entries that were not originally conceived as such in the lexicon. As lexical entries, they will be also linked to their corresponding `skos:Concepts`, which brings us back to the possibility of an onomasiological perspective on the data.

On the other hand, thinking in terms of LD forces us to **constantly question what is the nature of the relation between two pieces of information**. An LD-native dictionary will require a specification on the part of lexicographers of which kind of relations between which type of elements will be encountered when modelling lexicographic articles (details of this are provided in Chapter 8). This brings us to the difference between compiling dictionaries with only the human as target, and creating them for (both humans and) computers. The fact that an XML tag, for instance, can occur at different levels in the dictionary entry (e.g. a geographical usage indication attached to a pronunciation vs. a geographical usage indication attached to a sense,

as in the case of KD’s DTD) seems straightforward enough for a human, but an NLP application needs to be able to distinguish between a description of a string (e.g. [ka:] is the transcription of the British pronunciation of car) and the restriction on the usage of a sense (e.g. the *floor* with the meaning the *floor above the ground level floor* is only used in the UK). Modelling data as LLD thus entails a reflection of which information affects which elements, and which properties are the most suitable ones to be used in which case, taking all nuances and human implicit knowledge into account.

7.5 Reuse of the generated outcomes by external projects

Hypothesis H3.3 alludes to the reuse and integration of some of the generated RDF resources in external projects in the context of NLP and LLD. Those projects and services are currently on-going, but the outcomes of this thesis play a role in their development:

1. Lexicala API by K Dictionaries: <https://api.lexicala.com/intro>. The API of KD provides access to data from the whole Global series, which has been modelled to RDF following the application profiles of OntoLex-*lemon* suggested in this thesis. Currently, the data provided via the API adheres to the model presented in Chapter 5 and refined in Chapter 6 with the use of the *lexicog* module. The data offered in this API instantiates thus two of the contributions of the thesis: the application profile of OntoLex-*lemon* devised in 2016 and the *lexicog* extension update from 2019.
2. Legal Knowledge Graph for Multilingual Compliance Services (LYNX) (2017-2020)
 - (a) As part of its WP2-Data acquisition and management, the Lynx project is working on the integration of KD’s Global series in RDF into the Lynx platform to be consumed as domain independent vocabulary for the Legal Knowledge Graph. This in turn reuses the Lexical API, and involves a step of enrichment with additional sources in LD format (WordNet, DBpedia). The information from KD will be also reused, in necessary, to enable disambiguation mechanisms for semantic annotation in WP3 – Common services infrastructure.

- (b) Also as part of WP2 and, specifically, in the task focused on domain-dependent vocabulary management (T2.3), the Terminoteca RDF belongs to the list of data resources identified and specified to be relevant for the Lynx platform, so as to interlink and enrich the integrated resources in the Lynx platform through NLP systems developed as part of WP3.
3. The European Lexicographic Infrastructure (ELEXIS) (2018-2021). The ELEXIS project reuses OntoLex-*lemon* as one of its interoperability formats, taking into account the *lexicog* extension. Throughout 2019, its data model will be defined in a way such that it can be expressed by the *de facto* standard. In addition, KD’s data in JSON-LD format is provided as input for the ELEXIS Dictionary Matrix under development. Lastly, Chapter 8 proposes the notion of *LD-native dictionaries*, whose generation and publication will be addressed as part of ELEXIS (Declerck et al., 2019).

7.6 Queries over complementary data

Hypothesis H4 states that lexicographic resources represented as LLD allow linguists and lexicographers to formulate single queries that combine data extracted from different sources (to obtain data for their research), which would not be possible if the resources were not linked and did not share a common representation format. In this section we validate this scenario with a series of queries that combine complementary information from resources that were previously isolated.

One of the first aspects that comes to mind when dealing with multilingual resources represented as LLD is the possibility to obtain indirect translations. In this case, however, we do not have sense alignments between the different resources or a common index to make sure the translation relation holds, as it is not transitive. Still, and taking this risk into consideration, we can access a range of potential indirect translations of a given word by just relying on the translation in a given resource, and the translations present in another resource which involve that same lemma. Inferring translations across dictionaries is an active research line in itself (see for instance the Translation Inference Across Dictionaries (TIAD 2019) shared task).¹⁸²

¹⁸²<http://tiad2019.unizar.es/>

Example 7.9 shows a query that accesses both English translations extracted from Rayfield’s dictionary Georgian entries, as well as other potential translations of that Georgian entry gathered from Terminoteca RDF on the basis of the English lemma of the translation. These potential translations are retrieved by searching for English entries in Terminoteca RDF which share the same lemma as the English Rayfield translation. By pivoting on the English entries in Terminoteca, we access all other entries in different languages in Terminoteca RDF denoting the same concept as the English entries. Note that the results, shown in Figure 7.1, are likely to include errors due to the lack of sense alignments between the English lexicon extracted from Rayfield’s dictionary and the English lexicon generated from Terminesp and Termcat.

```
SELECT DISTINCT ?geoLabel ?originalTrans ?externalTrans
{
  ?dictGeo lime:language "kat";
    lime:entry ?geoEntry .
  ?geoEntry ontolex:sense ?geoSense ;
    ontolex:lexicalForm [ontolex:writtenRep ?geoLabel] .
  ?geoSense ontolex:isLexicalizedSenseOf ?concept .
  ?enSense ontolex:isLexicalizedSenseOf ?concept .
  ?enEntry ontolex:sense ?enSense ;
    ontolex:canonicalForm [ontolex:writtenRep ?originalTrans ] .
  ?enEntryTerminoteca ontolex:lexicalForm [ontolex:writtenRep ?originalTrans];
    ontolex:sense ?enSenseTerminoteca .
  ?enSenseTerminoteca ontolex:reference ?terminotecaConcept .
  ?otherSense ontolex:reference ?terminotecaConcept .
  ?otherEntry ontolex:sense ?otherSense ;
    ontolex:lexicalForm [ontolex:writtenRep ?externalTrans] .
} LIMIT 500
```

Example 7.9: SPARQL query to access English translations of Rayfield’s bilingual dictionary and other potential translations extracted from Terminoteca RDF

geoLabel	originalTrans	externalTrans
"აღლუმი"@kat	"review"@en	"Überprüfung"@de
"შემოთვლილობა"@kat	"instruction"@en	"Anweisung"@de
"ადგილსამყოფელი"@kat	"residence"@en	"Apartment-Anlage"@de
"აპარტამენტი"@kat	"apartment"@en	"Apartment"@de
"ავტორიტეტი"@kat	"authority"@en	"Behörde"@de
"ბიბლიოთეკა"@kat	"library"@en	"Bibliothek"@de
"სამწიგნობრო"@kat	"library"@en	"Bibliothek"@de
"წიგნისსაცავი"@kat	"library"@en	"Bibliothek"@de
"გამოსახულება"@kat	"image"@en	"Bild"@de
"თეჯირი"@kat	"screen"@en	"Bildschirm"@de
"ბურცი"@kat	"blister"@en	"Blase"@de
"ბუნგალო"@kat	"bungalow"@en	"Bungalow"@de
"ბუნგალო"@kat	"bungalow"@en	"Chalet"@de
"დიათერმია"@kat	"diathermy"@en	"Diathermie"@de
"ავტორიტეტი"@kat	"authority"@en	"Dienststelle"@de
"დიამეტრი"@kat	"diameter"@en	"Durchmesser"@de
"ელექტროენცეფალოგრამა"@kat	"electroencephalogram"@en	"EEG(Abkürzung)"@de
"ელექტროკარდიოგრამა"@kat	"electrocardiogram"@en	"EKG(Abkürzung)"@de
"ელექტროენცეფალოგრამა"@kat	"electroencephalogram"@en	"Elektroenzephalogramm"@de
"ელექტროკარდიოგრამა"@kat	"electrocardiogram"@en	"Elektrokardiogramm"@de
"ელექტროკარდიოგრაფია"@kat	"electrocardiography"@en	"Elektrokardiographie"@de
"ელექტროსკოპი"@kat	"electroscope"@en	"Elektroskop"@de
"ელექტრომკურნალობა"@kat	"electrotherapy"@en	"Elektrotherapie"@de
"ელექტროთერაპია"@kat	"electrotherapy"@en	"Elektrotherapie"@de
"ელექტროტონუსი"@kat	"electrotonus"@en	"Elektrotonus"@de

Figure 7.1: An extract of the results of the query shown in Example 7.9

Similarly, and focusing just on Terminoteca RDF and the differences between its terminologies (Termesp and Termcat), we can define a query to provide full forms of technical terms and their initialisms (Example 7.10). This relation is not encoded in Termesp, but it is present in Termcat. By linking the two terminologies, those terms in Termesp that were full forms but were not related to their initialism are now linked to them by leveraging Termcat data (if they were originally present in the latter one).

Example 7.11 presents a query that leverages the definitions in KD data to enrich the entries in Apertium RDF, and also provides the translations of those entries into Romanian (errors due to the lack of sense alignments). The results of this query are presented in Figure 7.2.

```

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX ontollex: <http://www.w3.org/ns/lemon/ontollex#>
PREFIX vartrans: <http://www.w3.org/ns/lemon/vartrans#>
PREFIX tr_cat: <http://purl.org/net/translation-categories#>
prefix lexinfo: <http://www.lexinfo.net/ontology/2.0/lexinfo#>
PREFIX lime: <http://www.w3.org/ns/lemon/lime#>
prefix dc: <http://purl.org/dc/elements/1.1/>

SELECT DISTINCT ?label ?initialism
WHERE {
    ?lexiconS lime:entry ?entry .
    ?entry ontollex:sense ?senseS .
    ?senseS ontollex:reference ?concept .
    ?entry ontollex:lexicalForm ?formS .
    ?formS ontollex:writtenRep ?label .
    ?entryT lexinfo:initialismFor ?entry .
    ?entryT ontollex:lexicalForm ?formT .
    ?formT ontollex:writtenRep ?initialism .
    ?lexiconS lime:language "en" .
} LIMIT 500

```

Example 7.10: SPARQL query that renders full forms of English technical terms and their initialisms

Full Form	Initialism
digital cassette	DAT
photoresistive cell	LDR
specific integrated circuit	ASIC
silicon bilateral switch	SBS
silicon-controlled switch	SCS
silicon unilateral switch	SUS
light-emitting diode	LED
charge-coupled device	CCD
surface mounted device	SMD
avalanche photodiode	APD
artificial intelligence	AI
complementary MOS	CMOS
printed circuit board	PCB
universal asynchronous receiver-transmitter	UART
universal synchronous and asynchronous receiver-transmitter	USART
silicon-controlled rectifier	SCR
uninterruptible power system	UPS
network interface card	NIC
gate turn-off thyristor	GTO
linear variable differential transformer	LVDT
unipolar transistor	FET
N-channel field-effect transistor	N-FET
carrier sense multiple access	CSMA
carrier sense multiple access with collision detection	CSMA/CD
code division multiple access	CDMA
time division multiple access	TDMA
computer-assisted learning	CAL
computer supported cooperative learning	CSCL

Table 7.6: Results of the query provided in Example 7.10

```

prefix lemon: <http://www.lemon-model.net/lemon#>
prefix lime: <http://www.w3.org/ns/lemon/lime#>
prefix ontollex: <http://www.w3.org/ns/lemon/ontollex#>
prefix owl: <http://www.w3.org/2002/07/owl#>
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix skos: <http://www.w3.org/2004/02/skos/core#>
prefix void: <http://rdfs.org/ns/void#>
prefix xml: <http://www.w3.org/XML/1998/namespace>
prefix xsd: <http://www.w3.org/2001/XMLSchema#>
prefix tr: <http://purl.org/net/translation#>

SELECT DISTINCT ?lemmaApertium ?KDdef ?tradApertium
{

?lexiconES a lime:Lexicon;
  lime:language "es" ;
  lime:entry ?lexicalEntry .
?lexiconESlemon a lemon:Lexicon;
  lemon:entry ?lemonEntry ;
  lemon:language "es" .

?lexicalEntry ontollex:lexicalForm [ontollex:writtenRep ?lemmaKD];
  ontollex:sense ?sense .
?sense ontollex:isLexicalizedSenseOf [skos:definition ?KDdef].
?lemonEntry lemon:lexicalForm [lemon:writtenRep ?lemmaApertium] .
?senseApertium lemon:isSenseOf ?lemonEntry .
?tr tr:translationSource ?senseApertium;
  tr:translationTarget ?targetApertium .
?targetApertium lemon:isSenseOf ?targetEntryApertium.
?targetEntryApertium lemon:lexicalForm [lemon:writtenRep ?tradApertium].

FILTER (lang(?tradApertium)="ro")
FILTER (?lemmaApertium = ?lemmaKD)

} LIMIT 500

```

Example 7.11: SPARQL query that shows entries from the Spanish Apertium lexicon and the definitions associated to entry with the same lemma in KD Spanish dictionary and translations to Romanian

"acreditar"@es	"dar crédito a una persona o cosa"@es	"acredita"@ro
"acreedor"@es	"persona que tiene derecho a cobrar una deuda"@es	"creanțier"@ro
"acrededor"@es	"persona que tiene derecho a cobrar una deuda"@es	"creditor"@ro
"acrededor"@es	"que merece un premio, recompensa, alabanza u otra cosa"@es	"creanțier"@ro
"acrededor"@es	"que merece un premio, recompensa, alabanza u otra cosa"@es	"creditor"@ro
"abdominal"@es	"relativo a la cavidad del cuerpo humano que va del pecho a las caderas"@es	"abdominal"@ro
"acribia"@es	"exactitud o minuciosidad en algo"@es	"acribie"@ro
"acrimonia"@es	"acritud"@es	"acrine"@ro
"acritud"@es	"cualidad de las cosas de sabor u olor áspero y penetrante"@es	"acreală"@ro
"acritud"@es	"aspeza en la forma de ser o expresarse"@es	"acreală"@ro
"acrobacia"@es	"ejercicio gimnástico de equilibrio o salto"@es	"acrobație"@ro
"acrobacia"@es	"salida ingeniosa para superar un problema"@es	"acrobație"@ro
"acróbata"@es	"persona que hace acrobacias en un espectáculo"@es	"acrobat"@ro
"abecedario"@es	"sistema de signos gráficos usados para escribir"@es	"abecedă"@ro
"acromegalia"@es	"enfermedad causada por una lesión de la glándula pituitaria"@es	"acromegalie"@ro
"abecedario"@es	"conjunto ordenado de las letras de un idioma"@es	"abecedă"@ro
"abedul"@es	"árbol de color blanco plateado de hasta 30 metros de altura"@es	"mesteacăn"@ro
"abeja"@es	"insecto que produce miel"@es	"albină"@ro
"abejar"@es	"lugar donde están las colmenas de las abejas"@es	"prisacă"@ro
"actitud"@es	"postura del cuerpo"@es	"atitudine"@ro
"actitud"@es	"disposición de ánimo para actuar"@es	"atitudine"@ro
"activación"@es	"puesta en funcionamiento de un mecanismo"@es	"activare"@ro
"activación"@es	"aumento de la intensidad o rapidez de algo"@es	"activare"@ro
"abejorro"@es	"insecto parecido a la abeja pero de cuerpo más gordo y velludo que emite un zumbido muy sonoro al volar"@es	"cărăbuș"@ro
"activar"@es	"poner en funcionamiento un mecanismo o máquina"@es	"activa"@ro
"activar"@es	"hacer más vivo e intenso algo"@es	"activa"@ro

Figure 7.2: An extract of the results of the query shown in Example 7.11

Chapter 8

Linked Data-based lexicography

This thesis has been focused so far on the modelling challenges and solutions to represent existent dictionaries as LLD, and the practical benefits this endeavour brings with respect to information preservation, enrichment of sources, or the easier integration of the data into NLP pipelines, as well as the benefits for their use in the representation of the macro- and the microstructure of lexicographic resources (see Chapter 7). In contrast, this chapter places LLD as a formalism to represent lexical data in the context of current work carried out in the branch of theoretical lexical semantics, and will propose the notion of *LD-native dictionaries*, i.e., dictionaries born in LLD and not converted from an original source format, their benefits, implications and challenges to tackle in order to make this a reality in the near future. Our goal with this chapter is therefore twofold: to locate LD in the context of lexicographic work in lexical networks, and to bring its benefits closer to the lexicographer as a basis for future endeavours. The analysis presented here has been published in Bosque-Gil et al. (2016a) and Gracia et al. (2017).

8.1 LD and lexical systems

WordNet already set a precedent (Miller (1995), Fellbaum (1998a)) as a graph-based lexico-semantic database where nodes represent the concepts (synsets or sets of cognitive synonyms) and hyponymy, meronymy and antonymy constitute the relations that link them together. Furthermore, there are other efforts in lexicography that emerge from a conception of the natural language lexicon as a network of entries rather than a

list, which is what the organisation of conventional dictionaries looks like. The entries are then viewed as part of a language system of related lexical elements. Polguère’s notion of a lexical system (Polguère, 2009, 2014) implemented in the French Lexical Network (Gader et al., 2012), falls into this category. However, in contrast to the projects developed in lexical semantics, LLD and the models proposed for converting resources into them (see Chapter 2), do not arise as initiatives to model the (mental) natural language lexicon, nor make such claim, even though they entail the use of classes and properties such as lexical entry, sense, lexical concept, syntactic frame, lexical form or definition. LD emerges as a technological means to better represent, share, integrate and discover linguistic knowledge scattered over the Web and its underlying RDF formalism is not conceived from a theoretical perspective as an alternative to study lexical information from the perspective of theoretical linguistics.

Nonetheless, knowing the direction into which lexical semantics and lexicography move, as well as the similarities between the representations suggested there and those proposed from an LD perspective, will help in building bridges for collaboration between experts from both sides. LD, as best practices for data representation, should be compatible with the representation of any lexical network, even though this implies the extension of vocabularies currently available on the LLOD cloud or models to encode all the data that the theory on which the resource is based addresses. An analysis of what modeling the lexicon as a graph in WordNet entails and which needs are met is given in Polguère (2014) and McCracken (2015): lexical entries were previously analysed and presented independently one from another and a novel approach reflecting what the structure of the lexicon might resemble was called for.

WordNet falls under the category of ontology-based lexical network (cf. Polguère (2014)), i.e. a network of lexical units with an ontology as backbone, including word senses arranged in a hierarchy and related by synonymy, hyponymy and meronymy relations. It is worth mentioning that LLD relies on linguistic ontologies or vocabularies, but the creation of an ontology of word senses or concepts is actually optional and it is not a required step in order to publish LD. Accordingly, we can state that the entry *enthusiasm* in an English lexicon has as part-of-speech `lexinfo:noun`, which is defined along with, for instance, `lexinfo:reflexivePersonalPronoun`, as an individual of type `lexinfo:PartOfSpeech` in LexInfo. We are thus linking two resources without establishing the concept denoted by *enthusiasm* in any hierarchy (e.g. as a child of *feeling*).

LD resources such as BabelNet, DBpedia, and WordNet RDF (see Chapter 2) have an underlying ontology, but this is not implied in the conversion of every resource to LLD. In relation to this, LLD builds upon the notion of semantics by reference (McCrae et al., 2012): the meaning of a word and the word itself (the *signifier* – *signified* opposition) are separated in two different layers, with `ontolex:LexicalEntry` and `skos:Concept` respectively, and the relation between the two is “reified” in a class that aims at encoding a sense (`ontolex:LexicalSense`). All the linguistic information pertaining to the word itself or to the use of that word with that specific meaning is separated from the actual meaning, which, ideally, is language independent. Hierarchic conceptual relations would be established, if they are, at the level of the concept.

Polguère places lexical systems on the other side of the balance: they are lexical networks that are not ontology-based. Lexical systems are conceived with the relations among the lexical elements as focus and relegate to the background the classification of units or property inheritance (Polguère, 2014). A key aspect of lexical systems is that relations are not limited to synonymy, hyponymy, etc. but they include paradigmatic and syntagmatic relations drawn from the set of lexical functions of the Meaning Text Theory (Mel’cuk, 1996). The result is a multi-dimensional graph with a wide range of relations linking the nodes (lexical elements), which instantly brings RDF to mind.

There are two important points to bear in mind when comparing lexical systems with resources migrated to LLD: first, the nodes in lexical systems are already ‘disambiguated’, each node represents one specific meaning of the lexical unit at hand (Polguère, 2014). The closest counterpart we have in RDF is `ontolex:LexicalSense`, which is a unique relation between a word and a meaning. Secondly, the nodes in a lexical system are not atomic and each one records the information we would find in a lexicographic article. Grammatical information, semantic label, syntactic government pattern (collocations are implemented by edges), etc. are stored inside the node. In LLD, some of these data would be linked to the entry at hand or to one of its `ontolex:LexicalSense` elements by means of specific properties (edges) and elements available in linguistic vocabularies, identifiable with their own URIs: lexical entries, word forms, senses, part-of-speech tags, gender, number, subcategorization, etc. To see which entries are related, SPARQL allows to perform queries on the graph and trace the connections between lexical senses or lexical entries. LexInfo, OntoLex-*lemon*, SKOS,

OLiA, etc. already provide a high number of relations, which can be extended with new ones or new vocabularies can be created as needed.

In sum, the idea of representing lexical information as a graph is not new, and LD are not presented as a novelty in this regard. However, they allow for the implementation of networks or the integration of already available ones on the basis of a homogeneous format. Thus LD meet the need for linking lexical elements that were previously isolated by using sets of relations and elements that are defined externally and can be extended as required, relying or not on an underlying ontology of word senses. This does not mean that LD is equivalent to any of the efforts mentioned above or forms a better option to the structures in which they are implemented, and it is not a model to understand natural language. RDF is, however, a model to represent data worth taking into consideration for lexicographic projects aiming at the creation of lexical networks because it provides a basis for their implementation while retaining all the benefits related to interoperability, visibility and NLP-services compliance. All in all, the LD paradigm is agnostic with respect to the different theories in modern lexicography, and it poses a number of tangible benefits that we enumerated in the previous sections in this thesis (Chapter 7).

8.2 LD-native lexicography

The ways in which dictionaries are compiled and used have evolved dramatically in recent years owing to the processes of digitization. This evolution has found in the Web an optimal means to empower the visibility and usability of dictionaries. In this context, we propose the notion of *LD-native dictionaries* as a natural next step in the evolution of lexicography. These future dictionaries could be LD-native and, as such, graph-based. Their nodes are not dependent on any internal hierarchy and are uniquely identified at a Web scale. In the following sections we analyse the advantages of such an approach and identify its possible impact on the dictionary representation, compilation, and usage processes. Some challenges related to interoperability and data aggregation issues are also discussed.

As a natural next step in the evolution of the dictionary concept, and grounded in the advantages of using LD to represent lexicographic content presented in this thesis and reported in the literature (see Section 2.5), we envisage **dictionaries that are**

born and evolve dynamically on the Web. While several experiences have been reported in the literature related to the conversion of different types of dictionaries as LD (Section 2.5), which illustrate the growing interest for LD in lexicography, the idea of developing dictionaries as LD **in a native way**, rather than converting already existent ones from their proprietary formats into LD, has received little attention so far. These will not be (only) the result of transforming lexicographic data from previous electronic formats into LD, but will ensue from compiling dictionaries as LD from scratch. Thus, such future dictionaries are LD-native and, as such, graph-based. Their nodes are not dependent on any internal hierarchy and are uniquely identified at a Web scale. This will enable the enhancement of a vast network of interconnected linguistic elements through semantically well-defined lexical, syntactic, pragmatic, etc. relations, through which lexicographers and users navigate to edit, query, or aggregate data. Links to other lexical resources, including other dictionaries, would thus be quickly and naturally established. However, compiling them from scratch in an RDF-based environment and directly following the LD principles will have an impact on the process of dictionary compilation, representation, and interoperation with other resources.

8.2.1 Building a graph-based reusable structure

A key aspect of an LD-based dictionary is that every lexical element (headword, sense, written form, grammatical attribute, etc.) is treated as a first-class node, being identified by its own URI at a Web scale, and being attached to its own descriptive information and linked to other relevant elements through RDF statements.

In an LD-native dictionary, a multiword expression such as *bow and scrape* will be a headword on its own with its own URI, and links will be drawn to relate it to the two parent entries *bow* and *scrape*, directly pointing to their suitable senses or homographs whenever appropriate. In that way, changes will be done in a single place, avoiding the need of copying information and reducing the risk of bad maintenance. This implies that an idiom or collocation, for instance, will not be encapsulated under the container of the entry in which it was originally defined, but will be related to it with the suitable property. Since the idiom now becomes a node, we are able to link it to any other node from any other entry in the dictionary. Similarly, LD solves the issue of maintaining cross-references mentioned in Chapter 7. Since entries and senses are now uniquely

identifiable throughout the dictionary data and graphs are not actually ordered, cross-references can be direct pointers to the entry or sense to which they are referring. Cross-references will not (only) be manual annotations for human consumption but real links between nodes in the dictionary graph.

Of course, the conversion of already existent XML-based dictionary data into LD might solve the aforementioned issues, and other similar ones, at the modelling level, but still not at the source. We argue that, by solving such issues at source, LD-native dictionaries will make lexicographers' work more efficient and will make the *consistency of lexicographic data* easier to maintain, given that redundancies are more easily avoided.

LD-native dictionaries will maximise *re-usability of lexical knowledge* during the lexicographic compilation process. For instance, a lexical entry can be characterised by synonyms. In a hierarchical arrangement, such synonyms are nested under their associated entry and there is no guarantee of their existence as lexical entries for themselves. In an LD set-up, each synonym is designed as a new node in the graph and then linked to the initial lexical entry through a synonymy relation. Such a new lexical entry only needs to be defined once, independently of the number of times it occurs in the dictionary.

External re-use of lexical knowledge is also enhanced via link declarations (in RDF) to other LD sources. That enables, for instance, to re-use grammatical categories already defined in external catalogues (e.g. LexInfo), import additional semantic descriptions from encyclopaedic resources such as BabelNet or DBpedia or connect different LD-based dictionaries.

Conceiving a dictionary as LD from scratch has also another advantage. In previous XML to LD conversion experiences, it was necessary to preserve as much information content as possible in order to keep the process reversible. This has led to the propagation of superfluous information into RDF, such as internal dictionary identifiers of the lexical elements or information related to how lexicographic data is displayed in a user interface. In the latter case, we argue that such information should be maintained apart from the purely lexical graph. In the former case, the definition of URIs for every lexical element makes the internal identifiers redundant. Further, well designed URIs will avoid collision of identifiers when integrating several dictionaries, which might be a risk if only internal dictionary identifiers were used.

8.2.2 Impact on the editorial process

In this section we analyse the impact LD-native dictionaries will have on the work of lexicographers in several aspects and their related challenges.

Modelling. For a dictionary to be created in LD, we first have to select the kind of information its entries will cover, and make sure this information is indeed representable as LD by available mechanisms. Once the information that an entry will capture is decided upon (syntactic, semantic, pragmatic, phonetic, etc.), the selection of available vocabularies, and the models to represent them, will proceed in order to create the model that will be the backbone of the editing tool that the lexicographer will be using to generate the data. Modelling challenges include the representation of the sense hierarchy, translations, examples, inflections, homographs or multimedia content in a way that stays true to the lexicographer’s view and maximises re-usability according to the LD principles. However, as mentioned above, the major shift that lexicography would experience involves a transition from a hierarchical ordering of the information recorded in a dictionary entry into a graph structure with its nodes uniquely identified by URIs, whose form should be also determined by the editor. The lexicographer will be required to identify the precise nature of the relation between two pieces of information by using ontological properties rather than unbounded textual descriptions. This echoes the difference between compiling dictionaries with only the human as target or creating them for both humans and computers.

Basic knowledge on LD. Even though expert knowledge of RDF and SPARQL should ideally not be required on the lexicographer’s part, they would need to assimilate the principles of LD lying at the heart of lexicographic compilation. By doing so, the editor will be able to unlock the potential of both using different URI naming strategies and linking to diverse external or internal resources to enrich their own data, for example.

Technical needs. Even though developing dictionaries natively in LD would allow them to be integrated into bigger knowledge systems and consumed by LD-aware NLP applications from the very beginning, the daily tasks that human users perform with the help of dictionary data should not be relegated to the background. In this respect, a clear challenge that we must face as we envision this ecosystem is the lack of

a well-established and solid mechanism for everyday dictionary users to query LD resources without the need to rely on Semantic Web and LD knowledge. In order to build LD-native dictionaries, tools for graph editing and visualisation would be required to enable the compilation without expert knowledge of Semantic Web formalisms. Natural language and guided interfaces on top of SPARQL would evolve into essential tools for the editor to query the different LD versions created during the editorial process and thus control the project's progress.

A paradigm-shift in lexicography would involve reconsidering the skills that are required both from lexicographers and editors as well as from the potential users of such linked dictionaries. Just as new natural language or guided interfaces will be called for in order for non-experts to query the datasets, their maintenance in terms of modification, enrichment and quality control on part of the editor will require new mechanisms as well.

Quality control. As reported in recent surveys on LD quality (Zaveri et al., 2016), there are aspects concerning data quality that are original to LD and therefore will need to be taken into account in LD-driven lexicography. Quality can be assessed through different dimensions, ranging from availability, licensing and security (accessibility dimensions) to data accuracy, consistency, etc. (intrinsic dimensions) and reputation and verifiability, among others (trust dimensions) (see Zaveri et al. (2016) for a state-of-the-art account on LD quality). Although each dictionary data provider may define its own criteria, they all share a common goal with respect to the intrinsic data, namely to provide lexical information that is semantically and syntactically correct, compact (i.e, without redundant data), complete (gathering all available data concerning an entry), and logically consistent (without contradictions or conflicting values). Processes aimed at evaluating the quality of the ontology in support of the dictionary editing phase, as well as for assessing the quality of the generated instances would need to take place as part of the regular lexicographic workflow.

8.2.3 Challenges of interconnecting LD dictionaries

In this subsection we discuss some of the challenges related to making LD-native dictionaries grow and interconnecting them. In particular, we discuss aspects related to interlinking and data integration.

Interlinking. As a first step, lexical or general conceptual resources would need to be identified as suitable linking targets (Vila-Suero et al., 2014b; Villazón-Terrazas and Corcho, 2011). Among the numerous datasets already available in the LLOD Cloud, BabelNet and DBpedia emerge as the conceptual encyclopaedic resources with the highest in-degree of links, thus acting as pivotal elements among multiple language datasets. LD-based systems aimed to support the automatic discovery and validation of such relations among language resources would be required to assist the lexicographer at this stage.

Data integration. Services should be developed on top of the LD-based ecosystem that, given a query, aggregate data from the different entries and offer users a unified representation. In this way, the system acts as a ‘single dictionary’ that is actually the sum and combination of many of them, which are in turn managed separately and developed independently.

The major challenge that we would meet here is the fact that information about the same dictionary entry would be sometimes repeated and scattered throughout the cloud of linked dictionaries. Each dictionary would be likely to show some differences in its underlying schema even though elements of the *de facto* standards had been re-used, especially if the editorial choice involved the use of a custom ontology. Some of the tasks that we would face in this stage have been already addressed in the literature in the LD integration context (Bleiholder and Naumann, 2009; Knap and Michelfeit, 2012), namely: schema matching, duplicate detection, and data fusion.

Schema matching refers to the detection of equivalent schema elements in the different sources (Bleiholder and Naumann, 2009). Proprietary schemas developed for the compilation of a dictionary often have equivalent counterparts in linguistic data category registries, such as LexInfo, but this is not always the case: mismatches between proprietary schema values for a specific DTD tag and individuals of an homologous class in an already available linguistic vocabulary can occur as well. Mappings between the dictionary editorial’s custom ontology and other models thus become crucial for overcoming these difficulties. **Duplicate detection** is the task of detecting equivalent resources to integrate data into one single and consistent representation (Bleiholder and Naumann, 2009). This means that information repeated across different linked dictionaries, e.g. the part of speech of a lexical item, should be presented only once in the answer to a query on the datasets. The problem arises when dissimilar values

are extracted from different dictionaries and conflicts need to be resolved as part of the **data fusion** step. Compatible values which however are different in granularity (e.g. *noun* and *common noun*) would need to be distinguished from different and contradictory ones for the same dictionary entry (e.g. *common noun* and *proper noun*). As reported in the literature (Bleiholder and Naumann, 2009; Knap and Michelfeit, 2012), these conflicts would need to be either avoided (in our *proper noun* and *common noun* example, no information about the part of speech would be given), ignored (both values are presented as parts of speech), or resolved with a set of conflict handling strategies, which, for example, identify some sources as more trustworthy than others.

8.3 Summary

LD is generating a rising interest in the area of lexicography, and many dictionaries have been already converted into LD. In this chapter, however, we have focused on what constitutes a step beyond by introducing the notion of LD-native lexicography. That is, dictionaries that will be compiled as LD from scratch. We have analysed the main advantages of this networked approach in contrast with the more traditional tree-oriented views, and have placed LLD in contrast to other approaches in lexicography based on lexical data representation as graphs. We have also discussed its potential impact on the editorial process and we have outlined some of the challenges to face in order to make this scenario a reality in the near future.

Chapter 9

Conclusion and Future Work

LLD emerged as an initiative to address the lack of interoperability across language resources and tools caused by the vast amount of data that is available in heterogeneous formats and scattered throughout the Web. Among these language resources, the number of dictionaries, lexica and terminologies continue to increase driven by the advances in digital humanities and the NLP community.

While numerous efforts have been devoted to the conversion of these resources to LLD for their future consumption by NLP tools, as well as for their better discovery and re-usability, until now there was no wide-scope exploration and assessment of the impact of the application of LLD to lexicography as a discipline: what the general representation needs are for the encoding of lexicographic resources as LLD, how current LLD models are to be applied or extended to meet such needs, the practical and theoretical benefits of doing so, the challenges and new possibilities that this paradigm holds if it is adopted as a means to create new lexicographic resources from scratch, and the open problems on the way for wider adoption.

Given that lexicographical resources are already increasingly being used in the context of NLP, this thesis has taken the opposite route and focuses on bringing LD technologies closer to lexicography in order to provide lexicographers with a motivation for the representation of their data in a way that maximises its discovery, reuse and integration with other external sources, and a walk-through on how to do so with solutions to account for information preservation as RDF during the process.

To attain this goal, our work has been guided by the following short-term objectives:

1. To advance the state of the art on models for the representation of LLD in order to support the representation of lexicographic data by minimising the loss of lexical information,
2. to facilitate the transformation of lexicographic resources to LLD by providing new application profiles for the *de facto* standard (OntoLex-*lemon*), extensions, practices, design decisions and a migrated set of representative examples in RDF,
3. to assess the impact of LLD in the lexicographic work-flow and in the use of lexicographic resources for data gathering and consultation by linguists and lexicographers.

In the following sections we review the contributions presented in this thesis, and how they help in attaining the above mentioned goals and in verifying the series of hypotheses presented in Chapter 3.3. In the process, we provide the conclusions we draw from the verification of hypotheses and the methodology followed to arrive at them.

9.1 Review of contributions

In order to attain the above mentioned goals, we divided our approach in four main research lines:

1. Identification of problems and challenges in the application of LLD to lexicographic resources,
2. proposal of modelling solutions and motivation of community initiatives to solve such problems,
3. conversion of lexicographic resources into LLD,
4. theoretical and query-based evaluation of the implications of LLD-based resources for lexicography, and for consultation and data gathering by linguists and lexicographers.

As lexicographic resources provide data that concern different levels of linguistic description, we first approached the analysis of the **state of the art** on LLD representation by proposing a **classification of models** in terms of the content they address.

Such a survey of models in LLD, conducted in the form of a systematic review by following a thematic approach based on the levels of linguistic analysis, was lacking. It allowed us to both detect the gaps in the literature as regards the application of LLD to lexicography, as well as to frame the different models in such a way that would facilitate to relate them to the different representation needs of lexicographic resources, following the path to attain our first goal.

This **survey** thus provides a new **frame to think of models to represent LLD** and clearly illustrates the proliferation of independently developed vocabularies that have a certain degree of overlap, specially with respect to lexical semantics and areas related to lexicography. This highlights that the representation of the syntax-semantics interface is still a non-trivial effort, as is the conversion of lexicographic resources as outcomes of a lexicographic compilation process to LLD.

Closely connected to the review of the state of the art, our second contribution refers to the **detection of challenges and limitations** of the current *de facto* standard for the representation of lexicographic resources, the OntoLex-*lemon* model, with a series of representative resources. In addition to the problems detected in the literature (Section 2.5), we addressed a range of information provided in dictionaries and terminologies, and assessed whether OntoLex-*lemon* and LexInfo accounted for such data for RDF representation (Chapters 4 and 5). While selecting specific resources limits the generalizability of the results, this approach served to compile a list of representation needs extracted from all of the resources to provide application profiles of the model and detect gaps. This, in turn, would serve as input for the extension (or generalisation) phase, involving the OntoLex community, in a later step.

Parallel to this contribution, we proposed **modelling solutions** to meet these representation needs in each different resource (Chapter 5). These two joint contributions thus address Limitation 2 (Section 3.1), the lack of guidelines to convert lexicographic resources as outcomes of a lexicographic compilation process to LLD. They allow us to facilitate the transformation of lexicographic resources to LLD, providing new application profiles, extensions, practices, design decisions and a migrated set of representative examples in RDF (Goal 2).

However, the analysis and modelling of resources was performed in a sequential order, starting with Apertium and Terminesp resources and finishing with K Dictionar-

ies and Rayfield’s Georgian-English Bilingual dictionary, which led to some differences among the RDF datasets.

Since these last two contributions (the detection of modelling difficulties and the proposal of application profiles to solve them) are focused on the specific selected resources, in order to identify further problems and challenges in the application of LLD to lexicographic resources, we **extended our analysis** and studied a series of entries of dictionaries in English, Spanish, and German (Chapter 6.1) in order to assess whether the detected modelling challenges also held for other resources beyond our converted set. This allowed us to confirm that some representation requirements identified in the resources we previously analysed were also common to other ones (e.g. the representation of dictionary structures), and it also provided us with more problematic cases with respect to the application of *OntoLex-lemon* to these resources (e.g. one dictionary entry with more than one part of speech not fitting the definition of `ontolex:LexicalEntry`).

With a list of gathered **issues**, we approached the community to **motivate a new module for the OntoLex-lemon model to represent lexicographic content**. The active participation, providing use cases, and signs of interest of the community in this endeavour, illustrates that the problems in converting lexicographic resources are a reality detected simultaneously in different efforts, and needed addressing in a standardised manner.

With **this module**, developed as a community initiative, we thus facilitate the transformation of lexicographic resources to LLD (Goal 2), and provide modelling examples of the use of that module to guide the conversion of lexicographic resources as outcomes of a lexicographic compilation process (addressing thereby Limitation 2). Though the module is not exhaustive, it does offer elements general enough to be reused in numerous cases.

Along with the profiles of *OntoLex-lemon* proposed in this thesis, we provide the data itself converted to RDF for Apertium, Terminoteca RDF, and Zhishi.lemon, which are open. K Dictionaries data is provided via its Lexicala API (closed data), and Rayfield’s Georgian-English Dictionary except as RDF will be managed by the Applied Computational Linguistics group at Goethe Universität Frankfurt according to the Linked Open Dictionaries (LiODi) project. As such, this set of resources is framed in the research line (3) mentioned above, the conversion of lexicographic data to LLD. It

also aims at serving as a pool of examples of the different application profiles suggested in this thesis.

This pool of examples served, in turn, to perform different kinds of queries targeted at the testing of hypotheses (see Chapter 7). One of these sets of queries was aimed at **evaluating** the features of dictionaries in LLD in terms of the different lexicographical access they allow, namely, **the onomasiological and semasiological views** (framed under the evaluation of Hypothesis H3.2). The results of these queries offer insight into how the same information available in the source lexicographic resource, which was originally compiled either concept-oriented or lemma-oriented, can now be easily accessed in both ways by using SPARQL and taking as entry point either the concept or the lemma. For this reason, this evaluation is a further step that helped us attain goal G3, to assess the impact of LLD in the lexicographic work-flow and in the use of lexicographic resources for data gathering and consultation by linguists and lexicographers. Though these queries provide a first basis to showcase this, a comparison with queries on the source (hierarchical) format (e.g. with XQuery¹⁸³) to evaluate whether the queries in SPARQL offer any advantage as opposed to queries in the source XML format is still lacking and left for future work. The advantages of querying graphs of lexicographic resources represented as LLD do not rely thus on the onomasiological and semasiological access options alone.

The previous evaluation was focused on the benefits of LLD representation for a single resource. The next evaluation, an query-based **evaluation** of the the implications of LLD-based resources for consultation and data gathering, highlights the **advantages of combining various resources as LLD with a shared model**, and it is framed on our Goal 3. A set of queries was used to shed light on how different complementary information can be accessed by a **single SPARQL query** for later processing. While these queries serve in this scenario to show the potential of LLD for this aim, provided a shared model, they are only a small subset of the potential queries that leverage all information in a series of datasets with a shared model. In order to insert this consultation in a real use case scenario for linguistic research, a previous step would be necessary: examples of data needs would have to be gathered from linguists and translated to SPARQL in order to compare these types of queries against the data com-

¹⁸³<https://www.w3.org/TR/xquery-31/>

pilation methods (via XML queries, interfaces, or manual consultation) that a control set of linguists would adopt in order to obtain the data needed for a particular study.

The last contribution motivates the theoretical benefits and challenges of LLD in the representation of lexicographic data as source format for new lexicographic projects (Chapter 8). The idea behind this contribution is to address the same limitation of the state of the art accounted for in the previous contribution from a different perspective – a theoretical one. This is done by focusing on the impact of the LLD conversion on the structure of the lexicographic resource and assessing whether the **new configuration and its features brings benefits or drawbacks** for the representation of the macro and micro-structure of the dictionary in comparison to the source format. Relying on the results of such an analysis, this contribution introduces the notion of *linked data dictionaries* and the challenges that they would give rise to. Even though this analysis is theoretical and it is not grounded on previous experiences in the generation of LLD native lexicographic resources, it provides a first overview of the challenges on the way and the next steps that future work on this line would need to follow.

9.2 Verification of hypotheses

In the following sections, we revisit each hypothesis, elaborate on the path towards its verification, and highlight the new knowledge obtained thereof.

H1. Current *de facto* standards for the representation of linguistic information as LLD and the available best practices for the generation of multilingual LLD provide mechanisms to capture lexical data contained in a range of NLP and human-oriented multilingual terminologies and NLP-oriented lexica, but fall short of preserving lexical data from lexicographic records (explicit, in the form of annotations or implicitly encoded through microstructural aspects) in mono- and multilingual general human-oriented trade dictionaries.

Discussion: H1 concerns the coverage of a series of features from lexicographic resources by current models. It is focused on the gaps detected in these models that prevent them from minimising the loss of lexical data in human-oriented dictionaries, i.e., which prove that they are not geared towards information preservation. Since H1

is focused on the detection of these gaps, is also attuned to H2, which addresses their closing. H1 was verified in Section 7.3.1 by assessing and reporting whether *lemon*, *OntoLex-lemon* and/or *LexInfo* covered a list of features extracted from the selected resources (and those studied during the extension phase of our methodology) and, if so, whether this coverage was shown as part of the Specification and documentation of these models and the BPMLOD guidelines, or as an application profile of the model provided in the literature. The result of such an analysis was provided in Table 7.1. From this analysis we can conclude that there are numerous features that are not covered by these models and that the literature has not provided much attention to the representation of such features following these models until now.

H2. New modelling solutions and application profiles of the *de facto* standard and extensions to it are able to represent both the microstructure of the lexicographic record as well as annotations found in mono- and multi-lingual general human-oriented trade dictionaries. Such modelling solutions and applications are able to preserve explicit lexical data in lexicographic entries and implicit lexical information reflected on structural aspects during their conversion into RDF.

Discussion: This hypothesis is verified both by proposing application profiles for *OntoLex-lemon* that meet the representation needs of the series of selected resources (to serve as guideline for future conversions), as well as by leading and creating, together with the *OntoLex* community, the *lexicog* extension to capture those features. This work is presented in Chapters 5 and 6, and the summary of the provided coverage is shown in Table 7.1 for comparison purposes with the previous state of the art. From the verification of this hypothesis we learn that the *OntoLex-lemon* model and *LexInfo* provide a solid basis, with elements that do in fact cover numerous features of lexicographic resources if used as suggested in the application profiles. In those cases in which the *lexicog* extension is required, we conclude that its elements (inspired by the micro-structure of lexicographic records) are general enough to allow for flexibility in the representation, while capturing the structure of the dictionary entry at the same time.

H3. The LD paradigm brings benefits to the lexicographic work: H3.1 By virtue of being based on graphs and URIs, the LD paradigm resolves the limitations of tree-based formats for the linking of lexicographic data in a resource by enforcing the reuse of its elements and the instantiation of typed relations. H3.2 By virtue of being based on graphs and URIs, the LD paradigm allows for onomasiological and semasiological access on the same lexicographic resource. H3.3 It is possible to reuse and integrate the content of lexicographic resources, if represented as RDF, in NLP systems for further lexicographic enrichment.

Discussion: H3 is verified by validating its three sub-hypotheses, H3.1, H3.2 and H3.3. The first one, H3.1, is addressed in Section 7.4. Based on our experiences in the modelling of lexicographic resources reported in Chapters 5 and 6.3, we provide an analysis of the benefits of LLD for the representation on the macro- and microstructure of lexicographic records that revisits the limitation of hierarchical formats and highlights the role of LD to attain reusability of elements across the resource. That same analysis, along with a set of queries performed on the RDF resources and provided in the same section, serves to illustrate the two different kinds of accesses that are now possible on the same resource, thereby validating H3.2. Lastly, H3.3. concerns the reuse of the RDF outcome of this thesis in NLP systems or in NLP external projects. This hypothesis is verified in Section 7.5 with the description of the projects and services that are currently reusing KD Global series RDF data and the Terminoteca RDF. After the validation of these three sub-hypotheses, we can conclude that LD brings benefits to lexicography due to the advantages for the representation lexicographic content as RDF (internal reusability, independence of the order of appearance of entries and their elements, onomasiological and semasiological access, etc.) without ruling out features that make of this format a suitable one for use of the data in NLP systems and services developed by external parties.

H4. Lexicographic resources represented as LLD allow linguists and lexicographers to formulate single queries that combine and filter lexical data from different sources to obtain data for their research, which would not be possible if the resources were not linked and did not share a common representation format.

Discussion: H4 is verified in Section 7.6 through a set of queries that provide a first glimpse on the potential ways to leverage complementary information in a series of datasets represented as LLD with a shared model, focusing on the extraction of indirect translations, definitions, and alternative forms for entries lacking these data in their description.

9.3 Implications and social impact

In this section we elaborate on the different ways in which our contributions lead to **further research on modern lexicography** and the main aspects that they leave open for future work on that same line.

1. The most evident implication concerns the raise of awareness among lexicographers about LD technologies, their benefits, the coverage of current models for the representation of lexicographic resources as RDF, and the availability of lexicographic works already converted to RDF. By focusing on the application of LD to lexicography from the looking glass of the lexicographer, the data owner of lexicographic works, or the user of dictionaries, LD are brought closer to the public non-expert in Semantic Web technologies, and presented independently of their practical benefits for the integration in NLP applications.
2. Chapter 8 introduces the basic notions of a new kind of dictionary, born as linked data, as seed of a LD-driven lexicography. This assumes a deviation from current widespread formats and a full adoption of the LD principles in the lexicographic workflow. However, this poses new challenges as well, such as the implementation of LD-data integration services and quality control mechanisms to ensure the lexical quality of the generated RDF and its conformance to the LD data models. In addition, interfaces on top of SPARQL would become crucial elements for such a scenario to become a reality.
3. With *OntoLex-lemon* being increasingly used to convert lexical resources to RDF, this thesis presents it (together with its extensions) as a solid mechanism to undertake lexicographic endeavours, and provides a general methodology to assess model coverage, to define application profiles, and to motivate and initiate module development as part of the *OntoLex-lemon* community. This keeps the door open

to more experts in lexicography testing the coverage of the model with their own resources (in their original format) and following a similar walk-through as the methodology presented in this work. However, this goes in parallel to the general tendency of the model being used for the conversion of resources, not only for ontology lexicalization.

4. Based on the representation needs in lexicographic resources that are met by the outcomes of this thesis, and starting from the queries to showcase the benefits of LD to query resources with isolated yet complementary data, future studies could address the different data needs of linguists in terms of use cases so as to assess whether LD brings benefits for hypothesis validation in empirical research in linguistics – provided that the representation needs of the lexical resources studied are covered. This line of work would be closely related to that of the LiODi project at Goethe Universität Frankfurt, and would be using LD-based lexicography as a means to investigate the implementation and use of LD technologies by linguists.

9.4 Future work

In the following, we summarise open lines of work to address as next steps.

1. Analysis of representative lexicographic resources from disparate type, various lexicographic traditions, and in languages of different family, **to broaden the list of representation needs and assess the generalizability of the proposed applications profiles and the *lexicog* module**. Most languages addressed in this thesis are majority (English, German, Spanish), and some of them well represented in the LLOD cloud. Others, as Georgian or Portuguese (translations from KD’s Spanish dictionary) still have scarce resources as LLD. Different languages and lexicographic traditions trigger assorted representations needs distinct from the ones we detected in Chapter 4 and 6 (e.g. see Jarrar et al. (2019) for different diacritization representations in Arabic lexical entries). The type of lexicographic work, e.g. a collocation dictionary in contrast to a historical one, also affects the information present in the microstructure and thus will need careful analysis so as to detect gaps in (1) our applications profiles and in the *lexicog* module, and (2) linguistic category registries such as LexInfo.

2. Analysis of lexicographic **resources conceived in the framework of a particular theory of language** in order to enable querying across them and the comparison of the theory-based linguistic models that support them. Some lexicographic projects such as The Explanatory Combinatorial Dictionary (ECD) (Mel’cuk and Polguère, 1987) or the Pattern Dictionary of English Verbs (PDEV)¹⁸⁴ belong to a theory of language from the area of lexical semantics. As Trippel (2006) points out, these resources do not often have wide coverage but their information and granularity is very rich, and they serve as implementation of these linguistic models of language with actual data. Converting this kind of resources to RDF (as the implementation of PDEV in *lemon* by El Maarouf et al. (2014)) is challenging in terms of modelling, and it often requires the development of custom ontologies. However, representing these resources as RDF and **aligning the underlying models** would enable queries to support **contrastive analysis of both the implementations as well as the theories themselves**.
3. In relation to the previous aspect, development of mechanisms to establish **schema levels links** (Anjomshoaa et al., 2014; Jain et al., 2010; Millard et al., 2010) **across linguistic ontologies**, which would help reasoners from inferring more knowledge about linguistic facts. A next step in this direction is the **application of ontology matching techniques** to establish links across the different linguistic models and the development of specific reasoners ready to deal with linguistic knowledge. Due to the difficulty in assessing mappings between linguistic models (as different conceptualizations can be based on disparate theories of language), the involvement of linguists in the work-flow becomes a crucial aspect.
4. Development of mechanisms to link the generated resources to **BabelNet at the sense level**, as well as to **infer translations from the language pairs** available in the generated RDF graph. The underlying aim is the creation of a pool of connected datasets that can be used as a reliable source of (indirect) translations and usage examples.
5. Research on **natural language-based and Web-based interfaces to SPARQL**, as this remains one of the major obstacles in the path for the ap-

¹⁸⁴<http://pdev.org.uk>

plication and widespread adoption of the LD paradigm beyond the Semantic Web community.

In this thesis we have addressed the interface between LLD and lexicography from the looking glass of the user interested in lexicography, the lexicographer, or the linguist interested in lexical semantics who consults lexicographic resources. The above mentioned contributions aim at detecting and resolving obstacles on the way for wider LLD adoption in lexicography, as well as at analysing and showcasing the benefits and implications of this format for dictionary representation, both as a target format as well as a potential source format for lexicographic projects in the future. This thesis thus constitutes a step towards a new kind of lexicography which has LD at its very core.

Bibliography

- Abromeit, F., Chiarcos, C., Fäth, C., and Ionov, M. (2016a). 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 (LDL-2016): Managing, Building and Using Linked Language Resources, Portoroz, Slovenia*, pages 11–19. 28, 31, 54, 73
- Abromeit, F., Chiarcos, C., Fäth, C., and Ionov, M. (2016b). Linking the tower of babel: modelling a massive set of etymological dictionaries as rdf. In *LDL 2016 5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources*, page 11. 108, 198
- Adamska-Sałaciak, A. (2014). Bilingual lexicography: Translation dictionaries. *International handbook of modern lexis and lexicography*, pages 1–11. 93
- Albertoni, R., De Martino, M., Di Franco, S., De Santis, V., and Plini, P. (2014). Earth: an environmental application reference thesaurus in the linked open data cloud. *Semantic Web*, 5(2):165–171. 102
- Alexiev, V. and Casamayor, G. (2016). FN goes NIF: Integrating FrameNet in the NLP Interchange Format. In *LDL 2016 5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources*, page 1. 44
- American Heritage Dictionary Online (2017). *American Heritage Dictionary*. Accessed 13.07.2017. 199
- 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. 70, 285

- Aparicio, J., Taulé, M., and Martí, M. A. (2008). Ancora-verb: A lexical resource for the semantic annotation of corpora. In *LREC*. 15
- Atkins, B. S. and Rundell, M. (2008). *The Oxford guide to practical lexicography*. Oxford University Press. xvii, 14, 15, 16, 17, 92, 93
- Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer. 118
- Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics. 15
- Berners-Lee, T. (2003). Www past & future. <http://www.w3.org/2003/Talks/0922-rsoc-tbl/>. 5
- Berners-Lee, T., Fielding, R., and Masinter, L. (2004). Uniform resource identifier (uri): Generic syntax. Technical report. 8
- Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The semantic web. *Scientific American*, 284(5):34–43. 4
- Bindi, R., Calzolari, N., Monachini, M., Pirrelli, V., and Zampolli, A. (1994). Corpora and computational lexica: Integration of different methodologies of lexical knowledge acquisition. *Literary and Linguistic Computing*, 9(1):29–46. 12
- Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked data-the story so far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22. 8
- Bleiholder, J. and Naumann, F. (2009). Data fusion. *ACM Computing Surveys (CSUR)*, 41(1):1. 273, 274
- Bond, F., Vossen, P., McCrae, J. P., and Fellbaum, C. (2016). Cili: the collaborative interlingual index. In *Proceedings of the Global WordNet Conference*, volume 2016, pages 50–57, Amsterdam, The Netherlands. Global WordNet Association. 49

- Borin, L., Dannells, D., Forsberg, M., and McCrae, J. P. (2014). Representing swedish lexical resources in rdf with lemon. In *Proceedings of the 2014 International Conference on Posters & Demonstrations Track - Volume 1272*, ISWC-PD'14, pages 329–332, Aachen, Germany, Germany. CEUR-WS.org. 43
- Bosque-Gil, J., Gracia, J., Aguado-de Cea, G., and Montiel-Ponsoda, E. (2015). Applying the ontolex model to a multilingual terminological resource. In *International Semantic Web Conference*, pages 283–294. Springer. 83, 103, 104, 144, 161
- Bosque-Gil, J., Gracia, J., and Gómez-Pérez, A. (2016a). Linked data in lexicography. *Kernerman Dictionary News*, 24:19–24. 198, 247, 265
- Bosque-Gil, J., Gracia, J., and Montiel-Ponsoda, E. (2017). Towards a Module for Lexicography in OntoLex. In *1st Workshop on the OntoLex Model (OntoLex-2017), 18 June, Galway, Ireland*. 200, 232
- Bosque-Gil, J., Gracia, J., Montiel-Ponsoda, E., and Aguado-de Cea, G. (2016b). Modelling multilingual lexicographic resources for the Web of Data: The K Dictionaries case. In *GLOBALEX 2016 Lexicographic Resources for Human Language Technology Workshop Programme*, pages 65–72. 28, 83, 109, 164, 198, 223, 232
- Bosque-Gil, J., Gracia, J., Montiel-Ponsoda, E., and Gómez-Pérez, A. (2018). Models to represent linguistic linked data. *Natural Language Engineering*, 24(6):811–859. 27
- Bosque-Gil, J., Lonke, D., Gracia, J., and Kernerman, I. (2019). Validating the ontolex-lemon lexicography module with k dictionaries' multilingual data. In *Proceedings of eLex 2019 conference*, page to appear. 220
- Bosque-Gil, J., Montiel-Ponsoda, E., Gracia, J., and Aguado-de Cea, G. (2016c). Terminoteca rdf: a gathering point for multilingual terminologies in spain. In *Proceedings of TKE 2016 the 12th International conference on Terminology and Knowledge Engineering*, pages 136–146, Copenhagen. Copenhagen Business School. 83, 102, 103, 106, 144
- Buitelaar, P., Arcan, M., Iglesias, C., Sanchez-Rada, F., and Strapparava, C. (2013). Linguistic linked data for sentiment analysis. In *Proceedings of the 2nd Workshop on*

- Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, terminologies and other language data*, pages 1–8. Association for Computational Linguistics. 65, 66
- Buitelaar, P., Cimiano, P., McCrae, J., Montiel-Ponsoda, E., and Declerck, T. (2011). Ontology lexicalisation: The lemon perspective. 118
- Buitelaar, P., Declerck, T., Frank, A., Racioppa, S., Kiesel, M., Sintek, M., Engel, R., Sonntag, D., Loos, B., Micelli, V., Porzel, R., and Cimiano, P. (2006). Linginfo: Design and applications of a model for the integration of linguistic information in ontologies. In *Proceedings of OntoLex06, a Workshop at LREC*, pages 28–32, Genoa, Italy. o.A. 35, 60
- Caracciolo, C., Stellato, A., Morshed, A., Johannsen, G., Rajbhandari, S., Jaques, Y., and Keizer, J. (2013). The agrovoc linked dataset. *Semantic Web*, 4(3):341–348. 102
- Chiarcos, C. (2012). Powla: Modeling linguistic corpora in owl/dl. In Simperl, E., Cimiano, P., Polleres, A., Corcho, O., and Presutti, V., editors, *The Semantic Web: Research and Applications*, pages 225–239, Berlin, Heidelberg. Springer Berlin Heidelberg. 64, 66
- Chiarcos, C. (2014). Towards interoperable discourse annotation. discourse features in the ontologies of linguistic annotation. In Chair), N. C. C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, Paris, France. European Language Resources Association (ELRA). 50, 53, 63
- Chiarcos, C. and Erjavec, T. (2011). OWL/DL formalization of the MULTTEXT-East morphosyntactic specifications. In *LAW V ’11: Proceedings of the 5th Linguistic Annotation Workshop*, pages 11–20, Stroudsburg, PA, USA. Association for Computational Linguistics. 63, 75
- 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 *LREC*, pages 4463–4471, Paris, France. European Language Resources Association (ELRA). 70

- Chiarcos, C., Fäth, C., and Sukhareva, M. (2016b). Developing and Using the Ontologies of Linguistic Annotation (2006-2016). In *LDL 2016 5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources*, pages 63–72. 63
- Chiarcos, C., Hellmann, S., and Nordhoff, S. (2012a). Linking linguistic resources: Examples from the open linguistics working group. In *Linked Data in Linguistics*, pages 201–216. Springer. 9, 28, 42
- Chiarcos, C., Hellmann, S., and Nordhoff, S. (2012b). The Open Linguistics Working Group of the Open Knowledge Foundation. In *Linked Data in Linguistics*, pages 153–160. Springer, Berlin, Heidelberg. 9, 28
- Chiarcos, C. and Ionov, M. (2019). Ligt: An llod-native vocabulary for representing interlinear glossed text as rdf. In *2nd Conference on Language, Data and Knowledge (LDK 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik. 65, 66
- Chiarcos, C., McCrae, J., and Cimiano, P. (2013). Towards open data for linguistics: Linguistic linked data. In *New Trends of Research in Ontologies and Lexical Resources*, pages 7–25. Springer Berlin Heidelberg. 28
- Chiarcos, C., Nordhoff, S., and Hellmann, S. (2012c). *Linked data in linguistics: Representing and connecting language data and language metadata*. Springer, Berlin, Heidelberg. 28
- Chiarcos, C., Pagé-Perron, É., Khait, I., Schenk, N., and Reckling, L. (2018). Towards a linked open data edition of sumerian corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. 94
- Chiarcos, C. and Sukhareva, M. (2014a). Linking Etymological Databases. A Case Study in Germanic. In *3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing*, pages 41–49, Stroudsburg, Pennsylvania, US. Association for Computational Linguistics. 28, 31, 54, 73
- Chiarcos, C. and Sukhareva, M. (2014b). 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)*, pages 41–49. 108

- Chiarcos, C. and Sukhareva, M. (2015a). OLiA - Ontologies of Linguistic Annotation. *Semantic Web*, 6(4):379–386. 28, 31, 61, 75
- Chiarcos, C. and Sukhareva, M. (2015b). OLiA—ontologies of linguistic annotation. *Semantic Web*, 6(4):379–386. 102
- Chiarcos, C., Walther, D., and Ionov, M. (2017). From language documentation data to lld: A case study in turkic lemon dictionaries. In *LDK Workshops*, pages 22–32. 109
- Cimiano, P., Buitelaar, P., McCrae, J., and Sintek, M. (2011a). LexInfo: A declarative model for the lexicon-ontology interface. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(1):29–51. 31, 34, 60, 75
- Cimiano, P., Buitelaar, P., McCrae, J., and Sintek, M. (2011b). Lexinfo: A declarative model for the lexicon-ontology interface. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(1):29–51. 102
- 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 Workshop OntoLex - From Text to Knowledge: The Lexicon/Ontology Interface; held in conjunction with ISWC 2007*, number Vol. 7, pages 1 – 12. 35, 60
- Cimiano, P., McCrae, J. P., Rodríguez-Doncel, V., Gornostay, T., Gómez-Pérez, A., Siemoneit, B., and Lagzdins, A. (2015a). Linked Terminologies: Applying Linked Data Principles to Terminological Resources. In *Proceedings of the eLex 2015 Conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom*, pages 504–517. Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd. 58, 74, 144
- Cimiano, P., McCrae, J. P., Rodríguez-Doncel, V., Gornostay, T., Gómez-Pérez, A., Siemoneit, B., and Lagzdins, A. (2015b). Linked terminologies: applying linked data principles to terminological resources. In *Proceedings of the eLex 2015 Conference*, pages 504–517. 163
- Collins Cobuild Advanced Learner’s English Dictionary (2003). *Collins Cobuild Advanced Learner’s English Dictionary*. 199

- Copestake, A. (1992). The acquilex lkb: representation issues in semi-automatic acquisition of large lexicons. In *Proceedings of the third conference on Applied natural language processing*, pages 88–95. Association for Computational Linguistics. 12, 13
- Corcho, O., Fernández-López, M., and Gómez-Pérez, A. (2003). Methodologies, tools and languages for building ontologies. where is their meeting point? *Data & knowledge engineering*, 46(1):41–64. 32
- Corcoglioniti, F., Rospocher, M., Aprosio, A., and Tonelli, S. (2016). PreMOn: a Lemon Extension for Exposing Predicate Models as Linked Data. In *LREC*, pages 877–884, Paris, France. European Language Resources Association (ELRA). 42, 44, 52
- Davide Picca, A. M. G. and Gangemi, A. (2008). Lmm: an owl-dl metamodel to represent heterogeneous lexical knowledge. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, B. M. J. M. J. O. S. P. D. T., editor, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)*, pages 2413–2419, Marrakech, Morocco. European Language Resources Association (ELRA). <http://www.lrec-conf.org/proceedings/lrec2008/>. 50, 53
- de la Clergerie, E. and Clément, L. (2005). Maf: a morphosyntactic annotation framework. pages 90–94. 34
- de Melo, G. (2015). Lexvo. org: Language-related information for the linguistic linked data cloud. *Semantic Web*, 6(4):393–400. 31
- Declerck, T. (2006). Synaf: Towards a standard for syntactic annotation. In *Proceedings of the 5th edition of the Language Resources and Evaluation Conference (LREC)*, pages 229–232, Paris, France. European Language Resources Association (ELRA). 34
- Declerck, T. and Krieger, H.-U. (2014). Harmonization of German Lexical Resources for Opinion Mining. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, pages 3872–3876, Paris, France. European Language Resources Association (ELRA). 65, 66
- Declerck, T., Lendvai, P., Mörtz, K., Budin, G., and Váradi, T. (2012). Towards linked language data for digital humanities. In *Linked Data in Linguistics*, pages 109–116. Springer, Berlin, Heidelberg, Germany. 28

- Declerck, T., McCrae, J., Navigli, R., Zaytseva, K., and Wissik, T. (2019). ELEXIS - European Lexicographic Infrastructure: Contributions to and from the Linguistic Linked Open Data. 257
- Declerck, T. and Mörtz, K. (2016). Towards a Sense-based Access to Related Online Lexical Resources. In *Proceedings of the XVII EURALEX International Congress 2016, 6 – 10 September, Tbilisi, Georgia*, pages 660–667. Ivane Javakhishvili Tbilisi State University. 55, 57
- Declerck, T. and Wandl-Vogt, E. (2015). Towards a Pan European Lexicography by Means of Linked (Open) Data. In *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference*, pages 342–355. Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd. 28, 54, 57, 73, 108, 116, 198
- 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*. 14
- 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 & policy*, 10(1):45–53. 32
- Duden (2017). *Duden Online Wörterbuch Online*. Accessed 13.07.2017. 199
- Dyba, T., Dingsoyr, T., and Hanssen, G. K. (2007). Applying systematic reviews to diverse study types: An experience report. In *Empirical Software Engineering and Measurement, 2007. ESEM 2007. First International Symposium on*, pages 225–234. IEEE. 32
- Eckart, T., Bosch, S., Goldhahn, D., Quasthoff, U., and Klimek, B. (2019). Translation-based dictionary alignment for under-resourced bantu languages. In *2nd Conference on Language, Data and Knowledge (LDK 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik. 42
- Eckle-Kohler, J., McCrae, J. P., and Chiacros, C. (2015). LemonUby - A large, inter-linked, syntactically-rich lexical resource for ontologies. *Semantic Web*, 6(4):371–378. 43, 52, 59, 60

- 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 Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 401–408. European Language Resources Association (ELRA). 60, 118
- 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): Multilingual Knowledge Resources and Natural Language Processing at the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland*, pages 88–93, Paris, France. European Language Resources Association (ELRA). 45, 52, 60, 285
- El Maarouf, I., Mousselly Sergieh, H., Alferov, E., Wang, H., Fang, Z., and Cooper, D. (2015). The guanxi network: a new multilingual llo for language learning applications. In *Proceedings of the Second Workshop on Natural Language Processing and Linked Open Data*, pages 42–51. INCOMA Ltd. Shoumen, BULGARIA. 28
- Falk, I. and Stein, A. (2016). "lvf-lemon — towards a linked data representation of "les verbes français"". In Chair), N. C. C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2401–2406, Paris, France. European Language Resources Association (ELRA). 45, 52
- Fang, Z., Wang, H., Gracia, J., Bosque-Gil, J., and Ruan, T. (2016). Zhishi. lemon: On publishing zhishi. me as linguistic linked open data. In *International Semantic Web Conference*, pages 47–55. Springer. 83, 119
- Farrar, S. and Langendoen, D. (2003). A linguistic ontology for the semantic web. *Glott International*, 7(3):97–100. 31, 41, 59, 75
- Fellbaum, C. (1998a). A semantic network of english: the mother of all wordnets. In *EuroWordNet: A multilingual database with lexical semantic networks*, pages 137–148. Springer. 265

Fellbaum, C. (1998b). *WordNet*. Blackwell Publishing Ltd, Hoboken, New Jersey, US.

33

Fellbaum, C. (2005). WordNet and wordnets. In Brown, K. and Others, editors, *Encyclopedia of Language and Linguistics, Second Edition*, pages 665–670. Elsevier, Oxford, second edition edition. 33

Flati, T. and Navigli, R. (2014). Three birds (in the LLOD cloud) with one stone: BabelNet, Babelfy and the Wikipedia Bitaxonomy. In *Proceedings of SEMANTiCS*, pages 10–13. CEUR-WS.org. 28

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)*, pages 145–155. The COLING 2016 Organizing Committee. 47, 53

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O’Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., and Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144. 7, 83, 95, 126

Forkel, R. (2014). The cross-linguistic linked data project. In *3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing*, pages 60–66, Paris, France. European Language Resources Association (ELRA). 58, 74

Francopoulo, G. and George, M. (2006). Lexical markup framework (LMF). In *Proceedings of the 5th edition of the Language Resources and Evaluation Conference (LREC)*, pages 233–236, Paris, France. European Language Resources Association (ELRA). 19, 34, 95

Gader, N., Lux-Pogodalla, V., and Polguère, A. (2012). Hand-crafting a lexical network with a knowledge-based graph editor. In *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon*, pages 109–126. 266

Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., and Schneider, L. (2002). Sweetening ontologies with DOLCE. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 166–181. Springer Berlin Heidelberg. 36, 50, 59

- Gillis-Webber, F. et al. (2019a). The shortcomings of language tags for linked data when modeling lesser-known languages. In *2nd Conference on Language, Data and Knowledge (LDK 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik. 63, 139
- Gillis-Webber, F., Tittel, S., and Keet, C. (2019b). A model for language annotations on the web. *Knowledge Graphs and Semantic Web. KGSWC 2019. Communications in Computer and Information Science*, 1029. 63, 75, 139
- 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*, pages 3–12, New York, NY, USA. ACM. 9, 11, 85, 91, 118, 123
- Gonzalez-Agirre, A., Laparra, E., and Rigau, G. (2012). Multilingual central repository version 3.0. In Chair), N. C. C., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 2525–2529, Paris, France. European Language Resources Association (ELRA). 49
- González-Blanco, E., del Rio, G., and Cantón, C. M. (2016). A proposal to solve interoperability issues between poetic repertoires. In *Proceedings 5th Workshop on Linked Data in Linguistics*, pages 77–81, Paris, France. European Language Resources Association (ELRA). 28
- Good, J. and Hendryx-Parker, C. (2006). Modeling contested categorization in linguistic databases. In *Proceedings of the EMELD 2006 Workshop on Digital Language Documentation: Tools and standards: The state of the art*, pages 20–22. Wayne State University, Eastern Michigan University. 63
- Gracia, J., Kernerman, I., and Gil, J. B. (2017). Toward linked data-native dictionaries. In *Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference*, pages 550–559. Lexical Computing. 247, 265
- 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. 69

- 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, Volume 936*, pages 41–45, Aachen, Germany. CEUR-WS. org. 69
- Gracia, J., Montiel-Ponsoda, E., Vila-Suero, D., and Aguado-De-Cea, G. (2014a). Enabling language resources to expose translations as linked data on the web. In Chair), N. C. C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 409–4013, Paris, France. European Language Resources Association (ELRA). 56, 97, 125
- Gracia, J., Montiel Ponsoda, E., Vila Suero, D., and Aguado de Cea, G. (2014b). Enabling language resources to expose translations as linked data on the web. 94, 103, 144, 145, 146
- Gracia, J., Villegas, M., Gómez-Pérez, A., and Bel, N. (2018). The apertium bilingual dictionaries on the web of data. *Semantic Web*, 9(2):231–240. 7, 37, 56, 60, 83, 91, 94, 95, 108, 126, 127, 249
- Gratta, R. D., Frontini, F., Khan, F., and Monachini, M. (2013). Converting the PAROLE SIMPLE CLIPS Lexicon into RDF with lemon. *Semantic Web*, 6(4):387–392. 47
- Gritz, M. (2018). Lexical meaning formal representations enhancing lexicons and associated ontologies. In *NL4AI@AI*IA*. 50
- Gromann, D. and Declerck, T. (2019). Towards the detection and formal representation of semantic shifts in inflectional morphology. In *2nd Conference on Language, Data and Knowledge (LDK 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik. 55, 73
- Hanks, P. (2004). Corpus pattern analysis. In *Euralex Proceedings*, volume 1, pages 87–98, Lorient, France. Université de Bretagne-Sud. 44
- Hanks, P. (2007). Pattern Dictionary of English Verbs (PDEV)–Project Page. 44

- Hanks, P. (2013). *Lexical analysis: Norms and exploitations*. Mit Press, Cambridge, MA, US. 45
- Hanks, P. and Pustejovsky, J. (2005). A pattern dictionary for natural language processing. *Revue Française de linguistique appliquée*, 10(2):63–82. 15, 45
- Hartmann, R. R. (2001). *Teaching and researching lexicography*. Routledge. 15, 16
- Hayes, B., Curtiss, S., Szabolcsi, A., Stowell, T., Stabler, E., Sportiche, D., Koopman, H., Keating, P., Munro, P., and Hyams, N. (2013). *Linguistics: An introduction to linguistic theory*. John Wiley & Sons, Hoboken, New Jersey, US. 32
- Hellmann, S. (2013). *Integrating natural language processing (NLP) and language resources using linked data*. PhD thesis, Universität Leipzig. 27
- Hellmann, S., Lehmann, J., Auer, S., and Brümmer, M. (2013). Integrating NLP using linked data. In *International Semantic Web Conference*, pages 98–113, Berlin, Heidelberg. Springer. 28, 59, 64, 66
- Hodge, G. (2000). Systems of knowledge organization for digital libraries: Beyond traditional authority files. clir pub91. april 2000. 102
- Horrocks, I., Parsia, B., Patel-Schneider, P., and Hendler, J. (2005). Semantic web architecture: Stack or two towers? In *International Workshop on Principles and Practice of Semantic Web Reasoning*, pages 37–41. Springer. 5
- Ide, N., Baker, C., Fellbaum, C., Fillmore, C., and Passonneau, R. (2008). Masc: The manually annotated sub-corpus of american english. In *6th International Conference on Language Resources and Evaluation, LREC 2008*, pages 2455–2460. European Language Resources Association (ELRA). 15
- Ide, N. and Pustejovsky, J. (2010). What does interoperability mean, anyway. *Toward an operational definition of interoperability*. 4
- Ide, N. and Romary, L. (2004). International standard for a linguistic annotation framework. *Natural language engineering*, 10(3-4):211–225. 32, 34

- Ide, N. and Suderman, K. (2007). GrAF: A graph-based format for linguistic annotations. In *Proceedings of the Linguistic Annotation Workshop*, pages 1–8. Association for Computational Linguistics. 34
- 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. 32
- Jain, P., Hitzler, P., Yeh, P. Z., Verma, K., and Sheth, A. P. (2010). Linked data is merely more data. In *AAAI Spring Symposium: linked data meets artificial intelligence*, volume 11, pages 82–86. AAAI Publications. 70, 285
- Jarrar, M., Amayreh, H., and McCrae, J. P. (2019). Representing arabic lexicons in lemon - a preliminary study. In Declerck, T. and McCrae, J. P., editors, *LDK (Posters)*, volume 2402 of *CEUR Workshop Proceedings*, pages 29–33. CEUR-WS.org. 284
- Kemps-Snijders, M., Windhouwer, M., Wittenburg, P., and Wright, S. E. (2008). Isocat: Corraling data categories in the wild. In *6th International Conference on Language Resources and Evaluation (LREC 2008)*. 102
- Khalfi, M., Nahli, O., and Zarghili, A. (2016). Classical dictionary al-qamus in lemon. In *2016 4th IEEE International Colloquium on Information Science and Technology (CiSt)*, pages 325–330. IEEE. 198
- Khan, A. (2018). Towards the representation of etymological data on the semantic web. *Information*, 9(12):304. 55, 73
- Khan, F., Bellandi, A., Boschetti, F., and Monachini, M. (2017). The challenges of converting legacy lexical resources to linked open data using ontolex-lemon: The case of the intermediate liddell-scott lexicon. In *Proceedings of the LDK workshops: OntoLex, TIAD and Challenges for Wordnets, Galway, Ireland*, pages 1–8. 94
- Khan, F., Bellandi, A., and Monachini, M. (2016a). 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)*, pages 164–171, Osaka, Japan. The COLING 2016 Organizing Committee. 54

- Khan, F., Boschetti, F., and Frontini, F. (2014). Using lemon to Model Lexical Semantic Shift in Diachronic Lexical Resources. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics (LDL): Multilingual Knowledge Resources and Natural Language Processing at the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, pages 49–53, Paris, France. European Language Resources Association (ELRA). 28, 31, 54, 73
- Khan, F., Díaz-Vera, J. E., and Monachini, M. (2016b). Representing polysemy and diachronic lexico-semantic data on the semantic web? In *SWASH@ ESWC*, pages 37–46. 94, 198
- Khan, F., Díaz-Vera, J. E., and Monachini, M. (2016c). Representing Polysemy and Diachronic Lexicosemantic Data on the Semantic Web. In *Proceedings of the Second International Workshop on Semantic Web for Scientific Heritage co-located with 13th Extended Semantic Web Conference (ESWC 2016)*, pages 37–46. CEUR-WS.org. 54, 55, 73, 177
- Khan, F., Frontini, F., Boschetti, F., and Monachini, M. (2016d). Converting the Liddell Scott Greek-English Lexicon into Linked Open Data using lemon. In *Digital Humanities 2016: Conference Abstracts*, pages 593–596, Kraków, Poland. Jagiellonian University and Pedagogical University. 55, 73
- 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 (GL2013)*, pages 62–69, Stroudsburg, PA, USA. Association for Computational Linguistics. 31, 47, 53
- 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. 64
- Kim, J.-D. and Wang, Y. (2012). Pubannotation - a persistent and sharable corpus and annotation repository. In *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pages 202–205, Stroudsburg, PA, USA. Association for Computational Linguistics. 64
- Kingsbury, P. and Palmer, M. (2002). From treebank to propbank. In *LREC*, pages 1989–1993. Citeseer. 15

- Klimek, B. (2017). Proposing an ontalex-mmooon alignment: Towards an interconnection of two linguistic domain models. In *Proceedings of the LDK workshops: OntoLex, TIAD and Challenges for Wordnets*, pages 68–73. CEUR-WS.org. 31, 41, 52, 59
- Klimek, B., Arndt, N., Krause, S., and Arndt, T. (2016). Creating linked data morphological language resources with mmooon - the hebrew morpheme inventory. In Chair), N. C. C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA). 28, 41
- Klimek, B. and Brümmer, M. (2015). Enhancing lexicography with semantic language databases. *Kernerman Dictionary News*, (23):5–10. 55, 57, 73, 198, 221, 223, 230, 232
- Klimek, B. and Brümmer, M. (2015). Enhancing lexicography with semantic language databases. *Kernerman DICTIONARY News*, 23:5–10. 248
- Knap, T. and Michelfeit, J. (2012). Linked data aggregation algorithm: Increasing completeness and consistency of data. *Provided by Charles University*. 273, 274
- Lenci, A., Bel, N., Busa, F., Calzolari, N., Gola, E., Monachini, M., Ogonowski, A., Peters, I., 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. 15, 46
- Liu, H. and Singh, P. (2004). Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226. 33
- 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. 51
- Marc Kemps-Snijders, Menzo Windhouwer, P. W. and Wright, S. E. (2008). Isocat: Corraling data categories in the wild. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, B. M. J. M. J. O. S. P. D. T., editor, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)*, Paris,

- France. European Language Resources Association (ELRA). <http://www.lrec-conf.org/proceedings/lrec2008/>. 31, 59, 75
- McCracken, J. (2015). The exploitation of dictionary data and metadata. In *The Oxford Handbook of Lexicography*. 266
- McCrae, J., 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*, 46:701–719. 7, 21, 30, 34, 41, 267
- McCrae, J., Aguado-de Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Others, Wunner, T., and ... & Wunner, T. (2010). The lemon cookbook. Technical report, Monnet Project. 34, 35, 42, 52
- McCrae, J., Chiarcos, C., Bond, F., Cimiano, P., and Declerck, T. (2016). The Open Linguistics Working Group: Developing the Linguistic Linked Open Data Cloud. In *Proceedings of 10th Language Resources and Evaluation Conference (LREC 2016)*, pages 2435–2441., Paris, France. European Language Resources Association (ELRA). 9, 27, 30
- McCrae, J., 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 (LDL): Multilingual Knowledge Resources and Natural Language Processing at the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Paris, France. European Language Resources Association (ELRA). 28, 48, 53
- McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., and Cimiano, P. (2017). The ontolx-lemon model: development and applications. In *Proceedings of eLex 2017 conference*, pages 19–21. 21
- McCrae, J. P. and Cimiano, P. (2015). Linghub: a linked data based portal supporting the discovery of language resources. In *SEMANTiCS (Posters & Demos)*, volume 1481, pages 88–91. CEUR-WS.org. 70
- McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., et al. (2013). Universal dependency

- annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 92–97. 15
- Mel’cuk, I. and Polguère, A. (1987). A formal lexicon in the meaning-text theory (or how to do lexica with words). *Computational Linguistics*, 13(3-4):261–275. 285
- Mel’čuk, I. M. (1997). *Vers une linguistique Sens-Texte: leçon inaugurale faite le vendredi 10 janvier 1997*. Collège de France, Paris, France. 47
- Mel’cuk, I. (1996). Lexical functions: a tool for the description of lexical relations in a lexicon. *Lexical functions in lexicography and natural language processing*, 31:37–102. 267
- Mel’čuk, I. (2006). Explanatory combinatorial dictionary. *Open problems in linguistics and lexicography*, pages 225–355. 15
- Mel’čuk, I. (1998). Collocations and lexical functions. In Cowie, A., editor, *Phraseology. Theory, Analysis, and Applications*, pages 23–53. Clarendon Press, Oxford, UK. 47
- Menke, P. (2016). *The Fiesta Data Model: A novel approach to the representation of heterogeneous multimodal interaction data*. BoD – Books on Demand, Norderstedt, Germany. 64, 66
- Menke, P., McCrae, J. P., 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*, pages 44 –52, Stroudsburg, PA, USA. Association for Computational Linguistics. 64, 66, 70
- Merriam-Webster Dictionary On-line (2017). *Merriam-Webster, Inc.* Accessed 13.07.2017. 199
- Miles, A. and Bechhofer, S. (2009). SKOS simple knowledge organization system reference. 35, 41

- Millard, I., Glaser, H., Salvadores, M., and Shadbolt, N. (2010). Consuming multiple linked data sources: Challenges and experiences. In *Proceedings of the First International Workshop on Consuming Linked Data (COLD2010)*, volume 665, pages 1–12. CEUR-WS.org. 70, 285
- Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41. 15, 265
- Montiel-Ponsoda, E., Aguado-de cea, G., Gómez-Pérez, A., and Peters, W. (2011). Enriching ontologies with multilingual information. *Natural Language Engineering*, 17:283–309. 35, 150
- Moran, S. (2012). Using Linked Data to create a typological knowledge base. In *Linked Data in Linguistics*, pages 129–138. Springer, Berlin, Heidelberg, Germany. 31, 41, 52
- Moran, S. and Wright, R. (2009). Phonetics information base and lexicon (PHOIBLE). 31, 41, 52
- Moussallem, D., Sherif, M. A., Esteves, D., Zampieri, M., and Ngomo, A.-C. N. (2018). Idioms: A multilingual linked idioms data set. *arXiv preprint arXiv:1802.08148*. 171
- Měchura, M. (2016). Data structures in lexicography: from trees to graphs. In *Proc. of 10th Workshop on Recent Advances in Slavonic Natural Language Processing (RASLAN 2016)*. 247
- Navigli, R. and Ponzetto, S. P. (2010). Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225. Association for Computational Linguistics. 118
- Neubert, J. (2009). Bringing the "thesaurus for economics" on to the web of linked data. *LDOW*, 25964. 102
- Niu, X., Sun, X., Wang, H., Rong, S., Qi, G., and Yu, Y. (2011). Zhishi. me-weaving chinese linking open data. In *International Semantic Web Conference*, pages 205–220. Springer. 83, 119

- Nordhoff, S. (2012). Linked data for linguistic diversity research: Glottolog/langdoc and asjp online. In *Linked Data in Linguistics*, pages 191–200. Springer, Berlin, Heidelberg, Germany. 28, 63
- Nordhoff, S. and Hammarström, H. (2011). Glottolog/Langdoc: Defining dialects, languages, and language families as collections of resources. In *Proceedings of the First International Workshop on Linked Science*, pages 53–58. CEUR-WS.org. 63, 75
- Nuzzolese, A. G., Gangemi, A., and Presutti, V. (2011). Gathering lexical linked data and knowledge patterns from framenet. In *Proceedings of the Sixth International Conference on Knowledge Capture, K-CAP '11*, pages 41–48, New York, NY, USA. ACM. 44, 52
- 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*. 50
- Oltramari, A. and Vetere, G. (2008). Lexicon and ontology interplay in senso comune. In *Proceedings of OntoLex 2008*, pages 24–30, Paris, France. European Language Resources Association (ELRA). 36
- Ouyang, L., Zou, B., Qu, M., and Zhang, C. (2011). A method of ontology evaluation based on coverage, cohesion and coupling. In *2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, volume 4, pages 2451–2455. IEEE. 241
- Oxford Dictionaries (2017). *Oxford Living Dictionaries*. Oxford University Press. Accessed 13.07.2017. 199
- Parvizi, A., Kohl, M., González, M., and Saurí, R. (2016). Towards a linguistic ontology with an emphasis on reasoning and knowledge reuse. In Chair), N. C. C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 441–448, Paris, France. European Language Resources Association (ELRA). 30, 36, 38, 41, 57, 59

- 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*, volume 7. Harvard University Press, Cambridge, MA, USA. 50
- Periñán-Pascual, C. and Arcas-Túnez, F. (2004). Meaning postulates in a lexico-conceptual knowledge base. In *Database and Expert Systems Applications, 2004. Proceedings. 15th International Workshop on*, pages 38–42. IEEE. 15
- Piotrowski, T. (1994). *Problems in bilingual lexicography*. Wydawn. Uniwersytetu Wrocławskiego. 93, 94
- Polguère, A. (2006). Structural properties of lexical systems: Monolingual and multilingual perspectives. In *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*, pages 50–59. Association for Computational Linguistics. 33
- Polguère, A. (2009). Lexical systems: graph models of natural language lexicons. *Language resources and evaluation*, 43(1):41–55. 266
- Polguère, A. (2014). From Writing Dictionaries to Weaving Lexical Networks. *International Journal of Lexicography*. 33
- Polguère, A. (2014). From writing dictionaries to weaving lexical networks. *International Journal of Lexicography*, 27(4):396–418. 266, 267
- PONS (2017). *PONS Online- Wörterbuch Deutsch als Fremdsprache*. Accessed 13.07.2017. 199
- Pradhan, S. S., Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2007). Ontonotes: A unified relational semantic representation. *International Journal of Semantic Computing*, 1(4):405–419. 51
- Przepiórkowski, A., Hajnicz, E., Patejuk, A., and Woliński, M. (2014). Extended phraseological information in a valence dictionary for nlp applications. In *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing*, pages 83–91. 12

- Pustejovsky, J. (1991). The generative lexicon. *Computational Linguistics*, 17(4):409–441. 31, 46
- Pustejovsky, J. and Bouillon, P. (1995). Aspectual coercion and logical polysemy. *Journal of semantics*, 12(2):133–162. 47
- Ramos, J. and Gómez-Pérez, A. (2008). Mappings for the semantic web. *IEEE Intelligent Systems-Trends & Controversies*, pages 79–81. 124
- Rayfield, D. (2006). *A comprehensive Georgian-English dictionary*. Garnett Press. 17, 25, 83, 95, 98
- Real Academia Española - DLE (2014). *Diccionario de la Lengua Española*, 23^a Edición. Accessed 13.07.2017. 199
- Romary, L. (2013a). Data models and the (blind?) query of lexical resources. In *Perspectives on querying TEI-annotated data-TEI Conference and Members Meeting 2013*. 12
- Romary, L. (2013b). Tei and lmf crosswalks. *arXiv preprint arXiv:1301.2444*. 21
- Rubiera, E., Polo, L., Berrueta, D., and El Ghali, A. (2012). Telix: An rdf-based model for linguistic annotation. In Simperl, E., Cimiano, P., Polleres, A., Corcho, O., and Presutti, V., editors, *The Semantic Web: Research and Applications*, pages 195–209, Berlin, Heidelberg, Germany. Springer. 30, 37, 38, 41, 64
- Ruimy, N., Corazzari, O., Gola, E., Spanu, A., Calzolari, N., and Zampolli, A. In *First International Conference on Language Resources and Evaluation*. 46
- Ruimy, N., Monachini, M., Distant, R., Guazzini, E., Molino, S., Olivieri, M., Calzolari, N., and Zampolli, A. (2002). CLIPS, a Multi-level Italian Computational Lexicon: a Glimpse to Data. In *International Conference on Language Resources and Evaluation (LREC 2002)*, Paris, France. European Language Resources Association. 47
- Rüther, M., Fock, J., Hübener, J., Bandholtz, T., and Schulte-Coerne, T. (2010). Linked environment data. In *EnviroInfo*, pages 470–479. 102

- Sánchez Rada, J. F. and Iglesias Fernandez, C. A. (2013). Onyx: Describing emotions on the web of data. In *Proceedings of the First International Workshop on Emotion and Sentiment in Social and Expressive Media*, pages 71–82. CEUR-WS.org. 65, 66
- 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)*. FEUP Ed. 58, 74
- Schuler, K. K. (2005). Verbnet: A broad-coverage, comprehensive verb lexicon. 15
- 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*, number 123, pages 62–70, Linköping, Sweden. Linköping University Electronic Press. 60, 75
- Sérasset, G. (2015). Dbnary: Wiktionary as a lemon-based multilingual lexical resource in rdf. *Semantic Web*, 6(4):355–361. 56, 60, 74, 109, 119
- Sherif, M. A. and Ngonga Ngomo, A.-C. (2015). Semantic Quran. *Semantic Web*, 6(4):339–345. 64, 66
- Sini, M., Soergel, D., and Johannsen, G. (2009). Proposal for restructuring Scientific Names and Common Names of Organisms in AGROVOC. 105, 150
- Skut, W., Uszkoreit, H., Brants, T., and Krenn, B. (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, August 17-28, Saarbrücken, Germany*. o.A. 42
- SM Dictionarios - CLAVE (2017). *Diccionario Clave (online)*. Accessed 13.07.2017. 199
- Soergel, D., Lauser, B., Liang, A., Fisseha, F., Keizer, J., and Katz, S. (2006). Reengineering thesauri for new applications: the AGROVOC example. *Journal of Digital Information*, 4(4). 105, 150

- Sperberg-McQueen, C. M. and Burnard, L. (1994). Guidelines for electronic text encoding and interchange. Technical report. 20, 34
- Spohr, D. (2012). *Towards a multifunctional lexical resource: Design and implementation of a graph-based lexicon model*, volume 141. Walter de Gruyter, Berlin, Germany; Boston, USA. 34, 41
- Stolk, S. (2017). Ontolex and onomasiological ordering: Supporting topical thesauri. In *LDK Workshops*, pages 60–67. 55
- Stolk, S. (2019). lemon-tree: Representing topical thesauri on the semantic web. In *2nd Conference on Language, Data and Knowledge (LDK 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik. 55, 74, 204
- Studer, R., Benjamins, V., and Fensel, D. (1998). Knowledge engineering: Principles and methods. *Data Knowledge Engineering*, 25(1):161–197. 4
- Su, X. (2002). A text categorization perspective for ontology mapping. *Norway: Department of Computer and Information Science, Norwegian University of Science and Technology*. 124
- Swanepoel, P. (1994). Problems, theories and methodologies in current lexicographic semantic research. In *Sixth International Euralex Congress held at the Free University of Amsterdam*, volume 30. 3, 12
- 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)*, pages 5–12, Paris, France. European Language Resources Association (ELRA). 56, 109
- Tolkien, J. R. R. (1997). *The Hobbit, or There and Back Again. 1937*. London: Harper-Collins. iv
- Toral, A. and Monachini, M. (2007). Simpleowl: a generative lexicon ontology for nlp and the semantic web. In *Workshop on Cooperative Construction of Linguistic Knowledge Bases (AIIA 2007)*. 46, 53

- Trippel, T. (2006). *The Lexicon Graph Model: A generic model for multimodal lexicon development*. AQ-Verlag, Saarbrücken, Germany. 15, 16, 17, 18, 33, 285
- Turdean, T. and Joshi, S. (2017). Triplifying a dictionary: Some learnings. *DICTIONARY News*, page 4. 191
- Vila-Suero, D., Gómez-Pérez, A., Montiel-Ponsoda, E., Gracia, J., and Aguado-de Cea, G. (2014a). *Publishing Linked Data: the multilingual dimension*, pages 101–118. Springer, Berlin, Heidelberg, Germany. 4, 9, 87, 160, 164
- Vila-Suero, D., Gómez-Pérez, A., Montiel-Ponsoda, E., Gracia, J., and Aguado-de Cea, G. (2014b). Publishing linked data: the multilingual dimension. *Towards the Multilingual Semantic Web*, pages 101–118. 273
- Villazón-Terrazas, B. and Corcho, O. (2011). Methodological guidelines for publishing linked data. *Una Profesión, un futuro: actas de las XII Jornadas Españolas de Documentación: Málaga*, 25(26):20. 273
- Villazón-Terrazas, B., Vila-Suero, D., Garijo, D., Vilches-Blazquez, L. M., Poveda-Villalon, M., Mora, J., Corcho, O., and Gomez-Perez, A. (2012). Publishing linked data-there is no one-size-fits-all formula. 9, 160
- Villegas, M. and Bel, N. (2015). PAROLE/SIMPLE ‘lemon’ontology and lexicons. *Semantic Web*, 6(4):363–369. 12, 47, 53
- Villegas, M., Melero, M., Bel, N., and Gracia, J. (2016). Leveraging rdf graphs for crossing multiple bilingual dictionaries. In Chair), N. C. C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 868–876, Paris, France. European Language Resources Association (ELRA). 28, 37
- Visser, P. R., Jones, D. M., Bench-Capon, T. J., and Shave, M. (1997). An analysis of ontology mismatches; heterogeneity versus interoperability. In *AAAI 1997 Spring Symposium on Ontological Engineering, Stanford CA., USA*, pages 164–72. 124

- 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. 49
- Westerski, A. and Sánchez-Rada, J. F. (2013). Marl Ontology Specification, V1. 0 May 2013. 65, 66
- Westphal, P., Stadler, C., and Pool, J. (2015). Countering language attrition with PanLex and the Web of Data. *Semantic Web*, 6(4):347–353. 56, 74
- Windhouwer, M. (2012). Relcat: a relation registry for isocat data categories. In Chair), N. C. C., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12)*, pages 3661–3664, Istanbul, Turkey. European Language Resources Association (ELRA). 59
- Windhouwer, M. and Wright, S. (2012). Linking to linguistic data categories in ISOcat. In *Linked Data in Linguistics*, pages 99–107. Springer, Berlin, Heidelberg, Germany. 31, 59
- Wolf, F. and Gibson, E. (2005). Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31(2):249–287. 51
- Yong, H. and Peng, J. (2007). *Bilingual lexicography from a communicative perspective*, volume 9. John Benjamins Publishing. 93
- Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., and Auer, S. (2016). Quality assessment for linked data: A survey. *Semantic Web*, 7(1):63–93. 272