

Paul Buitelaar · Philipp Cimiano *Editors*

# Towards the Multilingual Semantic Web

Principles, Methods and Applications

 Springer

# Towards the Multilingual Semantic Web



Paul Buitelaar • Philipp Cimiano  
Editors

# Towards the Multilingual Semantic Web

Principles, Methods and Applications

 Springer

*Editors*

Paul Buitelaar  
National University of Ireland  
Galway  
Ireland

Philipp Cimiano  
Universität Bielefeld  
Bielefeld  
Germany

ISBN 978-3-662-43584-7      ISBN 978-3-662-43585-4 (eBook)  
DOI 10.1007/978-3-662-43585-4  
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014952219

© Springer-Verlag Berlin Heidelberg 2014

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media ([www.springer.com](http://www.springer.com))

# Preface

The amount of Internet users speaking native languages other than English has seen a substantial growth in recent years. Recent statistics in fact show that the number of non-English Internet users is almost three times the number of English-speaking users. As a consequence, the Web is turning more and more into a truly multilingual platform in which speakers and organizations from different languages and cultural backgrounds collaborate, consuming and producing information at a scale without precedent. Originally conceived by Berners-Lee et al. (2001) as “an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation,” the Semantic Web has seen an impressive growth in recent years in terms of the amount of data published on the Web using the REsource Description Framework (RDF)<sup>1</sup> and OWL<sup>2</sup> data models. The kind of data published on the Semantic Web or Linked Open Data (LOD) cloud is mainly of a factual nature and thus represents a basic body of knowledge that is accessible to mankind as a basis for informed decision-making. The creation of a level playing field in which citizens from all countries have access to the same information and have equal opportunities to contribute to that information is a crucial goal to achieve. Such a level playing field will also reduce information hegemonies and biases, increasing diversity of opinion. However, the semantic vocabularies used to publish factual data in the Semantic Web are mainly in English, which creates a strong bias towards this language and English-based cultures.

As in the traditional Web, language represents an important barrier for information access in the Semantic Web as it is not straightforward to access information produced in a foreign language. A big challenge for the Semantic Web therefore is to develop architectures, frameworks, and systems that can help in overcoming

---

<sup>1</sup><http://www.w3.org/RDF/>.

<sup>2</sup><http://www.w3.org/OWL>.

language and national barriers, facilitating the access to information produced within different cultures and languages. An additional problem is that most of the information on the Web stems from a small set of countries where major languages are spoken. This leads to a situation in which the public discourse is mainly driven and shaped by contributions from those countries where these major languages are spoken. The Semantic Web vision bears an excellent potential to create a level playing field for users with different cultural backgrounds and native languages and originating from different geopolitical environments, the reason being that the information available on the Semantic Web is expressed in a language-independent fashion and thus bears the potential to be accessible to speakers of different languages if the right mediation mechanisms are in place. However, so far the relation between multilingualism and the Semantic Web has not received enough attention in the research community. The goal of this book is therefore to document the state of the art with respect to the above vision of a *Multilingual Semantic Web*, in which semantic information is accessible in multiple and across languages.

The *Multilingual Semantic Web*, as envisioned in this book, would allow for the following functionality:

- Answering information needs in any language with respect to semantically structured data available on the Semantic Web and Linked Open Data cloud
- Verbalizing and accessing semantically structured data, ontologies, or other conceptualizations in different languages
- Harmonization, integration, aggregation, comparison, and repurposing of semantically structured data across languages
- Aligning and reconciling ontologies or other conceptualizations across languages

This book has to some extent been the result of a Dagstuhl Seminar on the “*Multilingual Semantic Web*,” co-organized by Buitelaar et al. in September 2012. Several of the authors of the book chapters were present at this seminar.

The book is divided into three main parts: *Principles*, *Methods*, and *Applications*. The part on Principles discusses formalisms for building the Multilingual Semantic Web. The part on Methods describes algorithms for the construction of the Multilingual Semantic Web. The part on Applications describes the use of the Multilingual Semantic Web in the context of several real-life systems.

## Principles

The chapter by Hirst analyzes the original vision of a Semantic Web by Berners-Lee et al. (2001) and discusses what this vision implies for a Multilingual Semantic Web and the barriers that the nature of language imposes on it. The chapter essentially argues for the impossibility to represent knowledge interlingually by a symbolic language and argues for the exploitation of distributional semantics to represent multilingual content. In particular, the chapter contrasts a writer-oriented and a reader-oriented perspective of the Semantic Web, arguing that so far the

Semantic Web has focused on a writer-perspective and neglected issues related to the perspective of a reader who consumes information on the Semantic Web.

McCrae and Unger describe work at the ontology–lexicon interface and address the issue of how conceptual schemas and RDF datasets can be enriched with linguistic information to express how the elements of the data model can be expressed in different languages. In their work, they build on the *lemon* model and present a domain-specific representation language that builds on patterns to facilitate the creation of *lemon* lexica. This work will thus facilitate the enrichment of the Semantic Web with a lexical layer. They present the creation of a lexicon for DBpedia in English as a use case.

León Araúz and Faber discuss principled issues related to ontology localization. They argue that a lexical layer for the Semantic Web needs to have a suitable formalism for representing and handling cross-lingual variation including syntactic, lexical, conceptual, and semantic features but most importantly also contextual features that model which translation is appropriate in which context. Further, they also present a taxonomy of different types of cross-language equivalence relations.

Pretorius discusses in her chapter the opportunities that the vision of a Multilingual Semantic Web creates for under-resourced languages, in particular for the preservation of indigenous knowledge and thus cultural diversity. In her chapter, she takes a closer look at the challenges that under-resourced languages, in particular South African languages, face. She presents three use cases in which different types of linguistic resources, ranging from multilingual terminologies, indigenous knowledge on astronomy to a parallel corpus based on the South African constitution, are defined and made available as Linked Data.

van Grondelle and Unger present a paradigm for developing scoped human language technology (HLT) applications in the sense that these applications are aligned with a particular application domain and language. They propose a modular architecture for developing HLT applications by decomposing grammars into different modules that can be flexibly composed together in developing a specific application. With this approach, the development of HLT applications is facilitated by a plug-and-play philosophy, and the reuse of components and modules across applications is maximized. A proof-of-concept implementation of this architecture is presented.

Demey and Heath discuss issues related to the verbalization of  $n$ -ary relations given that popular Semantic Web formalisms natively support only binary relations. They propose an approach based on reification, which transforms  $n$ -ary relations into a set of binary relations. The authors discuss the case of English and Chinese and present a number of typical and representative verbalization patterns for  $n$ -ary relations.



## Methods

Vila-Suero et al. are concerned with the challenges in publishing Multilingual Linked Data. They present a methodology for the publication of Multilingual Linked Data that consists of the following steps: (1) specification, (2) modeling, (3) generation, (4) interlinking, and (5) publication. For each of these steps, they discuss aspects, issues, and design decisions, taking into account the multilingual nature of the data.

Alignment of ontologies or conceptualizations originating from different languages is a crucial research topic in the field of the Multilingual Semantic Web. Trojahn et al. discuss the state of the art in cross-lingual ontology matching. On the one hand, they formally define the problem, distinguishing the case of monolingual, multilingual, and cross-lingual ontology matching. On the other, they provide an overview of existing solutions and evaluation datasets and discuss the results of different tools on standard benchmarking datasets.

In a similar vein, Cabrio et al. analyze the synchronization level between language versions of DBpedia. They compare the coverage of the different DBpedia versions with respect to each other, concluding that the versions clearly vary in their completeness, granularity, and coverage, but complement each other. Further, they present an automatic approach to align the properties of different DBpedia language versions and show how these mappings can be exploited in the context of a cross-language question answering system, QAKIS.

Embley et al. present the ML-OntoES system, a semantic search system that supports searching information across languages by mapping them into a language-independent ontology that is shared across languages and into which content in different languages is mapped. A prototype implementation of this paradigm is discussed and shown to deliver satisfactory results.

A crucial aspect of the Multilingual Semantic Web is to enable different stakeholders to engage together and synchronize in the task of developing a joint conceptualization of some domain of common interest. Bosca et al. present an approach along these lines, based on the Moki toolkit, that allows experts to collaborate in creating and translating ontologies across languages. The features that support collaborative ontology management are discussed, focusing on challenging issues and their solution.

An important task within the Multilingual Semantic Web is to move from data models to linguistic representations (generation) and back (interpretation). Gerber and Ngonga Ngomo present a principled approach that is based on BOotstrapping linked data (BOA), a framework that supports the extraction of RDF data from text by inducing a set of lexical patterns. BOA can be used to extract RDF triples from text but also to generate linguistic descriptions from existing triples. A nice feature of BOA is that it follows a language-independent approach and thus can be adapted to different languages straightforwardly. The authors demonstrate the applicability of their approach across languages by training on four different corpora in two different languages (German and English). Further, they show how BOA can

be applied in different applications, e.g., in the task of extracting facts with high accuracy from textual data as well as in the task of validating RDF facts using textual data and in the context of the question answering system Template - based SPARQL Learner (TBSL).

Along similar lines, Damova et al. present an approach that allows one to query semantic knowledge bases in natural language and obtain results from the knowledge base as coherent texts. The solution builds on the Grammatical Framework and implements several transition steps to move from natural language to SPARQL and from a set of RDF triples to coherent natural language text in multiple languages.

Gromann and Declerck address the issue that labels in ontologies are often impoverished by sacrificing linguistic expressivity and completeness for compactness. However, in this way, domain semantics is lost, e.g., through ellipsis. They present a method to expand condensed labels by inferring implicit content from occurrences of ellipsis, which relies on cross-language comparison of labels.

Bond et al. present an approach to develop multilingual lexica linked to a formal ontology. The method is instantiated for WordNet, Global WordNet, and SUMO to create a rich Web of linguistic data linked to axiomatized knowledge.

Tanev and Zavarella present a semiautomatic, weakly supervised approach for lexical acquisition that is language independent and relies on the principle of distributional semantics. It learns semantic classes, modifiers, and event patterns from an unannotated text corpus. The authors discuss the application of this method to reports of natural disasters in Spanish and English.

## Applications

Cross-language and cross-border integration of knowledge is an important topic of research within the Multilingual Semantic Web. An important use case for this is the integration of financial information across countries and legal jurisdictions, in particular business reports that are typically created relative to financial taxonomies used in each country. The eXtensible Business Reporting Language (XBRL) has standardized the generation of and the access to financial statements like balance sheets, but language and XBRL-taxonomy diversity makes financial data integration across national borders and jurisdictions problematic. Integrating financial data in these circumstances requires that different multilingual jurisdictional taxonomies be aligned by finding correspondences between concepts. Thomas et al. present a method to align XBRL taxonomies originating from different countries. The method relies on semantic tagging of accounting concepts, thus narrowing down the possible mappings to a subset of all possible one-to-one mappings.

Thurmair presents an approach to acquire relevant domain knowledge and multilingual terminologies to support ontology-based search across languages. The chapter describes an effort in enhancing an existing system by a natural language query interface in which users can type in a free text query rather than navigate the

ontology to find relevant texts. The acquired multilingual terminologies are used to map a free-text query to the relevant ontology concepts, thus supporting multilingual search. A proof of concept of the ontology-based search approach is provided for the domain of assistive technology.

Murakami et al. present a service-oriented architecture that fosters the easy development of multilingual NLP services and enhances interoperability of language services and facilitates their composition. The chapter describes the architecture of the Language Grid and describes how the service domain model can be used to define service interfaces and service profiles.

## Acknowledgments

This book could not have been written without the support of the EU FP7 program in the context of the projects Monnet (Grant no.: 248458), LIDER (610782), EuroSentiment (296277), and Portdial (296170); the Science Foundation Ireland for the projects Lion2 (SFI/08/CE/I1380) and Insight (SFI/12/RC/2289); and the Deutsche Forschungsgemeinschaft (DFG) via the Excellence Center Cognitive Interaction Technology (CITEC).

We hope that you enjoy the book!

Galway, Ireland  
Bielefeld, Germany  
Spring 2014

Paul Buitelaar  
Philipp Cimiano

## References

- Berners-Lee, T., Hendler, J., & Lassila, O. (2001, May). The semantic web. *Scientific American*, 284(5), 34–43.
- Buitelaar, P., Choi, K. -S., Cimiano, P., & Hovy, E. H. (2012). The multilingual semantic web (Dagstuhl Seminar 12362). *Dagstuhl Reports*, 2(9), 15–94.

# Contents

## Part I Principles

<b>Overcoming Linguistic Barriers to the Multilingual Semantic Web</b> .....	3
Graeme Hirst	
<b>Design Patterns for Engineering the Ontology-Lexicon Interface</b> .....	15
John P. McCrae and Christina Unger	
<b>Context and Terminology in the Multilingual Semantic Web</b> .....	31
Pilar León-Araúz and Pamela Faber	
<b>The Multilingual Semantic Web as Virtual Knowledge Commons: The Case of the Under-Resourced South African Languages</b> .....	49
Laurette Pretorius	
<b>A Three-Dimensional Paradigm for Conceptually Scoped Language Technology</b> .....	67
Jeroen van Grondelle and Christina Unger	
<b>Towards Verbalizing Multilingual N-Ary Relations</b> .....	83
Yan Tang Demey and Clifford Heath	

## Part II Methods

<b>Publishing Linked Data on the Web: The Multilingual Dimension</b> .....	101
Daniel Vila-Suero, Asunción Gómez-Pérez, Elena Montiel-Ponsoda, Jorge Gracia, and Guadalupe Aguado-de-Cea	
<b>State-of-the-Art in Multilingual and Cross-Lingual Ontology Matching</b> .....	119
Cássia Trojahn, Bo Fu, Ondřej Zamazal, and Dominique Ritzke	

<b>Mind the Cultural Gap: Bridging Language-Specific DBpedia Chapters for Question Answering</b> .....	137
Elena Cabrio, Julien Cojan, and Fabien Gandon	
<b>Multilingual Extraction Ontologies</b> .....	155
David W. Embley, Stephen W. Liddle, Deryle W. Lonsdale, Byung-Joo Shin, and Yuri Tijerino	
<b>Collaborative Management of Multilingual Ontologies</b> .....	175
Alessio Bosca, Mauro Dragoni, Chiara Di Francescomarino, and Chiara Ghidini	
<b>From RDF to Natural Language and Back</b> .....	193
Daniel Gerber and Axel-Cyrille Ngonga Ngomo	
<b>Multilingual Natural Language Interaction with Semantic Web Knowledge Bases and Linked Open Data</b> .....	211
Mariana Damova, Dana Dannélls, Ramona Enache, Maria Mateva, and Aarne Ranta	
<b>A Cross-Lingual Correcting and Completive Method for Multilingual Ontology Labels</b> .....	227
Dagmar Gromann and Thierry Declerck	
<b>A Multilingual Lexico-Semantic Database and Ontology</b> .....	243
Francis Bond, Christiane Fellbaum, Shu-Kai Hsieh, Chu-Ren Huang, Adam Pease, and Piek Vossen	
<b>Multilingual Lexicalisation and Population of Event Ontologies: A Case Study for Social Media</b> .....	259
Hristo Tanev and Vanni Zavarella	
<b>Part III Applications</b>	
<b>Semantically Assisted XBRL-Taxonomy Alignment Across Languages</b> .....	277
Susan Marie Thomas, Xichuan Wu, Yue Ma, and Sean O’Riain	
<b>Lexicalizing a Multilingual Ontology for Searching in the Assistive Technology Domain</b> .....	295
Gregor Thurmair	
<b>Service-Oriented Architecture for Interoperability of Multilanguage Services</b> .....	313
Yohei Murakami, Donghui Lin, and Toru Ishida	
<b>Index</b> .....	329

# Contributors

- Guadalupe Aguado-de-Cea** Universidad Politécnica de Madrid, Madrid, Spain
- Francis Bond** Nanyang Technological University, Singapore, Singapore
- Alessio Bosca** Celi s.r.l., Torino, Italy
- Elena Cabrio** INRIA Sophia-Antipolis Méditerranée, Sophia Antipolis, France  
and EURECOM, Sophia Antipolis, France
- Julien Cojan** INRIA Sophia-Antipolis Méditerranée, Sophia Antipolis, France
- Mariana Damova** Ontotext AD, Sofia, Bulgaria
- Dana Dannélls** University of Gothenburg, Gothenburg, Sweden
- Thierry Declerck** DFKI GmbH, Saarbruecken, Germany  
ICLTT, Vienna, Austria
- Yan Tang Demey** Vrije Universiteit Brussel, Brussels, Belgium
- Chiara Di Francescomarino** FBK–IRST, Trento, Italy
- Mauro Dragoni** FBK–IRST, Trento, Italy
- David W. Embley** Brigham Young University, Provo, UT, USA
- Ramona Enache** University of Gothenburg, Gothenburg, Sweden
- Pamela Faber** University of Granada, Granada, Spain
- Christiane Fellbaum** Princeton University, Princeton, NJ, USA
- Bo Fu** University of Victoria, Victoria, BC, Canada
- Fabien Gandon** INRIA Sophia-Antipolis Méditerranée, Sophia Antipolis, France
- Daniel Gerber** Universität Leipzig, Leipzig, Germany

- Chiara Ghidini** FBK–IRST, Trento, Italy
- Asunción Gómez-Pérez** Universidad Politécnica de Madrid, Madrid, Spain
- Jorge Gracia** Universidad Politécnica de Madrid, Madrid, Spain
- Dagmar Gromann** Vienna University of Economics and Business, Vienna, Austria
- Clifford Heath** Data Constellation, Roseville, NSW, Australia
- Graeme Hirst** University of Toronto, Toronto, ON, Canada
- Shu-Kai Hsieh** National Taiwan University, Taipei, Taiwan
- Chu-Ren Huang** Hong Kong Polytechnic University, Hong Kong, China
- Toru Ishida** Kyoto University, Kyoto, Japan
- Pilar León-Araúz** University of Granada, Granada, Spain
- Stephen W. Liddle** Brigham Young University, Provo, UT, USA
- Donghui Lin** Kyoto University, Kyoto, Japan
- Deryle W. Lonsdale** Brigham Young University, Provo, UT, USA
- Yue Ma** Technische Universität Dresden, Dresden, Germany
- Maria Mateva** Ontotext AD, Sofia, Bulgaria
- John P. McCrae** Bielefeld University, Bielefeld, Germany
- Elena Montiel-Ponsoda** Universidad Politécnica de Madrid, Madrid, Spain
- Yohei Murakami** Kyoto University, Kyoto, Japan
- Axel-Cyrille Ngonga Ngomo** Universität Leipzig, Leipzig, Germany
- Sean O’Riain** National University of Ireland, Galway, Ireland
- Adam Pease** Articulate Software, San Francisco, CA, USA
- Laurette Pretorius** University of South Africa, Pretoria, South Africa
- Aarne Ranta** University of Gothenburg, Gothenburg, Sweden
- Dominique Ritze** University of Mannheim, Mannheim, Germany
- Byung-Joo Shin** Kyungnam University, Kyungnam, South Korea
- Hristo Tanev** European Commission, Joint Research Centre, Ispra, Italy
- Susan Marie Thomas** SAP AG, Karlsruhe, Germany
- Gregor Thurmair** Liguattec GmbH, Munich, Germany
- Yuri Tijerino** Kwansai Gakuin University, Kobe-Sanda, Japan

**Cássia Trojahn** University of Toulouse 2 and IRIT, Toulouse, France

**Christina Unger** Bielefeld University, Bielefeld, Germany

**Jeroen van Grondelle** Be Informed, Apeldoorn, The Netherlands

**Daniel Vila-Suero** Universidad Politécnica de Madrid, Madrid, Spain

**Piek Vossen** Vrije Universiteit, Amsterdam, The Netherlands

**Xichuan Wu** SAP AG, Karlsruhe, Germany

**Ondřej Zamazal** University of Economics, Prague, Czech Republic

**Vanni Zavarella** European Commission, Joint Research Centre, Ispra, Italy



# **Part I**

## **Principles**

# Overcoming Linguistic Barriers to the Multilingual Semantic Web

Graeme Hirst

**Abstract** I analyze Berners-Lee, Hendler, and Lassila’s description of the Semantic Web, discussing what it implies for a Multilingual Semantic Web and the barriers that the nature of language itself puts in the way of that vision. Issues raised include the mismatch between natural language lexicons and hierarchical ontologies, the limitations of a purely writer-centered view of meaning, and the benefits of a reader-centered view. I then discuss how we can start to overcome these barriers by taking a different view of the problem and considering distributional models of semantics in place of purely symbolic models.

**Key Words** Distributional semantics • Near-synonymy • Ontologies • Reader-centered view of meaning • Semantic Web • Writer-centered view of meaning

## 1 Introduction

The Semantic Web . . . in which information is given well-defined meaning, better enabling computers and people to work in cooperation. — Berners-Lee et al. (2001, p. 37)<sup>1</sup>

Sometime between the publication of the original paper with this description of the Semantic Web and Berners-Lee’s (2009) “Linked Data” talk at TED, the vision of the Semantic Web contracted considerably. Originally, the vision was about “information”; now it is only about data. The difference is fundamental. Data, even if it is strings of natural language, has an inherent semantic structure and a stipulated interpretation, even if that too is a label in natural language. Other kinds of information, however, are semi-structured or unstructured and may come with no interpretation imposed. In particular, textual information gains an interpretation only in context and only for a specific reader or community of readers (Fish 1980).

---

<sup>1</sup>I will refer to these authors, and metonymously to this paper, as *BLHL*.

G. Hirst (✉)

Department of Computer Science, University of Toronto, Toronto, ON, Canada M5S 3G4

e-mail: [gh@cs.toronto.edu](mailto:gh@cs.toronto.edu)

I do not mean to criticize the idea of restricting Semantic Web efforts to data *pro tem*. It is still an extremely challenging problem, especially in its multilingual form (Gracia et al. 2012, this volume *passim*), and the results will still be of enormous utility. At the same time, however, we need to keep sight of the broader goal that BLHL’s vision implies in order to make sure that our efforts to solve the preliminary problem are not just climbing trees to reach the moon. In this chapter, I will perform a hermeneutical analysis of BLHL’s description, with discussion of what it implies for the Multilingual Semantic Web and the barriers that the nature of language itself puts in the way of that vision. I will then discuss how we can start to overcome these barriers.

I assume in this chapter the standard received notion of the Multilingual Semantic Web as one in which web pages contain (inter alia) natural language text but are also marked up with semantic annotations in a logical representation that enables inferences to be made, that is independent of any particular natural language, and that draws on shared ontologies that are also language-independent. And consequent upon that, the Multilingual Semantic Web, in response to users’ queries and searches, expressed in a natural language or by other means, is able to bring together multiple pages in multiple languages, matching the query to semantic annotations, drawing inferences as necessary, and presenting the results in whatever language the user wants, translating from one language to another as necessary.

## 2 Well-Defined Meaning and Multilinguality

In BLHL’s vision, “information is given well-defined meaning,” implying paradoxically that information did not have well-defined meaning already. Of course, the phrase “well-defined meaning” lacks well-defined meaning, but BLHL are not really suggesting that information on the non-Semantic Web is meaningless; rather what they want is *precision* and the *absence of ambiguity* in the semantic layer. In the case of information expressed linguistically, this implies semantic interpretation into a symbolic knowledge representation language of the kind that they talk about elsewhere in their paper. Developing such representations was a goal that exercised, and ultimately defeated, research in artificial intelligence and natural language understanding from the 1970s through to the mid-1990s (Hirst 2013) (see Sect. 5) and which the Semantic Web has made once more a topic of research (e.g., Cimiano et al. 2014).

One of the barriers that this earlier work ran into was the fact that traditional symbolic knowledge representations proved to be poor representations for linguistic meaning and hierarchical ontologies proved to be poor representations for the lexicon of a language (Hirst 2009a).<sup>2</sup> Models such as LexInfo and *lemon*

---

<sup>2</sup>Wilks (2009), echoed by Borin (2012), suggests that, *a fortiori*, “ontologies” as presently constructed are nothing more than substandard lexicons disguised as something different.

(Cimiano et al. 2011; McCrae et al. 2012) attempt to associate multilingual lexical and syntactic information with ontologies, but they necessarily retain the idea that “the sense inventory is provided by a given domain ontology” (Cimiano et al. 2011, fn 9), under the assumption that the domain of a text, and hence the requisite unique ontology, is known *a priori* or can be confidently identified prior to semantic analysis. In practice, however, this leads to an inflexible and limiting view of word senses. For example, languages tend to have many groups of near-synonyms that form clusters of related and overlapping meanings that do not admit a hierarchical differentiation (Edmonds and Hirst 2002). And quite apart from lexical issues, any system for fully representing linguistically expressed information must itself have the expressive power of natural language, which is far greater than the first-order and near-first-order representations that are presently used; but the higher-order and intensional representations required for this degree of expressiveness (Montague 1974) are computationally infeasible (Friedman et al. 1978).

All these problems are compounded when we add multilinguality as an element. For example, different languages will often present a different and mutually incompatible set of word senses, as each language lexicalizes somewhat different categorizations or perspectives of the world and each language has *lexical gaps* relative both to other languages and to the categories of a complete ontology (Hirst 2009a, pp. 278–279). The consequence of these incompatibilities for the Multilingual Semantic Web is that the utility of ontologies for interpreting linguistic information is thereby limited, and so, conversely, is the ability of lexicons to express ontological concepts. This leads to practical limitations on models of lexicons for ontologies, such as McCrae et al.’s (2012) *lemon* model, that put an emphasis on *interchangeability*—the idea that one ontology can have many different lexicons, for example, for different languages or dialects. This wrongly assumes that *translation-equivalent words* have identical meanings. In fact, it is rare even for words that are regarded as translation equivalents to be completely identical in sense, and such cases are limited mostly to cross-language borrowings and monosemous technical terms in highly structured domains (Adamska-Sałaciak 2013). For example, the sport of soccer, which Cimiano et al. (2014) use as a domain to exemplify an ontology with interchangeable lexicons, is sufficiently technical and well-structured for the approach to succeed; so are the deliberately very narrow domains considered by Embley et al. (this volume). But interchangeability might fail even in ontologies for well-structured domains (cf. Léon-Araúz and Faber, this volume). For example, regarding the domain of university administrative structures, Schogt (1988, p. 97) writes: “When I want to talk about aspects of the intricate administrative system of the University of Toronto to Dutch academics it is very difficult to use Dutch because there are no Dutch terms that correspond to those used in Toronto, the Dutch set-up not sharing the functions and divisions that characterize the Toronto system.”

More usually, translation-equivalent words are merely cross-lingual near-synonyms (Hirst 2009a, p. 279). For example, in the concept space of differently sized areas of trees, the division between the French *bois* and *forêt* occurs at a “larger” point than the division between the German *Holz* and *Wald*

(Hjelmslev 1961; Schogt 1976, 1988). Similarly, English, German, French, and Japanese all have a large vocabulary for different kinds of mistakes and errors, but they each divide up the concept space quite differently. For example, the Japanese words that translate the English words *mistake* or *error* include *machigai*, *ayamari*, and *ayamachi*; Fujiwara et al. (1985) note:

*Machigai* implies a straying from a proper course or the target, and suggests that the results are not right. *Ayamari* describes wrong results objectively. Focus of attention is given solely to the results; concerns, worries, or inadvertence in the course of action are not taken into consideration as in *machigai*. *Ayamachi* implies serious wrongdoing or crime. Also, it is used for accidental faults. *Ayamachi* is concerned with whether the results are good or bad, based on moral judgement, while *ayamari* is concerned with whether the results are right or wrong.<sup>3</sup>

To translate the same two English words *mistake* and *error* to German, Farrell (1977, p. 220) notes that even though *error* “expresses a more severe criticism than *mistake*”, both are covered by *Fehler*, except that *Irrtum* should be used if the mistake is a misunderstanding or other mental error and *Mangel* if the mistake is a “deficiency [or] absence” rather than a “positive fault or flaw” or if it is a visible aesthetic flaw.

These kinds of translation misalignments are common across languages. However, in the *lemon* model, we cannot, for example, just have a concept in our ontology for a smallish area of trees, which *bois* and *Holz* map to, and one for a bigger area, which *forêt* and *Wald* map to. Rather, to properly represent the meanings of these words, we must have four separate language-dependent concepts in our ontology. (*lemon* allows language-dependent concepts to be defined for use within a specific lexicon, provided, of course, that the new concept is expressible in terms of the existing external ontology (Cimiano et al. 2014).) Additional languages complicate the picture further; for example, Dutch gives a spectrum of three words, *hout*, *bos*, and *woud* (Henry Schogt, p.c.). A language-independent ontological representation of the different kinds of errors that are lexically reified by various languages, a small sample of which was shown above, would be even more complex. Of course, an ontology may be “localized” to a particular language, as posited by Gracia et al. (2012), but cross-lingual mappings between localized ontologies will be very difficult in practice; the example given by Gracia et al. covers only one easy case where a term in one language neatly subsumes two in another (English *river*, French *fleuve* and *rivière*).

Edmonds and Hirst (2002) have proposed that instead of thus making the ontology ever more fine-grained as additional languages are taken into account, only relatively coarse-grained ontological information should be used in the lexicon, along with explicit differentiating information for nonhierarchically distinguished near-synonyms, both within and across languages—much as we saw in the examples above from Fujiwara et al. and Farrell, albeit in a formal representation. Drawing on this model, Inkpen and Hirst (2006) used the explicit differentiating information

---

<sup>3</sup>Thanks to Kazuko Nakajima for the translation of this text from Japanese.

in conventional dictionaries and dictionaries of near-synonym explication to develop knowledge bases of lexical differentiation for English and (minimally) for French. Inkpen and Hirst demonstrated that using this knowledge of lexical differences improved the quality of lexical choice in a (toy) translation system, using aligned French–English sentence pairs from the proceedings of the Canadian Parliament as test data. Nonetheless, differentiating information on nonhierarchically distinguished near-synonyms, within or across languages, might need to be used in inferences. A Multilingual Semantic Web cannot rely on only an ontology as an interlingual representation or as a nonlinguistic representation for inference; there is, in practice, no clean separation between the conceptual and the linguistic.

### 3 Given Meaning by Whom?

In BLHL’s vision, “information is given well-defined meaning”—but by whom? BLHL’s answer was clear: it would be done by the person who provides the information. “Ordinary users will compose Semantic Web pages and add new definitions and rules using off-the-shelf software that will assist with semantic markup” (BLHL, p. 36). That is, semantic markup—and even the creation of new ontological definitions and rules—is assumed to occur at page-creation time, either automatically or, more usually, semi-automatically with the assistance of the author, who is an “ordinary user”—the writer of a blog, perhaps. Hence, in this view a Semantic Web page has a single, fixed, semantic representation that (presumably) reflects its author’s view of what he or she wants or expects readers of the page to get from it. The markup is created in the context of the author’s personal and linguistic worldview.

This is a *writer-centered view of meaning*. It assumes that the context, background knowledge, and agenda that any potential user or reader of the page will draw on in understanding its content are the same as those of the author and that therefore the meaning that the user will take from the page is the same as the meaning that its creator put in. This is so both in the case that the user is a human looking at natural language text and in the case that the user is software looking at the semantic markup. It is a version of the *conduit metaphor* of communication (Reddy 1979), in which text (or markup) is viewed as a container into which meaning is stuffed and sent to a receiver who removes the meaning from the container and in doing so comprehends the text and thereby completes the communication. This view may also be thought of as *intention-centered*, in that, barring mistakes and accidents, the meaning received is the meaning that the author intended to convey.

Many potential uses of the Semantic Web fit naturally into the paradigm of markup for writer-based meaning and an intention-centered view. These uses are typically some kind of *intelligence gathering*, in the most general sense of that term—understanding what someone else is thinking, saying, or doing. That is, the user’s question, looking at some text, is “What are they trying to tell me?”

(Hirst 2007, 2008). Tasks that fit this paradigm, in addition to simple searches for objectively factual information, include sentiment analysis and classification, opinion extraction, and ideological analysis of texts—for example, finding a well-reviewed hotel in a particular city. In each of these tasks, determining a writer’s intent is the explicit goal, or part of it, and the markup will help to do this.

Future methods of automatic translation of Semantic Web pages also fall under this paradigm. The goal of translation is to reproduce the author’s intent as well as possible in the target language. Translation systems will be able to use both the original natural language text and the author’s markup in order to produce a translation that is more accurate and more faithful to the author’s intent than a system relying on the text alone could produce.

However, this writer- and intention-centered view is too constraining and restrictive for fully effective use of the Semantic Web—in fact, for many of the primary uses of the Semantic Web. Consider, for example, the limitations that this view puts on search. For a search to usefully take domain circumscriptions and shared ontologies into account, the user must be thinking and searching in the same terms as those of the author of the information that the user wishes to find. If there is a conceptual mismatch, then the information sought might not be found at all—an outcome no better than a simple string-matching search with unfortunately chosen terms.<sup>4</sup> And this leads to my next point.

## 4 Work Together for Whose Benefit?

In BLHL’s vision, the Semantic Web will “better [enable] computers and people to work in cooperation [with each other].” But for whose benefit is this? The Semantic Web vision rightly emphasizes the benefit of the *information seeker*, whose task will be made easier and who will be given a greater chance of success. The benefit to the *information provider*, who wants to bring their information to the notice of the world for commercial, administrative, or other purposes, is secondary and often indirect.

And this is why a strictly writer-based view of meaning is inappropriate for the Semantic Web. Much of the potential value of querying the Semantic Web is that the system may act on behalf of the user, finding relevance in, or connections between,

---

<sup>4</sup>For example, contemporary researchers in biodiversity have trouble searching the legacy literature in the field because diachronic changes both in the terminology and in the conceptual understanding of the domain result in there being no shared ontologies. “Even competent and well-intentioned researchers often have difficulties searching this literature. Simple Google-style keyword searches are frequently insufficient, because in this literature, more so perhaps than most other fields of science, related concepts are often described or explained in different terms, or in completely different conceptual frameworks, from those of contemporary research. As a result, interesting and beneficial relations with legacy publications, or even with whole literatures, may remain hidden to term-based methods” (Hirst et al. 2013).

interpretations of texts that go beyond anything that the original authors of those texts intended. For example, if the user wants to find, say, evidence that society is too tolerant of intoxicated drivers or evidence that the government is doing a poor job or evidence that the Philippines has the technical resources to commence a nuclear-weapons program, then a relevant text need not contain any particular set of words nor anything that could be regarded as a literal assertion about the question (although it might), and the writer of a relevant text need not have had any intent that it provide such evidence (Hirst 2007, p. 275).

But for a Semantic Web system to find situations in which a document *unintentionally* answers an information seeker's query, it must embody also a *reader-centered view of meaning*. It must be able to ask, on behalf of the user, "What does this text mean to me?" (Hirst 2007, 2008). In its most general form, this is a postmodern view of text, in which the interpretation of each reader, or each community of like-minded readers, may be different (Fish 1980). Here, however, we need only a more limited view: that the system understand the user's goal or purpose in their search and, ideally, the user's viewpoint, beliefs, or ideology and "anything else known about the user" (Hirst 2007, p. 275). That is, a *user model* is available to the system, and, moreover, an agent local to the user's search interface has possibly inferred (or been explicitly told) the broader context or purpose of the user's current activity. The elements of the user model might, in turn, be partially derived or inferred from the system's observation of the user's prior reading and prior searches, in addition to feedback and possibly explicit training from the user. It would start as a generic model and then adapt and accommodate itself to the individual user, becoming more precise and refined (Hirst 2009b). In particular, the user model might include aspects of the user's beliefs and values and their reflections in ontology and lexis—for example, which shared ontologies the user accepts and which ones they reject. These factors may then be used in the search to answer the user's query, perhaps becoming part of the query itself and being used in matching and inference processes to interpret Semantic Web pages.

Consider, for example, a user who wants to know whether they should spend their time and money on a certain movie. A writer-centered Semantic Web would require them to ask a *proxy question* such as "Did other people like this movie?", whereas a reader-centered Semantic Web would allow them to ask their real question, "Will *I* like this movie?". If the system knows, from its model of the user, that they prefer quiet, intelligent movies, then a disgruntled review criticizing the movie for its lack of action could be interpreted as a positive answer to the question. More generally, a reader-centered perspective is particularly useful for abstract, ideological, wide-ranging, or unusual questions and for tasks such as nonfactoid question-answering and query-oriented multi-document summarization where interpretation is an essential part of the task.

Of course, as the movie example above suggests, it may still, in the end, be the writer's annotation to which a reader-centered matching process will be applied. However, the writer's annotation need no longer be the only annotation considered. Whenever a user's query matches a page, the retrieval software may add an annotation to that page with the reader-centered interpretation and inferences that



are produced and the reader characteristics upon which they are based. This will facilitate future matching by similar readers with similar queries. Thus, in time a Semantic Web page might bear many different annotations reflecting many different interpretations, not merely that of the writer.<sup>5</sup> In particular, for the Multilingual Semantic Web, these annotations may include translations and glosses that future processes may use.

None of this is to say that the writer-centered view isn't valuable too; as we noted earlier, many intelligence-gathering tasks fit that paradigm. The ideal Semantic Web would embody both views. And the ontological resources, markup, and inference mechanisms of a writer-centered Semantic Web are a prerequisite for the additional mechanisms of a reader-centered view.

## 5 Overcoming Linguistic (and Representational) Barriers

The discussion above gives us a starting point for thinking about what our next steps should be toward a monolingual or Multilingual Semantic Web that includes textual information. First, it implies that we must, in some ways, lower our expectations. We must give up, at least *pro tem*, the goal of creating a Semantic Web that relies on high-quality knowledge-based semantic interpretation and translation or understanding across languages. We must accept that any semantic representation of text will be only partial and will be concentrated on facets of the text for which a first-order or near-first-order representation can be constructed and for which some relatively language-independent ontological grounding has been defined. Hence, the semantic representation of a text may be incomplete, patchy, and heterogeneous, with different levels of analysis in different places (Hirst and Ryan 1992). We must also accept that the Semantic Web will be limited, at least in the initial stages, to a static, writer-centered view of meaning.

However, we should *not* take the view that the Semantic Web will remain “incomplete” until BLHL's vision is realized. Rather, we should say that at each step along the way it will on the one hand be a useful artifact and on the other hand will remain “imperfect.” The difference is that an *incomplete* Semantic Web would be missing certain features or abilities but would be fully realized in other respects; the underlying metaphor is one of piece-by-piece construction from components that are each already individually complete and perfect at the time that they are added, and the construction is complete when, and only when, the final component

---

<sup>5</sup>The collaborative annotation of a Semantic Web page with semantic interpretations generated by software agents that are beyond the control of its author raises many issues that are outside the scope of this chapter. The annotations might be objectionable to the author or counterproductive to his or her goals; they could be willfully misleading or outright vandalism. While these issues may also arise with the present-day public tagging or bookmarking of sites by users (Breslin et al. 2009), their scale is greatly magnified when the annotations become a central part of the Semantic Web retrieval mechanism rather than merely some user's advisory opinion.

is put in place (even if partial usability is achieved at an earlier stage). By contrast, none or almost none of the features and abilities of an *imperfect* Semantic Web will be fully realized, and it will only imperfectly reflect BLHL's vision; the underlying metaphor is one of growth or evolution, in which even an immature organism is, in an important sense, complete even if not fully functional.

The practical difference between these two views of the development of the Semantic Web is that they lead to different research strategies. And, crucially, we should recognize that the second view is not a lowering but a *raising* of expectations. Why? It reflects the change of view that occurred in computational linguistics and natural language processing in the mid-to-late 1990s, and these fields have been enormously successful since they gave up the too-far-out (or maybe impossible) goal of high-quality knowledge-based semantic interpretation (Hirst 2013) (see Sect. 2). Contemporary NLP and CL have little reliance on symbolic representations of knowledge and of text meaning and far less emphasis on precise results and perfect disambiguation. We have realized that imperfect methods based on statistics and machine learning frequently have great utility; not every linguistic task requires humanlike understanding with 100% accurate answers; many tasks are highly tolerant of a degree of fuzz and error.

Many other areas of artificial intelligence and knowledge representation came to a similar realization in the last decade or so—just about the time that BLHL's paper was published, but not in time to influence it. In simple terms, BLHL's vision of the Semantic Web is Old School. There needs to be space in the Multilingual Semantic Web for the kinds of imperfect methods now used in NLP and for the textual representations that they imply. In particular, research on vector-based (or tensor-based) distributional semantics (e.g., Turney and Pantel 2010; Clarke 2012; Erk 2012) has reached the point where compositional representations of sentences are now in view (Baroni et al. 2014), and research on distributional methods of semantic relatedness (e.g., Mohammad and Hirst 2006; Hirst and Mohammad 2011) is being extended to cross-lingual methods (e.g., Mohammad et al. 2007; Kennedy and Hirst 2012).

Distributional representations don't meet the "well-defined meaning" criterion of being overtly precise and unambiguous. But it's exactly because of this that they also offer hints of the possibility of reader-centered views of the Semantic Web. Broad distributional representations of a user's search goal, possibly further refined by specific knowledge of other aspects of the user, may match representations of Semantic Web pages that would not be matched by a more precise, symbolic representation of the same goal.

Nonetheless, this can work only if there is agreement on how these representations are constructed from text, including the corpora from which the distributional data are derived. We can envision the development of some kind of standardized lexical or ontological vector representation and principles of composition and, moreover, a method of extending the representation across languages. In particular, taking the matter of near-synonymy across languages seriously, we would require that cross-lingual near-synonyms have recognizably similar representations, and hence cross-lingual sentence paraphrases would too.

We should expect to see symbolic representations of textual data increasingly pushed to one side as monolingual and cross-lingual methods are further developed in distributional semantics and semantic relatedness (and a few Semantic Web researchers have already begun some very preliminary investigations Nováček et al. 2011; Freitas et al. 2013). I say this with some caution, as the kind of compositional distributional semantics that could represent phrase and sentence meaning, not just word meaning, and could support useful inference is still at a very early stage of development (e.g., Mitchell and Lapata 2010; Erk 2013; Baroni et al. 2014). In particular, there is no hint yet of a theory of inference for these representations. The whole enterprise might yet fail. But even if this turns out to be so, the broader point remains—that the future of semantic representations for the Multilingual Semantic Web is likely to lie in imperfect nonsymbolic methods that work well enough in practice for most situations.

**Acknowledgements** This work was supported financially by the Natural Sciences and Engineering Research Council of Canada. For helpful comments, I am grateful to Lars Borin, Philipp Cimiano, Nadia Talent, the anonymous reviewers, and the participants of the Dagstuhl Seminar on the Multilingual Semantic Web.

## References

- Adamska-Sałaciak, A. (2013). Equivalence, synonymy, and sameness of meaning in a bilingual dictionary. *International Journal of Lexicography*, 26(3), 329–345. doi:10.1093/ijl/ect016.
- Baroni, M., Bernardi, R., & Zamparelli, R. (2014). Frege in space: A program for compositional distributional semantics. *Linguistic Issues in Language Technology*, 9(6) (February 2014).
- Berners-Lee, T. (2009). The next Web. In *TED Conference*, Long Beach, CA. [www.ted.com/talks/tim\\_berners\\_lee\\_on\\_the\\_next\\_web.html](http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html).
- Berners-Lee, T., Hendler, J., & Lassila, O. (2001, May). The Semantic Web. *Scientific American*, 284(5), 34–43.
- Borin, L. (2012). Core vocabulary: A useful but mystical concept in some kinds of linguistics. In D. Santos, K. Lindén, & W. Ng'ang'a (Eds.), *Shall we play the Festschrift game?* (pp. 53–65). Berlin: Springer. doi:10.1007/978-3-642-30773-7\_6.
- Breslin, J. G., Passant, A., & Decker, S. (2009). *The social Semantic Web*. Berlin: Springer. doi:10.1007/978-3-642-01172-6.
- Cimiano, P., Buitelaar, P., McCrae, J., & Sintek, M. (2011). LexInfo: A declarative model for the lexicon–ontology interface. *Journal of Web Semantics*, 9(1), 29–51. doi:10.1016/j.websem.2010.11.001.
- Cimiano, P., Unger, C., & McCrae, J. (2014). *Ontology-based interpretation of natural language*. San Rafael: Morgan & Claypool Publishers.
- Clarke, D. (2012). A context-theoretic framework for compositionality in distributional semantics. *Computational Linguistics*, 38(1), 41–71. doi:10.1162/COLI\_a\_00084.
- Edmonds, P., & Hirst, G. (2002). Near-synonymy and lexical choice. *Computational Linguistics*, 28(2), 105–144. doi:10.1162/089120102760173625.
- Erk, K. (2012). Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10), 635–653. doi:10.1002/lnco.362.
- Erk, K. (2013). Towards a semantics for distributional representations. In *Proceedings, 10th International Conference on Computational Semantics (IWCS-2013)*, Potsdam. [www.aclweb.org/anthology/W13-0109](http://www.aclweb.org/anthology/W13-0109).

- Farrell, R. B. (1977). *German synonyms*. Cambridge: Cambridge University Press.
- Fish, S. (1980). *Is there a text in this class? The authority of interpretive communities*. Cambridge: Harvard University Press.
- Freitas, A., O’Riain, S., & Curry, E. (2013). A distributional semantic search infrastructure for linked dataspaces. In *The Semantic Web: ESWC 2013 Satellite Events. Lecture Notes in Computer Science* (Vol. 7955, pp. 214–218). Berlin: Springer. doi:10.1007/978-3-642-41242-4\_27.
- Friedman, J., Moran, D. B., & Warren, D. S. (1978). Explicit finite intensional models for PTQ. *American Journal of Computational Linguistics, microfiche 74*, 3–22. [www.aclweb.org/anthology/J79-1074](http://www.aclweb.org/anthology/J79-1074)
- Fujiwara, Y., Isogai, H., & Muroyama, T. (1985). *Hyogen Ruigo Jiten*. Tokyo: Tokyodo Publishing.
- Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Gómez-Pérez, A., Buitelaar, P., & McCrae, J. (2012). Challenges for the multilingual Web of data. *Journal of Web Semantics, 11*, 63–71. doi:10.1016/j.websem.2011.09.001
- Hirst, G. (2007). Views of text-meaning in computational linguistics: Past, present, and future. In G. Dodig Crnkovic & S. Stuart (Eds.), *Computation, information, cognition — The Nexus and the Liminal* (pp. 270–279). Newcastle: Cambridge Scholars Publishing. [ftp.cs.toronto.edu/pub/gh/Hirst-ECAPbook-2007.pdf](http://ftp.cs.toronto.edu/pub/gh/Hirst-ECAPbook-2007.pdf).
- Hirst, G. (2008). The future of text-meaning in computational linguistics. In P. Sojka, A. Horák, I. Kopeček, & K. Pala (Eds.), *Proceedings, 11th International Conference on Text, Speech and Dialogue (TSD 2008). Lecture Notes in Artificial Intelligence* (Vol. 5246, pp. 1–9). Berlin: Springer. doi:10.1007/978-3-540-87391-4\_1.
- Hirst, G. (2009a). Ontology and the lexicon. In S. Staab & R. Studer (Eds.), *Handbook on ontologies. International Handbooks on Information Systems* (2nd ed., pp. 269–292). Berlin: Springer. doi:10.1007/978-3-540-92673-3\_12.
- Hirst, G. (2009b, July). Limitations of the philosophy of language understanding implicit in computational linguistics. *Proceedings, 7th European Conference on Computing and Philosophy*, Barcelona (pp. 108–109). [ftp.cs.toronto.edu/pub/gh/Hirst-ECAP-2009.pdf](http://ftp.cs.toronto.edu/pub/gh/Hirst-ECAP-2009.pdf).
- Hirst, G. (2013). Computational linguistics. In K. Allan (Ed.), *The Oxford handbook of the history of linguistics*. Oxford: Oxford University Press.
- Hirst, G., & Mohammad, S. (2011). Semantic distance measures with distributional profiles of coarse-grained concepts. In A. Mehler, K. U. Kühnberger, H. Lobin, H. Lungen, A. Storrer, & A. Witt (Eds.), *Modeling, learning, and processing of text technological data structures. Studies in Computational Intelligence Series* (Vol. 370, pp. 61–79). Berlin: Springer. doi:10.1007/978-3-642-22613-7\_4.
- Hirst, G., & Ryan, M. (1992). Mixed-depth representations for natural language text. In P. S. Jacobs (Ed.), *Text-based intelligent systems* (pp. 59–82). Hillsdale, NJ: Lawrence Erlbaum Associates. [ftp.cs.toronto.edu/pub/gh/Hirst+Ryan-92.pdf](http://ftp.cs.toronto.edu/pub/gh/Hirst+Ryan-92.pdf).
- Hirst, G., Talent, N., & Scharf, S. (2013, 27 May). Detecting semantic overlap and discovering precedents in the biodiversity research literature. In *Proceedings of the First International Workshop on Semantics for Biodiversity (S4BioDiv)* (CEUR Workshop Proceedings, Vol. 979), *10th Extended Semantic Web Conference (ESWC-2013)*, Montpellier, France. [ceur-ws.org/Vol-979/](http://ceur-ws.org/Vol-979/).
- Hjelmslev, L. (1961). *Prolegomena to a theory of language* (rev. ed.). (F. J. Whitfield, Trans.). Madison: University of Wisconsin Press. (Originally published as *Omkring sprogteoriens grundlæggelse*, 1943.)
- Inkpen, D., & Hirst, G. (2006). Building and using a lexical knowledge-base of near-synonym differences. *Computational Linguistics, 32*(2), 223–262. [www.aclweb.org/anthology/J06-2003](http://www.aclweb.org/anthology/J06-2003)
- Kennedy, A., & Hirst, G. (2012, December). Measuring semantic relatedness across languages. In *Proceedings, xLiTe: Cross-Lingual Technologies Workshop at the Neural Information Processing Systems Conference*, Lake Tahoe, NV. [ftp.cs.toronto.edu/pub/gh/Hirst-ECAP-2009.pdf](http://ftp.cs.toronto.edu/pub/gh/Hirst-ECAP-2009.pdf).

- McCrae, J., Aguado-de-Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., et al. (2012). Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 46(4), 701–719. doi:10.1007/s10579-012-9182-3.
- Mitchell, J., & Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science*, 34(8), 1388–1429. doi:10.1111/j.1551-6709.2010.01106.x.
- Mohammad, S., Gurevych, I., Hirst, G., & Zesch, T. (2007). Cross-lingual distributional profiles of concepts for measuring semantic distance. In *2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, Prague (pp. 571–580). [www.aclweb.org/anthology/D07-1060](http://www.aclweb.org/anthology/D07-1060).
- Mohammad, S., & Hirst, G. (2006, July). Distributional measures of concept-distance: A task-oriented evaluation. In *Proceedings, 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, Sydney, Australia (pp. 35–43). [www.aclweb.org/anthology/W06-1605](http://www.aclweb.org/anthology/W06-1605).
- Montague, R. (1974). *Formal philosophy*. New Haven: Yale University Press.
- Nováček, V., Handschuh, S., & Decker, S. (2011). Getting the meaning right: A complementary distributional layer for the web semantics. In *Proceedings, 10th International Semantic Web Conference (ISWC-2011)* (Vol. 1, pp. 504–519). *Lecture Notes in Computer Science*, Vol. 7031. Berlin: Springer. doi:10.1007/978-3-642-25073-6\_32.
- Reddy, M. J. (1979). The conduit metaphor: A case of frame conflict in our language about language. In A. Ortony (Ed.), *Metaphor and thought* (pp. 284–324). Oxford: Oxford University Press. [Reprinted unchanged in the second edition, 1993, pp. 164–201.]
- Schogt, H. G. (1976). *Sémantique synchronique: synonymie, homonymie, polysémie*. Toronto: University of Toronto Press.
- Schogt, H. G. (1988). *Linguistics, literary analysis, and literary translation*. Toronto: University of Toronto Press.
- Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37, 141–188. doi:10.1613/jair.2934.
- Wilks, Y. (2009). Ontotherapy, or how to stop worrying about what there is. In N. Nicolov, G. Angelova, & R. Mitkov (Eds.), *Recent advances in natural language processing V* (pp. 1–20). Amsterdam: John Benjamins.

# Design Patterns for Engineering the Ontology-Lexicon Interface

John P. McCrae and Christina Unger

**Abstract** In this paper, we combine two ideas: one is the recently identified need to extend ontologies with a richer lexical layer, and the other is the use of ontology design patterns for ontology engineering. We combine both to develop the first set of design patterns for ontology-lexica, using the ontology-lexicon model, *lemon*. We show how these patterns can be used to model nouns, verbs and adjectives and what implications these patterns impose on both the lexicon and the ontology. We implemented these patterns by means of a domain-specific language that can generate the patterns from a short description, which can significantly reduce the effort in developing ontology-lexica. We exemplify this with the use case of constructing a lexicon for the DBpedia ontology.

**Key Words** Design patterns • Lexicon • Ontology • Ontology engineering • Ontology-lexica

## 1 Introduction

Ontology design patterns (Gangemi and Presutti 2009) are a method of formalising commonly used structures in ontologies and in particular have been proposed for Web Ontology Language (OWL) (McGuinness and Van Harmelen 2004) ontologies. Recently, there has been interest in extending the lexical context of ontologies, to create what has been dubbed an *ontology-lexicon* (Prévoit et al. 2010). As such, a number of models have been proposed for representing this *ontology-lexicon interface* (Montiel-Ponsoda et al. 2008; Cimiano et al. 2011; Buitelaar et al. 2009; Reymonet et al. 2007), in particular the *Lexicon Model for Ontologies* (McCrae et al. 2012a, *lemon*). We take this model as our basis and consider how we model ontology-specific semantics of lexical entries and their linguistic properties, so that they can be used in NLP applications. We approach this by the use of design patterns

---

J.P. McCrae • C. Unger (✉)

AG Semantic Computing, CITEC, Bielefeld University, Bielefeld, Germany  
e-mail: [jmccrae@cit-ec.uni-bielefeld.de](mailto:jmccrae@cit-ec.uni-bielefeld.de); [cunger@cit-ec.uni-bielefeld.de](mailto:cunger@cit-ec.uni-bielefeld.de)

that define how certain lexico-semantic phenomena should be modelled and also a small complementary meta-ontology, which we call *lemonOILS* (*Lemon Ontology for the Interpretation of Lexical Semantics*). This meta-ontology captures basic semantic concepts such as events and scalar qualities but differs strongly from *top-level ontologies* (Gangemi et al. 2002), in that it is orientated towards engineering ontology-lexica for Natural Language Processing (NLP) applications. As such, we describe different forms of modelling in the ontology-lexicon interface, rather than philosophical distinctions.

Our goal in creating such a catalogue of *ontology-lexicon design patterns* is to ameliorate the process of developing ontology-lexica, by replacing complex combinations of frame semantics and first-order logic axioms with simple patterns with few parameters. This would allow the quick development of lexica for ontologies existing on the Semantic Web and so enable these lexica to be developed quickly in multiple languages, which would in turn enable the development of tools such as question answering systems (Unger et al. 2010) to enable this content to be accessed by users. The patterns that we propose in this paper are designed to be useful not only for English languages but to be portable to any language, and as such we have designed the patterns to be language independent. Thus, we believe that the use of these patterns should not only help the process of developing monolingual lexica but also in the translation of these lexica to new languages.

To enable the use of these patterns, we start by developing the patterns in terms of a domain-specific language (Fowler and Parsons 2010; Wampler and Payne 2008) that can generate the suitable axioms and frames in Resource Description Framework (RDF) using OWL and *lemon*. Furthermore, we apply these patterns to the creation of a lexicon for the DBpedia ontology (Bizer et al. 2009) and demonstrate how this improves the ontology-lexicon engineering process (McCrae et al. 2012b).

## 2 Lexicon-Ontology Modelling with *Lemon* and OWL

The *lemon* model (McCrae et al. 2012a) is a model for representing lexica and machine-readable dictionaries relative to ontologies by a principle called *semantics by reference* (Cimiano et al. 2013). This means that the meaning of a word is given by reference to an ontology, resulting in a clean separation between the lexical and semantic layer. *lemon* consists of a small core model and a number of additional modules. The core model consists of the following elements:

- **Lexical entry:** The object which represents the entry in the lexicon
- **Lexical form:** An object representing an inflected form of an entry
- **Representation:** The character string representing a form in a given orthography

- **Lexical sense:** The object representing the meaning of the object and its properties that depend on both the meaning and the form of the entry, such as register or translations
- **Ontological reference:** The interpretation of the sign in a logical form (ontology)

There are a number of additional modules,<sup>1</sup> but for this paper we will focus on the *syntax and mapping module*, which describes how frames are constructed and linked to ontology predicates. In *lemon*, an entry may have any number of *frames*, each of which has a number of *arguments* linked by means of *syntactic role* properties. Classes of frames are characterised by the syntactic roles, for example, a transitive frame is sufficiently a frame with a subject and direct object argument. Following the *lemon* philosophy of being descriptive not prescriptive, the set of syntactic roles and frame classes are defined in an external ontology. For this we use the LexInfo2 ontology (Cimiano et al. 2011).

On the ontological side, it is assumed that there are ground symbols (OWL individuals), unary predicates (OWL classes) and binary predicates (OWL properties), and the arguments of each frame are linked to each sense by means of one of the following properties:

- `subjOfProp`: Indicates the first argument of a binary predicate and the subject of a triple
- `objOfProp`: The second argument of a binary predicate and the object of a triple
- `isA`: Indicates the argument of a unary predicate and the subject of a `rdf:type` triple

As the formalism provided by OWL does not allow the direct modelling of higher arity predicates, predicates with arity greater than two are modelled by composing frames by means of *compound senses* composed of a number of *atomic senses*. Each of these atomic senses refers to a property in the ontology, and the compound sense may refer to the class in the ontology of the argument that is shared by all predicates.<sup>2</sup> For example, we may decompose the predicate `Give` as a reified event, where each thematic role is represented through a predicate:

$$\text{Give}(x, y, z) \equiv \exists e : \text{GivingEvent}(e) \wedge \text{Giver}(e, x) \wedge \text{Recipient}(e, y) \wedge \text{Given}(e, z)$$

Note that it is not required that the introduced argument is the subject of all subsenses but this is the most frequent case and the only case dealt with by our patterns.

---

<sup>1</sup>For a complete list, see <http://lemon-model.net/>.

<sup>2</sup>Note that properties may consist of chains of properties, giving multiple unbound arguments. In this case, OWL 2 property chains should be used to reduce to one unbound argument.



### 3 Design Patterns

Our catalogue of patterns includes noun, verb and adjective patterns. We start by looking at the case of common and proper nouns, breaking them down into cases where they denote classes and cases where they denote relations. Next, we turn to verbs, where we consider the division of verbs into activities, achievements, accomplishments and states (Vendler 1957), and argue that stative verbs should be fundamentally separated from event verbs, as they represent the most common form of verbs modelled by ontologies. Finally, we turn to the case of adjectives, as studied by Raskin and Nirenburg (1995), and following Bouillon and Viegas (1999) we split adjectives into four classes: *intersective*, *property-modifying*, *relational* and *scalar* adjectives. As a novel contribution, we show how these can be modelled using *lemon* and OWL, show fundamental limits of description logics (Baader 2003) and consider how the modelling may be extended in more flexible logic formalisms.

#### 3.1 Names and Nouns

We start our catalogue by making a fundamental distinction between common nouns and proper nouns. For proper nouns (names), we define a preferred pattern that is a single entry annotated with `partOfSpeech=properNoun` and linked to an ontology entity of OWL type `NamedIndividual`, as shown in Fig. 1.

For common nouns, we distinguish *class nouns* and *relational nouns*. Class nouns, as pictured in Fig. 2, represent the class of nouns that indicate the genus of an object in the world, such as “mountain”; here the pattern is simply made with an entity with `partOfSpeech=commonNoun` and OWL type `Class`.

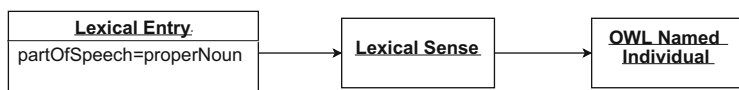


Fig. 1 The design pattern for names

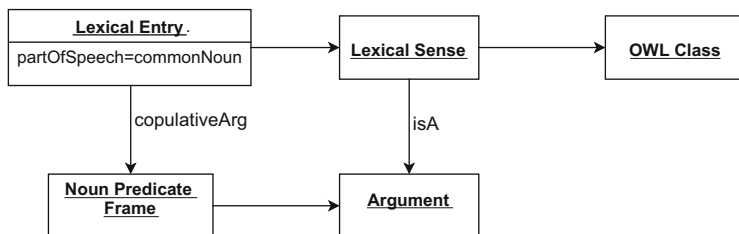


Fig. 2 The design pattern for class nouns. The `isA` role indicates that the argument refers to the “is instance of” relation to the class

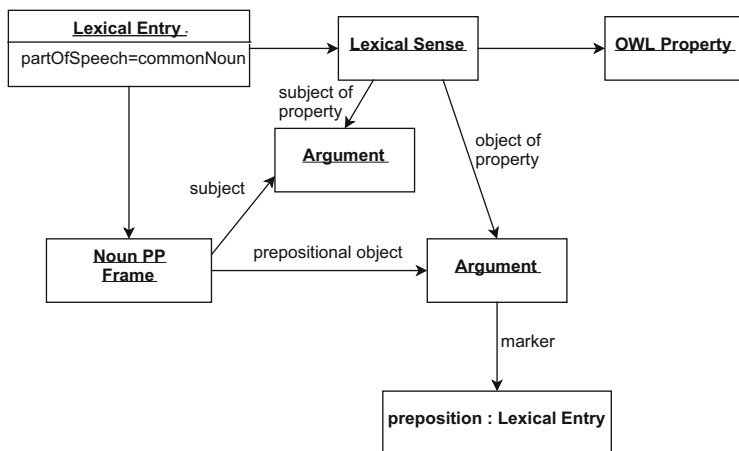


Fig. 3 The design pattern for relational nouns

In addition, we include a single frame to indicate the attributive usage of the noun, for example, “*X* is a mountain”, and its reference to an ontological concept, for example, *Mountain* in the DBpedia ontology (Bizer et al. 2009). This frame’s argument is called the *copulative argument*, which is used in place of the more general *subject*, in order to allow for languages that use a zero copula and do not have or frequently use a verb equivalent to “to be”.

Relational nouns, on the other hand, indicate a relation between two entities. We further divide them into the following two classes.

- **Bivalent:** Here the noun corresponds to some property, as in the case of “capital of” lexicalising the DBpedia property *capital*, for example. As OWL only allows predicates with at most two arguments, we limit ourselves to this case. The modelling is shown in Fig. 3.<sup>3</sup>
- **Multivalent:** Here the noun corresponds to a property with potentially many arguments. This pattern is similar to the bivalent pattern but is modified to overcome the limitations of OWL. We introduce a new class to model this, called *oils:Relationship*.<sup>4</sup>

In addition, we define a convenience pattern that combines both the class noun pattern and the (bivalent) relational noun pattern. This is common for nouns such as “father”, where there are both a property between fathers and children and a class of all people who are fathers of some child (usually encoded by an ontology axiom).

<sup>3</sup>Depending on whether the argument of the noun is a prepositional object, as in “marriage with someone”, or also allows for a possessive construction, as in “a country’s capital”, the syntactic frame is specified either as *NounPPFrame* or as *NounPossessiveFrame*.

<sup>4</sup>The *oils* name space is <http://lemon-model.net/oils#>.

### 3.2 *State Verbs*

For verbs we argue that the most important distinction is between state and event verbs. State verbs are of primary interest for several reasons. Firstly, in existing ontologies, verb labels nearly always indicate a state. Secondly, states are useful in applications that do not model time, for example, in business rules systems (Halle and Ronald 2001), which model processes in terms of alethic and deontic statements, such as “ $X$  must possess  $Y$ ” or “ $X$  should be capable of  $Y$ ”. While states are frequently also temporal entities<sup>5</sup> that model a certain property of something that holds in a certain time interval, binary properties in OWL ontologies are typically specified atemporally. Finally, state verbs conform to the intended usage of properties in OWL, which was to indicate the relationships between resources on the Web or properties of these resources, and these are considered to be true only within a context (such contexts extend the triple model to a quad model (Tappolet and Bernstein 2009)).

We introduce two patterns for modelling state verbs, one for bivalent and one for the multivalent cases, which in practice are very similar to the corresponding noun patterns.

### 3.3 *Event Verbs*

We argue that events have fundamentally different semantics to states and take an approach based on *Davidsonian event semantics* (Davidson 1967). To that extent, we introduce a class into *lemonOILS*, called `oils:Event`, that can take any number of arguments. Furthermore, we allow two aspect properties (Comrie 1976) to be specified:

**Telicity:** Indicates whether the event has a clear end or is a continuous activity, for example, “to score a goal”, which has a clear end, or “to (be able to) play a musical instrument”, which does not. The design pattern for telic verbs is depicted in Fig. 4.

**Durativity:** Indicates whether the event occurs for a fixed period of time or whether the action is an instantaneous event, for example, “to travel to” has a clear duration, whereas “to arrive in” does not.

If these properties are set, appropriate axioms are introduced to the event class based on the *lemonOILS* properties `oils:begin`, `oils:end`, `oils:duration` and `oils:time`.

---

<sup>5</sup>For example, the state of being larger is atemporal for natural numbers but temporal for the height of children.

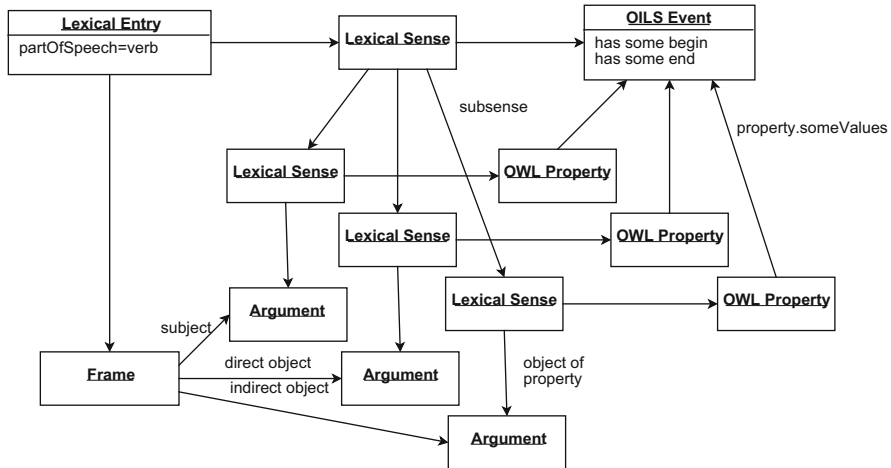


Fig. 4 The design pattern for telic verbs

### 3.3.1 Consequences of an Event

Finally, we introduce an extra pattern for states which arise from the completion of telic events. As an example of this, we take the frame “X was born in Y”, which directly refers to some birth event but whose arguments are theme and location properties of this event. In fact, such a modelling is uncommon in existing ontologies, but it can be stated that the property chain  $theme^{-1} \circ location$  implies a single property `birthPlace`. The consequence pattern introduces both a telic event frame for the event and a sense that refers directly to the consequent factive state. We then introduce an axiom in the ontology that indicates the link as follows:

$$theme^{-1} \circ location \sqsubseteq birthPlace$$

If the theme and location properties are not stated in the ontology, new properties are introduced. This means that the process of interpreting the sentence “Lenin was born in Ulyanovsk” involves first inferring an event for the birth of Lenin, with a theme as “Lenin” and a location of “Ulyanovsk”. From this it is possible to infer the `birthPlace` property. Modelling this property directly as a state verb would not indicate that the past tense should be used to express the property.

### 3.4 Adjectives

Adjectives are split into four main categories. Firstly, we consider intersective adjectives, which have an intersective semantics and are defined in the ontology by a class. This is the most straightforward group of adjectives and covers those that are logically defined by an intersection of the adjective and the known class (e.g. defined by the noun in an attributive construction). For example, “Belgian” is an intersective adjective, as “Belgian women” is the intersection of being from Belgium and being a woman. For the design patterns, we quickly noted that these adjectives are often in fact defined by a property and value in existing OWL ontologies, for example, “Belgian” may be represented by an object property `nationality` with the resource `Belgium` as object or by a datatype property `citizenship` with the literal “be” as object. As such we use three separate patterns for intersective adjectives with frame for both predicative and attributive usage:

- **Intersective class adjectives:** The simple case where there is a named class.
- **Intersective object property adjective:** The lexical entry is interpreted in the ontology as an axiom of the form  $\exists prop.\{i\}$  (for some individual  $i$ ).
- **Intersective data property adjective:** Similar to above, the axiom is of the form  $\exists prop.\{v\}$  (for some data value  $v$ ).

#### 3.4.1 Property-Modifying Adjectives

Property-modifying adjectives are considered to be those that modify the meaning of the class they apply to; an example of this is “former”, which may be described by an ontological property `heldRole`. This is taken to be the meaning of the lexical entry, and a frame is created to describe this attributive usage, “ $X$  is a former  $Y$ ”. As these adjectives require that there is a class noun to modify, these adjectives only have attributive frames. We do not provide modelling for adjectives that indicate semi-intersective subtypes such as “polar bear”; instead we assume that these are modelled as multi-word expressions using a class noun pattern.

#### 3.4.2 Relational Adjectives

Relational adjectives describe a relationship between two individuals such as “similar”, as in “ $X$  is similar to  $Y$ ”. The pattern for these describes a simple relationship to an object property, and a frame describing the predicative usage is described. In addition, a similar frame called *class relational adjective* that admits a frame for the attributive usage and associates it with some class can be used for modelling such as for “useful (for)”.

### 3.4.3 Scalar Adjectives

Scalar adjectives are generally very hard to define in formal logic. This problem is caused by the intuition that scalar adjectives, such as “big”, represent a fuzzy concept. Following Raskin and Nirenburg (1995), we wish to model this in a manner that is decidable in description logic, and as in Cimiano et al. (2011) we do this by defining a threshold on a per-class basis, that is, axioms of the form “Buildings taller than 5 stories are ‘big’”. Of course, this is unsatisfactory as a modelling in the ontology-lexicon, and a solution is to use membership degrees (Raskin and Nirenburg 1998) to give a general definition of scalar adjectives, that is to say “Big things are those that are in the top quartile of size for their class”. We reject this for the case of OWL as such a statement is inherently non-monotonic (as knowledge of more objects will change the quartile boundaries) and this is incompatible with the monotonic nature of description logic.<sup>6</sup> Furthermore, we note that this pattern should be used with extra caution as it can lead to inconsistent modelling. For example, if we define “big” for dogs and also define breeds of dogs as subclasses of dogs, then we can lead to contradictions, as follows: If every “Shih Tzu” is a “dog”, then every “big Shih Tzu” must be a “big dog”; however, it is clear that a “big Shih Tzu” cannot be considered to be a “big dog”.

In our implementation of scalar adjectives, we provide two forms, one that generates predicative and attributive frames and one that further generates comparative and superlative predicative frames for languages that have a particle comparative.<sup>7</sup>

While scalar classes are uncommon in ontologies, scalar adjectives are frequently used to lexicalise datatype properties. For example, the frame “*X* is *Y* (unit) high” may lexicalise a property `elevation`. We model this as a scalar adjective whose object is the adverbial phrase giving the value of the property and its unit.

## 4 Using the Pattern Catalogue for Ontology-Lexicon Engineering

To enable these patterns to be easily used, we created a *domain-specific language* (Wampler and Payne 2008, DSL) that enables lexica to be stated using a simple sublanguage of the Scala programming language or as an independent language defined by the BNF converter (Forsberg and Ranta 2003). The standard form of the pattern catalogue generates an RDF/XML model from the DSL, which can then be published on the Web or integrated with other *lemon* and RDF tools.

---

<sup>6</sup>Of course, handling the non-monotonic natural of language would be desirable; however, the introduction of non-monotonicity should occur at the ontology level, by means of extensions to the OWL language.

<sup>7</sup>In WALs this constitutes only 13 % of languages documented (Stassen 2011); however, as this includes the Romance and Germanic language families, we found this to be especially useful.

```

@prefix dbo: <http://dbpedia.org/ontology/> .
@prefix dbr: <http://dbpedia.org/resource/> .

Lexicon(<http://www.example.com/lexicon_en#>,"en",

  ClassNoun("company",dbo:Company)
           with plural "companies",

  RelationalNoun("owner",dbo:owner,
                propSubj = PossessiveAdjunct,
                propObj  = CopulativeArg),

  StateVerb("own",dbo:owner,
            propSubj = DirectObject,
            propObj  = Subject),

  IntersectiveObjectPropertyAdjective("Belgian",
                                       dbo:locationCountry,dbr:Belgium)
)

Lexicon(<http://www.example.com/lexicon_de#>,"de",

  ClassNoun("Firma",dbo:Company)
           feminine with plural "Firmen",
  ClassNoun("Unternehmen",dbo:Company)
           neuter with plural "Unternehmen",

  RelationalNoun("Inhaber",dbo:owner,
                propSubj = PrepositionalObject("von"),
                propObj  = CopulativeArg),

  StateVerb("gehören",dbo:owner,
            propObj  = IndirectObject),

  IntersectiveObjectPropertyAdjective("Belgisch",
                                       dbo:locationCountry,dbr:Belgium)
)

```

**Fig. 5** Examples of modelling using the DSL for *lemon* patterns

A formal grammar for the pattern DSL as well as the tools for compiling it is available at:

<http://github.com/jmccrae/lemon.patterns>

The DSL captures the patterns introduced in Sect. 3. An example showing the use of these pattern for a class noun, a relational noun, a state verb and an intersective adjective, lexicalising concepts of Belgian companies and their owners in German and English, is given in Fig. 5, which shows how with the DSL we can succinctly capture the multilingual lexical relationship between the lexicalisations. Following

the *lemon* principle of separating semantics and the lexicon, the links between English and German lexicalisations are achieved solely by reference to the ontology. Furthermore, note that much of the English code can be directly ported to German with only minor modifications, such as changing the forms and adding gender information.

## 5 Evaluation of the Patterns

We applied the developed patterns to the task of lexicalising a section of the DBpedia ontology (Bizer et al. 2009). In particular, we selected all classes (only excluding a few abstract ones or ones without instances) and all those properties that have more than 10,000 occurrences in the DBpedia dataset, yielding a set comprising 354 classes and 300 properties. We manually created a *lemon* lexicon for these classes and properties using the design patterns presented in this chapter, constructing 1,290 lexical entries. In particular, eight patterns were applied across 1,235 entries, while 56 entries (roughly 4 % of all entries) could not be represented as patterns (for a discussion, see below).

This work has led to the creation of the first *lemon* lexicon for DBpedia, available at [http://lemon-model.net/lexica/dbpedia\\_en](http://lemon-model.net/lexica/dbpedia_en) (Unger, et al. 2013).

In Table 1, we show the breakdown in the usage of each pattern. Class noun patterns are most frequent, representing the fact that there are more classes than properties in the selected part of DBpedia and that the part-of-speech variety in lexicalising those classes is very low. In fact, all classes are modelled using the `ClassNoun` pattern, except for 26 classes (one DBpedia class and 25 additionally defined restriction classes) which are modelled using the `IntersectiveAdjective` pattern, for example, verbalisations of nationalities such as “Russian”. All other patterns were used for lexicalising properties. Among those, about 60 % are verb

**Table 1** The usages of the design patterns in relation to a section of the DBpedia ontology

Pattern	Uses
Noun patterns	<b>955</b>
ClassNoun	692
RelationalNoun	263
Verb patterns	<b>207</b>
StateVerb	171
ConsequenceVerb	36
Adjective patterns	<b>173</b>
RelationalAdjective	62
IntersectiveAdjective	26
IntersectiveObjectPropertyAdjective	57
IntersectiveDataPropertyAdjective	28
Total	<b>1, 235</b>



patterns, mostly state verbs, and 40 % are adjective patterns, all of which are either relational or intersective.

That some of the lexicalisations could not be represented as patterns is mainly due to the following reasons. First, constructions are not yet covered by *lemon* but prove necessary for some verbalisations. For instance, “*X* has *Y* inhabitants” is a very common verbalisation of the property `population`, and “*X* consists of *Y* percent of water” is a straightforward verbalisation of `percentageOfAreaWater`. These cases account for about half of all entries that could not be represented as patterns.

Second, 18 entries required a compound sense comprising several subsenses. For instance, the adjective entry “active from *X* until *Y*” verbalises a combination of two properties, `activeStartDate` and `activeEndDate`.

And third, some of the entries require a syntactic behaviour that is not covered by the patterns. An example is the preposition “like”, as in “*X* is like *Y*”, verbalising the property `similar`. Prepositional frames are not covered by the patterns, as we argue that in *lemon* the main role of prepositions is the one they acquire in a specific frame. In isolation they have a domain-independent meaning that is usually very vague and can be fixed only in a specific linguistic context. The preposition “in”, for example, can generally be used denoting spatial relations (e.g. “Mount Everest is in Nepal”), temporal relations (e.g. “in 1963”) or a range of others (e.g. “Sofia Coppola is in *The Godfather*” and “2461 verses are in the book of Psalms”). Trying to list all possible usages of “in” with respect to DBpedia in the lexicon would not only be tedious but very likely also remain incomplete.

## 6 Related Work

Recently, there have been a number of developments in attempting to formally define the boundary between the ontology and the lexicon. These have been characterised as complementary resources (Buitelaar 2010), as the ontology forms a shared conceptualisation (Gruber 1995) and the lexicon describes the lexical encoding of that conceptualisation in words (Prévot et al. 2010). In recent years, there has been significant development in the creation and application of ontologies and more recently in the context of the ontology-lexicon interface.

The advent of the Semantic Web has led to a large degree of agreement in the representation of ontologies, in particular in the form of the OWL (Web Ontology Language) (McGuinness and Van Harmelen 2004) model, which is based on description logics (Horrocks et al. 2003). The fact that these ontologies can be represented and linked on the Web has led to an explosion of available semantic data, in particular in the form of large-scale resources, such as DBpedia (Bizer et al. 2009).

There have been several attempts to apply ontological principles to linguistic data: for example, *OntoWordNet* (Gangemi et al. 2003) aimed to take the existing information in *Princeton WordNet* (Fellbaum 2010) and extend it with ontological

information while fixing ontological errors in the modelling of WordNet. Similarly, the General Ontology of Linguistic Description (Farrar and Langendoen 2003, GOLD) aims to capture and represent formal linguistic concepts using an OWL ontology, based on studies in typology and field linguistics. The Lexical Markup Framework (LMF) (Francopoulo et al. 2006) is an ISO-standard model for the representation of lexica, which originated from a number of previous efforts in the harmonisation of dictionaries in disparate formats and with differing terminology. In spite of providing a common XML format, LMF does not establish interoperability between different lexica, as it does not introduce data categories (Ide and Romary 2006), that is, agreed-upon terms referring to clearly defined linguistic concepts. ISocat, a repository of so-called data categories, has however been set up to provide a common vocabulary of linguistic description, thus fostering interoperability between different resources. As many categories do not have a simple link, this resource has been further extended by means of relational links between concepts (Windhouwer 2012).

There have been a number of recent attempts to create models for describing the ontology-lexicon interface. LexInfo (Cimiano et al. 2011; Buitelaar et al. 2009) was proposed as a model that unifies the Lexical Markup Framework with the OWL ontology model. A complementary model, called the Linguistic Information Repository (Montiel-Ponsoda et al. 2008), was proposed, and the combination of these two models led to the *lemon* model used in this paper. The Ontology and Terminological Resources (Reymonet et al. 2007, OTR) meta-model is a similar model that focused on the use of terminological resources and in grounding the terms to instances where they are used in texts. *Senso Comune* (Oltremari et al. 2010), a dictionary of Italian terms, employed a similar model to the one discussed, except that instead of using a domain ontology, the meaning of terms was grounded relative to the top-level ontology DOLCE (Gangemi et al. 2002). A similar attempt to organise unique identifiers for words and link them to ontologies exists as part of the LexVo.org project (De Melo and Weikum 2008), which includes sense and string links to ontologies such as DBpedia. Our approach to representing the syntax-semantics interface is to provide minimal generalisable constructs so that the model can describe or be extended to a number of formalisms, such as to the Generative Lexicon (Pustejovsky 1991) as in Khan et al. (2013). Similarly, these patterns are applicable alongside other functional theories, such as underspecification (Egg et al. 2001) or by means of joint constraint resolution on both the syntactic and semantic constraints (Debusmann et al. 2004).

Ontology design patterns, first introduced by Gangemi and Presutti (2009), have been shown to be a useful part of the ontology design process (Presutti et al. 2009), and the use of patterns for mapping of linguistic structures to ontological predicates has been used to construct ontologies (Buitelaar et al. 2004). As we consider the case of patterns for models such as *lemon*, we require new kinds of patterns that model the ontology-based lexical semantics of lexical entries. There have been a number of endeavours to provide ontology-conform logical representations of the meaning of words such as by the Mikrokosmos project (Nirenburg et al. 1996). Raskin and Nirenburg (1995) have in particular discussed in depth how to model

the ontology-based semantics of adjectives beyond simple frames or predicates. A more recent project (Lefrançois and Gandon 2011, ULiS) has also attempted to provide a complete description of the lexicon and ontology based on meaning-text theory (Mel'cuk 1981) in combination with OWL ontologies.

## 7 Conclusion

We have presented a method for developing lexica for ontologies represented in the Semantic Web by means of defining a set of design patterns representing the most common lexicalisations of labels found in ontologies. As such, this method presents a principled method for the quick development of lexica for any ontology on the Semantic Web, and the use of generic patterns allows these lexica to be ported to new languages. Our application of this methodology to DBpedia has shown that the patterns we identified correspond well to the most frequently used in practice. As future work, we aim to further extend this set of patterns and consider the integration of this within a complete ontology-based language resource development work flow.

## References

- Baader, F. (2003). *The description logic handbook: Theory, implementation, and applications*. Cambridge: Cambridge University Press.
- Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., et al. (2009). DBpedia-A crystallization point for the Web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3), 154–165.
- Bouillon, P., & Viegas, E. (1999). The description of adjectives for natural language processing: Theoretical and applied perspectives. In *Proceedings of Description des Adjectifs pour les Traitements Informatiques. Traitement Automatique des Langues Naturelles* (pp. 20–30).
- Buitelaar, P. (2010). Ontology-based semantic lexicons: Mapping between terms and object descriptions. In *Ontology and the Lexicon* (pp. 212–223). Cambridge: Cambridge University Press.
- Buitelaar, P., Cimiano, P., Haase, P., & Sintek, M. (2009). Towards linguistically grounded ontologies. In *The Semantic Web: Research and applications* (pp. 111–125). Berlin: Springer.
- Buitelaar, P., Olejnik, D., & Sintek, M. (2004). A Protégé plug-in for ontology extraction from text based on linguistic analysis. In *The Semantic Web: Research and applications* (pp. 31–44). Berlin: Springer.
- Cimiano, P., Buitelaar, P., McCrae, J., & Sintek, M. (2011). LexInfo: A declarative model for the lexicon-ontology interface. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(1), 29–51.
- Cimiano, P., McCrae, J., Buitelaar, P., & Montiel-Ponsoda, E. (2013). On the role of senses in the ontology-lexicon. In *New trends of research in ontologies and Lexical resources* (pp. 7–25). Heidelberg: Springer.
- Comrie, B. (1976). *Aspect: An introduction to the study of verbal aspect and related problems*. Cambridge: Cambridge University Press.
- Davidson, D. (1967). The logical form of action sentences. In N. Rescher (Ed.), *The logic of decision and action* (pp. 81–95). Pittsburgh: University of Pittsburgh Press.

- De Melo, G., & Weikum, G. (2008). Language as a foundation of the Semantic Web. In *Proceedings of the 7th International Semantic Web Conference (ISWC 2008)* (Vol. 401).
- Debusmann, R., Duchier, D., Koller, A., Kuhlmann, M., Smolka, G., & Thater, S. (2004). A relational syntax-semantics interface based on dependency grammar. In *Proceedings of the 20th International Conference on Computational Linguistics* (pp. 176–182).
- Egg, M., Koller, A., & Niehren, J. (2001). The constraint language for lambda structures. *Journal of Logic, Language and Information*, 10(4), 457–485.
- Farrar, S., & Langendoen, D. T. (2003). A linguistic ontology for the semantic web. *GLOT International*, 7(3), 97–100.
- Fellbaum, C. (2010). *WordNet*. Berlin: Springer.
- Forsberg, M., & Ranta, A. (2003). The BNF converter: A high-level tool for implementing well-behaved programming languages. In *NWPT'02 Proceedings, Proceedings of the Estonian Academy of Sciences* (pp. 1–16).
- Fowler, M., & Parsons, R. (2010). *Domain-specific languages*. Upper Saddle River: Addison-Wesley Professional.
- Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M., et al. (2006). Lexical markup framework (LMF). In *Proceedings of the Fifth International Conference on Language Resources and Evaluation* (pp. 233–236).
- Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., & Schneider, L. (2002). Sweetening ontologies with DOLCE. In *Knowledge engineering and knowledge management: Ontologies and the Semantic Web* (pp. 166–181). Berlin: Springer.
- Gangemi, A., Navigli, R., & Velardi, P. (2003). The OntoWordNet project: Extension and axiomatization of conceptual relations in WordNet. In *On the move to meaningful Internet systems 2003* (pp. 820–838). Berlin: Springer.
- Gangemi, A., & Presutti, V. (2009). Ontology design patterns. In *Handbook on ontologies* (pp. 221–243). Berlin: Springer.
- Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human Computer Studies*, 43(5), 907–928.
- Halle, B., & Ronald, G. (2001). *Business rules applied: Building better systems using the business rules approach*. New York: Wiley.
- Horrocks, I., Patel-Schneider, P. F., & Van Harmelen, F. (2003). From SHIQ and RDF to OWL: The making of a web ontology language. *Web Semantics: Science, Services and Agents on the World Wide Web*, 1(1), 7–26.
- Ide, N., & Romary, L. (2006). Representing linguistic corpora and their annotations. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation* (pp. 3205–3212).
- Khan, F., Frontini, F., Grata, R. D., Monachini, M., & Quochi, V. (2013). Generative lexicon theory and linguistic linked open data. In *Proceedings of the 6th International Conference on Generative Approaches to the Lexicon* (pp. 62–69).
- Lefrançois, M., & Gandon, F. (2011). ULiS: An expert system on linguistics to support multilingual management of interlingual knowledge bases. In *9th International Conference on Terminology and Artificial Intelligence* (p. 108).
- McCrae, J., Aguado-de Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gomez-Perez, A., et al. (2012a). Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 46(4), 701–719.
- McCrae, J., Montiel-Ponsoda, E., & Cimiano, P. (2012b). Collaborative semantic editing of linked data lexica. In *Proceedings of the 2012 International Conference on Language Resource and Evaluation* (pp. 2619–2625).
- McGuinness, D., & Van Harmelen, F. (2004). OWL web ontology language overview. *W3C Recommendation*, 10, 2004–03. <http://www.w3.org/TR/owl-features/>
- Mel'cuk, I. (1981). Meaning-text models: A recent trend in Soviet linguistics. *Annual Review of Anthropology*, 10, 27–62.
- Montiel-Ponsoda, E., Aguado de Cea, G., Gómez-Pérez, A., & Peters, W. (2008). Modelling multilinguality in ontologies. In *Proceedings of the 22nd International Conference on Computational Linguistics* (pp. 67–70).

- Nirenburg, S., Beale, S., Mahesh, K., Onyshkevych, B., Raskin, V., Viegas, E., et al. (1996). Lexicons in the Mikrokosmos project. In *Proceedings of the Society for Artificial Intelligence and Simulated Behavior Workshop on Multilinguality in the Lexicon* (pp. 26–33).
- Oltramari, A., Vetere, G., Lenzerini, M., Gangemi, A., & Guarino, N. (2010). Senso comune. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation* (pp. 3873–3877).
- Presutti, V., Daga, E., Gangemi, A., & Blomqvist, E. (2009). eXtreme design with content ontology design patterns. In *Workshop on Ontology Patterns* (p. 83).
- Prérot, L., Huang, C., Calzolari, N., Gangemi, A., Lenci, A., & Oltramari, A. (2010). Ontology and the lexicon: A multi-disciplinary perspective. In *Ontology and the Lexicon: A natural language processing perspective* (pp. 3–24). Cambridge: Cambridge University Press.
- Pustejovsky, J. (1991). The generative lexicon. *Computational Linguistics*, 17(4), 409–441.
- Raskin, V., & Nirenburg, S. (1995). Lexical semantics of adjectives. *New Mexico State University, Computing Research Laboratory Technical Report, MCCS-95-288*.
- Raskin, V., & Nirenburg, S. (1998). An applied ontological semantic microtheory of adjective meaning for natural language processing. *Machine Translation*, 13(2), 135–227.
- Reymonet, A., Thomas, J., & Aussenac-Gilles, N. (2007). Modelling ontological and terminological resources in OWL DL. In *Proceedings of OntoLex07 Workshop at the 6th International Semantic Web Conference*.
- Stassen, L. (2011). Comparative constructions. In M. Dryer & M. Haspelmath (Eds.), *The world atlas of language structures online* (Chap. 121). Available online at <http://wals.info/chapter/121>.
- Tappolet, J., & Bernstein, A. (2009). Applied temporal RDF: Efficient temporal querying of RDF data with SPARQL. In *The Semantic Web: Research and applications* (pp. 308–322). Berlin: Springer.
- Unger, C., Hieber, F., & Cimiano, P. (2010). Generating LTAG grammars from a lexicon-ontology interface. In *Proceedings of the 10th International Workshop on Tree Adjoining Grammars and Related Formalisms (TAG+10)* (pp. 61–68).
- Unger, C., McCrae, J., Walter, S., Winter, S., & Cimiano, P. (2013). A lemon lexicon for DBpedia. In *Proceedings of the NLP+DBpedia Workshop Co-located with the 12th International Semantic Web Conference*.
- Vendler, Z. (1957). Verbs and times. In *The philosophical review*, 66(2), 143–160.
- Wampler, D., & Payne, A. (2008). *Programming Scala* (Chap. 11). Sebastopol: O'Reilly.
- Windhouwer, M. (2012). RELcat: A relation registry for ISOcat data categories. In *Proceedings of the Eight International Conference on Language Resources and Evaluation* (pp. 3661–3664).

# Context and Terminology in the Multilingual Semantic Web

Pilar León-Araúz and Pamela Faber

**Abstract** One of the main challenges of the Multilingual Semantic Web (MSW) is ontology localization. This first needs a representation framework that allows for the inclusion of different syntactic, lexical, conceptual and semantic features, but it also needs to account for dynamism and context from both a monolingual and multilingual perspective. We understand dynamism as the changing nature of both concepts and terms due to contextual constraints, whereas context is defined by the different pragmatic factors that modulate such dynamism (e.g. specialized domains, cultures, communicative situations). Context is thus an important construct when describing the concepts and terms of any domain in monolingual resources. However, in multilingual resources, context also affects interlingual correspondences. When dealing with multilingual ontologies, context features must be extended to include translation relations and degrees of equivalence.

**Key Words** Concept dynamics • Context • Term variants • Terminology • Translation relations

## 1 Introduction

Ontology localization has been identified as one of the main challenges of the Multilingual Semantic Web (MSW) (Espinoza et al. 2009; Gracia et al. 2012). It has been defined as “the process of adapting a given ontology to the needs of a certain community, which can be characterized by a common language, a common culture or a certain geo-political environment” (Cimiano et al. 2010). This adaptation first needs a representation framework that allows for the inclusion of different syntactic, lexical, conceptual and semantic features, but it also needs to account for dynamism and context, which happen to influence all of these features at different levels. We understand dynamism as the changing nature of both concepts and terms due to contextual constraints, whereas context is defined by the different

---

P. León-Araúz (✉) • P. Faber

Department of Translation and Interpreting, University of Granada, Buensuceso, 11 18002 Granada, Spain

e-mail: [pleon@ugr.es](mailto:pleon@ugr.es); [pfaber@ugr.es](mailto:pfaber@ugr.es)

© Springer-Verlag Berlin Heidelberg 2014

P. Buitelaar, P. Cimiano (eds.), *Towards the Multilingual Semantic Web*,

DOI 10.1007/978-3-662-43585-4\_3

pragmatic factors that modulate such dynamism (e.g. specialized domains, cultures, communicative situations). As a consequence of their natural dynamism, concepts may be recategorized and have their relational behaviour constrained, whereas terms may show several types of variants with different cognitive, semantic and usage consequences. Context is thus an important construct when describing the concepts and terms of any domain in monolingual resources. However, in multilingual resources, context also affects interlingual correspondences. When dealing with multilingual ontologies, context features must be extended to include translation relations and degrees of equivalence. As a result, a believable and useful knowledge representation needs to account for and classify context types as well as the result they may cause.

The remainder of this chapter is structured as follows: Sect. 2 introduces the areas where multilingual terminology analysis can contribute to the MSW, especially within the *lemon* (McCrae et al. 2010) framework, mainly based on the description of the contextual features related to concept and term dynamics as well as translation relations; Sect. 3 shows how context is composed of linguistic, conceptual and pragmatic facets without considering multilingualism; Sect. 4 gives an overview of the problems that may arise when establishing cross-lingual correspondences and how, as a result, different translation relations may apply.

## 2 Terminology and the MSW

Traditionally, Terminology has dealt with the description and/or standardization of the concepts and terms of a given specialized domain as well as their relations. Recently, it has also evolved towards the development of certain standard vocabularies and formats for data interoperability in combination with ontologies. Thus, the link between Terminology and knowledge representation is more than obvious. In fact, its relevance in the field has been widely acknowledged (Buitelaar et al. 2011), especially with the advent of multilingual ontologies. Nevertheless, most terminological resources are published in application-specific formats and are difficult to access (McCrae et al. 2012).

As emphasized by Fu et al. (2010), the promise of the Semantic Web is that of a new way to organize, present and search information that is based on meaning and not just text. Ideally, this would imply that language-independent knowledge could be accessible across different natural languages, which is how the MSW is envisioned. However, the Semantic Web is still essentially monolingual, and there is a growing need for multilingual resources that help overcome communication barriers. This entails creating more high-quality multilingual resources and providing ways to link and share them. EuroWordNet was a good attempt to enhance both multilinguality and interoperability. Each language module represents an autonomous and unique language-specific system of language-internal relations between synsets, which are connected through the Inter-Lingual-Index (ILI) (Vossen 2004). In this way, language-specific synsets linked to the same ILI-record are

considered equivalents across languages. However, EuroWordNet is not suitable for specialized knowledge, since synset members acquire different meanings in specialized contexts (León Araúz et al. 2012).

There have been several other initiatives allowing for the sharing of multilingual specialized knowledge, such as the Simple Knowledge Organization System (SKOS) and *lemon*, among many others. Even though SKOS was not specifically conceived for multilingual purposes, it has been widely used for semantic interoperability among different multilingual terminological resources, such as General Multilingual Environmental Thesaurus (GEMET). Since SKOS is concept oriented, correspondences are established through conceptual mappings based on the relations *skos:closeMatch*, *skos:exactMatch*, *skos:broadMatch*, *skos:narrowMatch* and *skos:relatedMatch*. As for terms and variants, SKOS proposes *skos:prefLabel*, *skos:altLabel* and *skos:hiddenLabel*. However, these relations, though useful, are not sufficient to capture the complexity of interlinguistic correspondence. For example, a *skos:prefLabel* in one language will not necessarily correspond to the *skos:prefLabel* in another language. SKOS only aims at establishing conceptual correspondences across different resources through binary mappings and taxonomies. In its current form, it can be very useful within the Linked Data initiative (Berners-Lee 2006) for certain purposes, but it might not be the best way to deal with the intricacies of multilinguality. In this line, Leroi and Holland (2010) propose a set of guidelines to enable multilinguality in SKOS through the mapping of both concepts and terms. They state that equivalences in a multilingual context can be of three kinds: semantic, cultural and structural. Semantic equivalence refers to the meaning of the concept, cultural equivalence refers to the use of a term in a given language or culture and structural equivalence refers to the semantic relations between concepts. Nonetheless, in the following sections, we show how this classification can be extended.

Knowledge, as regarded in Terminology, is something more complex than a thesaurus-like structure. In this sense, ontologies are better suited for accounting for multilinguality and contextual constraints. However, they are often considered multilingual when the concepts are accompanied by a simple *rdf:label*, even though cross-lingual differences have led to the awareness of dynamic conceptualizations. According to Cimiano et al. (2010), while the translation of labels is an important aspect in ontology localization, conceptualizations may also need to be adapted to different cultural or geopolitical contexts, as was attempted in EuroWordNet. In fact, it has been criticized that the pivotal role of English often leads to the translation of labels instead of proper localizations (Declerck and Gromann 2012). Furthermore, terminological resources provide more information than only *rdfs:labels*, but it is often lost in the final representation because of the required univocity of each label (Declerck and Gromann 2012). All these limitations have led researchers to propose the inclusion of terminological and linguistic information in different ontological modules.

In this sense, *lemon* is an RDF ontology-lexicon model that defines specific modules for different types of linguistic and terminological descriptions that are separate from the ontology. Apart from the subsumption relation in the ontology,



*lemon* represents the relation between two senses with the property *senseRelation* and enriches the representation with the subproperties of *equivalent*, *incompatible*, *broader* or *narrower*, which are similar to those in SKOS. The *lexicalSense* class also provides the restrictions (*usage*, *context*, *register*) that make a certain lexical entry appropriate for naming a certain concept in the specific context of the lexicalized ontology (Aguado de Cea and Montiel-Ponsoda 2012). In the lexical module, *lemon* covers the pragmatic preference of terms with the subproperties *prefSem*, *altSem* and *hiddenSem*. *lexicalVariant* relates different lexical entries (acronym, full form, etc.), and *formVariant* relates different forms of the same lexical entries (McCrae et al. 2010). Polysemy and synonymy are thus considered at the sense level and not only at the lexical level. The dynamics of terms and variants have been also included in the classes *CanonicalForms* and *PreferredLexicalizations*, where syntactic variants of the *LexicalForm* are differentiated.

Therefore, *lemon* covers the dynamics of both concepts and terms through different properties related to the sense and lexical levels depending on context. Furthermore, it allows to model contextual conditions with two properties: *context* and *condition*. *Context* constrains the domains under which the interpretation of the lexical entry as the concept in question is permissible, whereas *condition* is used to describe the circumstances that need to be fulfilled so that the lexical entry can be interpreted as the ontological concept in question (Cimiano et al. 2012). However, context should be made more explicit and be modelled according to a set of more concrete criteria, whereas preferred lexicalizations and canonical forms should be framed against these types of contexts and conditions, since they are interdependent. Consequently, *lemon* also has an extension that enables the representation of translations in a separate linguistic layer, thus leaving the original ontologies or data sources separate (Montiel-Ponsoda et al. 2011). Translations are regarded as variants and are linked through the property *isTranslationOf* and the translation relation types *isDescriptiveTranslationOf* and *isCulturalEquivalentTranslationOf*. Nevertheless, these distinctions may also depend on the lexical and sense modules and, thus, on context. As a result, translation relation types can be extended accordingly. In fact, the *lemon* developers acknowledge that further specifications of the translation relation would contribute to a true MSW (Montiel-Ponsoda et al. 2011).

Terminological resources are often designed as a support for human translators, who must then consider contextual factors. Thus, before systematizing translation correspondences, these factors should be modelled in regard to each language. The *lemon* developers also acknowledge that, although *lemon* includes features for the assignment of words to a pragmatic context, there is the need to define a pragmatic taxonomy (Montiel-Ponsoda et al. 2010). The following sections propose a set of pragmatic constraints that should be accounted for from both a monolingual and multilingual perspective.

### 3 Monolingual Dynamics: The Role of Context

Context is generally regarded as the parts of a written or spoken statement that precede or follow a specific word or phrase and which can influence its meaning or effect. It is also the situation, events or information that are related to something and which help a person to understand it. Context can have a wider or narrower scope and can include external factors (situational and cultural) as well as internal cognitive factors, all of which interact with each other (House 2006). In many cases, context is the only factor that can be used for word sense disambiguation, and it also influences the choice of a word form over its variant. In computing, the idea that contextual information is important is not new. Proposals for the incorporation of this type of information have been made to enhance similarity measurements of data-mining results (Singh and Vajirkar 2003) or to make them more context sensitive (Dong et al. 2010). The key to success lies in parametrizing contexts so that the system can be aware of situational meaning constraints. If this is done for each language separately, cross-lingual mappings would be enhanced.

#### 3.1 Term Dynamics

Although Terminology initially aspired to having one linguistic designation for each concept for greater precision, it is true that the same concept can often have many different types of linguistic designations according to context. There are certain types of variation that are often used with no significant impact on communication, such as morphological variants, orthographic variants, ellipted variants, abbreviations, graphical variation, variation by permutation, etc. (Bowker and Hawkins 2006). However, terminological variation often occurs for considerably more complex reasons. Freixa (2006) classifies the causes for variation in the following categories: (1) dialectal, based on origin; (2) functional, based on register; (3) discursive, based on style; (4) interlinguistic, based on the contact between languages; and (5) cognitive, based on different conceptualizations. Variants of the last type involve a change in semantics, as they give a particular vision of the concept. In this line, Montiel-Ponsoda et al. (2012) and Aguado de Cea and Montiel-Ponsoda (2012) propose a model based on how variants affect ontology semantics: (1) variants that are semantically the same, but formally different; (2) variants that are semantically and formally different, but still refer to the same concept; and (3) variants that are totally different and point to two related, but different, concepts.

Based on the above-mentioned approaches, our experience and other foundational work on term variation (Daille 2005; Fernández-Silva et al. 2011), we propose the following extended classification, since all of its types may affect semantics, pragmatics and—in a later stage—linguistic interlingual correspondences:

- Orthographic variants (with no geographic origin, e.g. *aesthetics*, *esthetics*). They do not affect semantics or the communicative situation.

- Diatopic variants
  - Orthographic variants (e.g. *groyne*, *groin*). They do not affect semantics.
  - Dialectal variants (e.g. *gasoline*, *petrol*). They may affect semantics if culture-bound factors highlight or suppress any of the semantic features (see Sect. 3.3.2).
  - Culture-specific variants (e.g. *sabkha*, *dry lake*). They affect both semantics and the communicative situation when referring to a particular entity that, in a specific culture, adds more specific features (see Sect. 3.3.2).
  - Calques. They may affect semantics and the communicative situation and are the result of an interlinguistic borrowing for different reasons, such as the influence of a particular language on a specialized domain.
- Short-form variants. They do not affect semantics but only the communicative situation.
  - Abbreviation
  - Acronym (e.g. *laser*, *light amplification by stimulated emission of radiation*).
- Diaphasic variants
  - Science-based variants. They do not affect semantics but only the communicative situation.
    - Scientific names (e.g. *Dracaena draco*, *drago*). They refer to specialized nomenclatures and are especially useful in botany, zoology, chemistry, etc.
    - Expert neutral variants (e.g. *ocellaris clownfish*, *Amphiprion ocellaris*). They would be the default term choice in a specialized scenario.
    - Jargon. Sometimes experts have their own informal way of referring to specialized concepts (e.g. in medicine, *lap-appy* would correspond to *laparoscopic appendectomy*, but no lay user would use this term).
    - Formulas (e.g. H<sub>2</sub>O, *water*; CaCO<sub>3</sub>, *pearl*). They do not affect semantics but only the communicative situation.
    - Symbols (e.g. \$, *dollar*).
  - Informal variants. They do not necessarily affect semantics but especially the communicative situation.
    - Lay-user variants (e.g. *dragon tree*, *drago*). They would be the default term choice in non-specialized scenarios.
    - Colloquial variant (e.g. *fracking*, *hydraulic fracturing*).
    - Generic variants (e.g. *sea*, *ocean*; *erosion*, *weathering*). Very informal variants can activate terms pointing to different levels of conceptual granularity and thus affecting semantics.
  - Domain-based variants (e.g. *sludge*, *mud*). They may affect semantics and/or the communicative situation (see Sect. 3.3.1) when term preferences change across specialized domains.

- Dimensional variants (e.g. *Gutenberg's discontinuity, core-mantle boundary*). They are usually multi-word terms and affect semantics, since they convey different dimensions of the same concept (the person who first named it and the two parts it delimits).
- Metonymic variants (e.g. *escollera, espigón*). They affect semantics because the metonymic variant designates the concept according to its parts.
- Diachronic variants. They only reflect historical term usages.
- Non-recommended variants (e.g. in medicine, *mental retardation* now has negative connotations and has been substituted by *intellectual disability*).
- Morphosyntactic variants (e.g. *the action of the waves, wave action*). They do not affect semantics but depend on collocates, term selection preferences and the communicative situation.

The nature and scope of these variants are very diverse and may have different consequences in communication. Nevertheless, terms can activate more than one variant type, which might make term choice more difficult. For example, H<sub>2</sub>O and/or *water* may be domain-based variants since the first is more frequently used in chemistry and water treatment domains than in oceanography. However, their use also depends on the communicative situation (i.e. formal or informal). On the contrary, the same type of variant can be expressed by more than one term. Diaphasic variants, in particular, form a continuum from more formal to informal (e.g. *thermal low-pressure system, thermal low, thermal trough and heat low*).

### 3.2 *Concept Dynamics*

Conceptual contexts in texts involve the conceptual relations activated by the words in a relatively short span before and after the keyword/phrase. In Terminology, this affects opaque noun compounds, which can be more difficult to process. For instance, when *sediment* is the head word, in N+N compounds the slot activated is usually < *location* > (e.g. *intertidal zone sediment, streambed sediment, aquifer sediment*, etc.), whereas in A+N compounds the < *material* > slot is opened up (*lithogenous sediment, biogenous sediment, hydrogenous sediment, cosmogenous sediment*). The analysis of heads and slots can contribute to the extraction of hyponyms. Nevertheless, it can also be useful in the study of dimensional variants that show the dynamics of concepts, where synonyms designate the same concept but add or suppress semantic slots (e.g. *Gutenberg's discontinuity, core-mantle boundary*). Thus, multi-word terms, whose formation is dynamic by nature, show that concepts may be classified according to multidimensional facets (*location, material*, etc.) and can be a rich source for semantic feature modelling. However, semantic features should not always be stable in representations.

In Terminology, multidimensionality (Rogers 2004) is often regarded as a way of enriching traditional static representations. However, not all dimensions are always part of a unique conceptualization, since their activation is context dependent

(León Araúz et al. 2013). Because of multidimensionality, a given concept may have two hyperonyms in the same domain according to the well-known phenomenon of multiple inheritance. However, contextual multidimensionality can also be a source of non-monotonic inheritance, because shared properties are incompatible or because their activation depends on perspective (e.g. SAND<sup>1</sup> as a *type\_of* SEDIMENT, as a *type\_of* BEACH FILL, as a *type\_of* MORTAR MATERIAL, as a *type\_of* SOIL or as a *type\_of* PROPPANT). In these cases, although *sand* is a term that designates the same referent, the concept's relational behaviour is constrained depending on its hypernym. As a result, when it is a MORTAR MATERIAL, it will not be related to the same concepts as when it is a type of BEACH FILL.

In *lemon*, semantics are stored both in the ontology and in the lexicon. In this way, other possible (less prototypical) hyperonyms are stored in the lexicon module expressed as *narrower* (e.g. student as a type of person could pose certain problems in the ontology but must be somehow represented) (McCrae et al. 2010). However, certain constraints should be imposed on the context when this narrower relation should be activated or not, as no specialized knowledge concept can be activated in isolation but rather as part of an event where perception, culture and many other dynamic factors may trigger different conceptualizations.

### 3.3 Pragmatic Dynamics

As follows from the previous sections, pragmatics is at the core of the dynamics of terms and concepts, since changes in conceptualization and in the lexicon are clearly not independent from each other but interact in a number of unforeseeable ways (Cimiano et al. 2010). Precisely for that reason, the context of concepts and terms should be described according to how they change across disciplines, cultures, etc., as well as the fuzzy category boundaries they establish. Domain and culture-based constraints are an example of how context can emerge.

#### 3.3.1 Domain-Based Constraints

According to Picht and Draskau (1985), multidimensionality depends on who is the classifier as well as the different knowledge sources that may reflect different criteria when organizing the same domain. This kind of dynamism stems from the fact that various disciplines deal with concepts in a different way and use different sets of terms to designate them. In EcoLexicon,<sup>2</sup> a multilingual terminological knowledge base on the environmental domain, domain-based multidimensionality has produced

---

<sup>1</sup>In order to differentiate terms and concepts, we use small capitals for concepts and italics for terms.

<sup>2</sup><http://ecolexicon.ugr.es>.

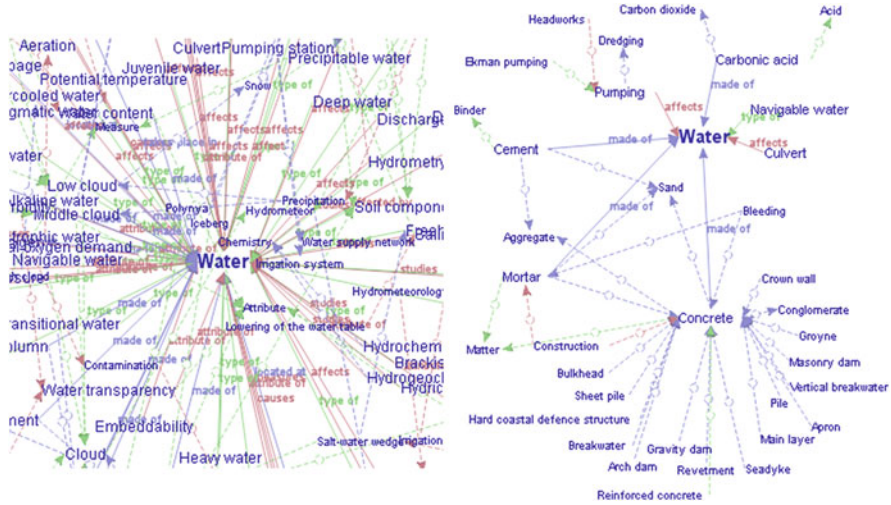


Fig. 1 Domain-free and domain-based semantic network of WATER

an information overload, especially in the conceptualizations of top-level concepts, such as WATER (left side of Fig. 1).

WATER can certainly be related to all of these concepts; however, it rarely, if ever, activates all of them at the same time, since this would evoke completely different and incompatible scenarios (León Araúz et al. 2013). Our claim is that any specialized domain contains subdomains in which conceptual dimensions become more or less salient, depending on the activation of specific contexts. The area of environmental knowledge was thus divided into a set of domain-based contexts (e.g. HYDROLOGY, GEOGRAPHY, OCEANOGRAPHY, CIVIL ENGINEERING, ENVIRONMENTAL ENGINEERING, etc.), and the relational power of concepts was constrained accordingly. Thus, when constraints are applied, the network of WATER within the CIVIL ENGINEERING domain is recontextualized and becomes more meaningful (right side of Fig. 1). This is not only important for representations but also because the activation of specific domain-based semantic features constrains the potential meaning of the terms. Furthermore, this may give rise to domain-based term variation that may result in non-semantic variants (e.g. *sludge*, *mud*) or dimensional variants, where term preferences usually point to multi-word terms that highlight particular semantic features according to the most important facets in the domain [e.g. *beach sediment* (material), *beach fill* (function)].

### 3.3.2 Culture-Based Constraints

One might think that natural landforms are more or less the same all over the world. Until recently, it was believed that natural entities, such as MOUNTAIN,

were universals. However, even within the same language, there are significant differences as to how scientific concepts are categorized. For instance, the concept WATERSHED in American English covers a whole river basin, whereas in British and Australian English it is more narrowly defined and only refers to the dividing line between two river systems. This means that within the whole of an American *watershed*, British and Australian scientists see several watersheds. *Drainage basin* and *catchment area* are other term variants that designate the American sense of *watershed*. They are sometimes used interchangeably and other times used as a hyperonym of WATERSHED. BOGS or FENS are usually grouped together and referred to as *mires* in Europe, but not in the USA. MARSHES in Europe are often called *reed swamps*, but SWAMPS in the USA are not dominated by reeds but by trees. *Carr* is the northern European way of referring to the Southeast American *wooden swamp*, which in the UK is also called *wet woodland*. There are also specific types of wetlands that are only predominant in certain geographic areas that are not lexicalized in all cultures, such as the Australian *billabong*, the African *dambo* or the Canadian *muskeg*. Thus, when one of these terms is activated in a text, the location-related category features of the concept are constrained.

Artificial geographic objects are also susceptible to cultural variation as much as natural geographic objects. For instance, the concept PIER is often designated as *jetty* in the Great Lakes, while a JETTY is generally a structure designed to prevent the shoaling of a channel and not a recreational area. However, in British English, *jetty* is the synonym of a *wharf*. In contrast, in American English, *pier* may also be a synonym of *dock*. Nevertheless, in British English a DOCK is the area of water used for loading or unloading cargo in a harbour, which in American English is called a *port*. Geographical variation in this category domain is often conceptually motivated and mainly based on the dimensions of location and function. For instance, a DIKE may be called a *levee* when it is located on a river, whereas a BREAKWATER may be called a *mole* when it is covered by a roadway. On the contrary, when a BREAKWATER serves as a PIER, it is called a *quay* in British English and a *wharf* in American English.

## 4 Multilingual Dynamics: The Role of Equivalence

At the heart of any discussion of translation is the issue of correspondence or equivalence. Although a great deal has been written about translation equivalence, much of it is repetitive and not very useful for systematization purposes. Apart from the traditional opposition of faithful/free, other pairs of terms such as semantic/communicative (Newmark 1981) and formal/dynamic (Nida and Taber 1969) have also been proposed as descriptions for the degree of perceived similarity at the level of form and/or function between a source text and a target text. However, it is to be lamented that these changes of label have not been accompanied by significant new insights into the nature of equivalence.

## 4.1 Cross-Linguistic Problems

Part of this complexity is due to the fact that the rules do not remain the same, but change with each new translation context. Espinoza et al. (2009) highlight three translation contexts in ontology localization: (1) existence of an exact equivalent, (2) existence of several context-dependent equivalents and (3) existence of a conceptualization mismatch. They state that the first situation is specific to specialized and engineering fields. However, this does not seem to be the case in the environmental domain. As for the second and third types, we believe that the boundaries between different types of translation problem are somewhat less clear-cut. In our view, when dealing with cross-lingual meaning and vagueness, even in specialized domains, the following problems arise in regard to both concept and term dynamics:

1. The entity exists in both cultures, but the term for it in one language culture is more general or more specific (e.g. *shingle* in English is a term that covers several more specific terms in Spanish).
2. The entity exists in both cultures, but only one language culture has a term for it. The other has not regarded it as sufficiently salient to name (e.g. *river* and the French *fleuve* and *rivière* and *espigón* and *jetty* and *groyne*).
3. The entity exists in both cultures yet the terms are not exact correspondents because they highlight different aspects of the concept or focus on it from different perspectives (e.g. the French *fleuve* and the English *main stem*).
4. The entity exists in both cultures, and both language cultures have terms for it, but only in one language the concept has been lexicalized in several variants with different communicative or conceptual consequences (e.g. the Spanish *intestinos* and the English synonyms *intestines* and *bowels* or *rubble-mound* *breakwater* and the Spanish synonyms *dique de escollera* and *dique en talud*) (see next page).
5. The entity exists in both cultures, and both language cultures have terms for it, which approximately correspond. However, the lexical categories appear to have different structures in each culture and thus seem to operate on different design principles (e.g. *dock*, *quay* and *wharf*, and the Spanish *muelle*, *embarcadero* and *dársena*).
6. The entity exists in both cultures, but its cultural role (utility, affordances and hindrances) in each one is different. This leads to a conceptual mismatch and lack of correspondence (e.g. *pier* and *embarcadero*).
7. The entity exists in only one of the cultures, but its name has been adopted in the other culture to refer only to the foreign culture-specific concept (e.g. the Australian *billabong*, the African *dambo* or the Canadian *muskeg*).
8. The entity exists in both cultures, but one culture has recycled a term from the other culture to refer to another totally different concept (e.g. *playa* in West USA as *dry lake* and not as the usual Spanish equivalent *beach*, but *salar*).
9. The entity exists in only one of the cultures and is totally unknown in the other without any designation (e.g. *pejerrey*, a fish that only can be found in South America).



10. The entity exists in both cultures, but one of the cultures may refer to it with a metonymic designation and be ambiguous (e.g. *groyne* as the equivalent of the Spanish *escollera*, the material it is usually made of).

In order to define translation strategies that successfully address these problems, all the previous senses of context must be considered and interrelated. According to Montiel-Ponsoda et al. (2011), when there are several terms in each language, it is desirable to unambiguously express which term variant in language A is the translation of which term variant in language B. At this point, translation relations acquire significance. Nevertheless, even when all possible contextual constraints of both source and target terms and concepts are defined, this still does not establish 1:1 correspondences. Instead, a wide range of interrelated variables must still be considered.

For example, if a concept is designated by an informal term variant, it should not always be translated by its informal counterpart in another language and vice versa because one must also consider the nature of the communicative situation. Furthermore, pragmatic conventions can also change from culture to culture and might be even more important than semantics. For instance, even if a term pair such as *intestinos* and *intestines* are full equivalents, *bowels* would be more appropriate in an English doctor–patient situation. In this line, Cimiano et al. (2010) state that unintended shifts in meaning may occur when the term chosen as a translation equivalent has different connotations in the target community. As previously mentioned, multidimensionality has an impact not only on how concepts are classified but also on how term variants emerge. This may thus impair translation equivalence. In Spanish, there are two ways to designate the concept RUBBLE-MOUND BREAKWATER: *dique de escollera* or *dique en talud*. *Dique de escollera* would be the direct semantic equivalent of the English term *rubble-mound breakwater*, because both of them focus on the material dimension (*escollera*, *rubble-mound*), whereas *dique en talud* focuses on the place where it is located (on a slope). Since all rubble-mound breakwaters are built on a slope, two conceptualizations are possible, but only in Spanish do they emerge as lexicalized term variants. However, even if *rubble-mound* and *escollera* are equivalents in Spanish, *dique en talud* is the most frequently used term. Thus, unless there are certain contextual constraints pointing to the material these structures are made of, *dique en talud* would be the most reliable translation even if it is the less intuitive choice at first sight.

Another important factor in translation is directionality, since translation relations are not necessarily symmetric. For instance, when translating from French into English, both *fleuve* and *rivière* can be translated as *river*, but *fleuve* and *rivière* are not interchangeable when translating from English into French. Furthermore, translation pragmatics also imposes certain constraints on symmetry. The concepts PRIME MINISTER in the British political system and PRESIDENTE DEL GOBIERNO in the Spanish political system are not exact equivalents, but can be considered the closest cultural equivalents. However, while *Spanish prime minister* is the usual

translation of the Spanish *presidente del gobierno*, *prime minister* should never be translated as *presidente del gobierno británico*, since a PRESIDENT is usually a head of state.

## 4.2 Translation Relations

As previously stated, Montiel-Ponsoda et al. (2011) propose representing in *lemon* two translation relations (i.e. descriptive and cultural translations). However, based on our experience and a selected combination of the equivalence strategies proposed by translation studies scholars (Newmark 1981; Nida and Taber 1969; Nord 1997), we believe that a more extensive classification should be devised. The following translation relations would address the problems discussed in Sect. 4.1:

- Canonical translations apply when no equivalence problems arise and the translation relation may be symmetric. *River* and *río* would be canonical symmetric equivalents, but this does not mean that when canonical translations are found, no other relations are possible, since context can impair the degree of equivalence.
- Generic-specific translations would address problems 1, 2 and 3—which are related to cross-lingual categorization differences—, depending on the communicative situation and directionality. A specific-generic translation would apply when translating the term *shingle*, which in Spanish can be translated by its hypernym *material de grano grueso* (*coarse material*). In the same way, when the context describes a beach nourishment scenario in Spanish, the term *material de grano grueso* can be translated by its canonical form *coarse material* but also by its specific translation *shingle*. Alternatively, the following relation may apply.
- Extensional translation would address problems 1 and 2 and is a kind of generic-specific translation, because the original term is translated by all of the hyponyms of the concept in the target culture. In this way, *shingle* can also be translated by the enumeration of its subtypes (*arena y grava*).
- Communicative translations would address problem 4 establishing register correspondence among domain-specific and diaphasic variants. The canonical translation of *lodo* is usually *mud*, but in a water treatment domain, experts have a preferred designation: *sludge*. Furthermore, depending on the communicative situation, certain terms can be translated as the expert neutral variant or the lay-user variant in the target language (e.g. *intestines* or *bowels* for *intestinos*).
- Functional translations would address problems 5, 6 and 7 and involve deculturalizing original terms, so that receivers can relate to the concept. *Muskeg* can be translated as *turbera* and *malecón* as *seawall*. These equivalents lose their cultural traits but are the closest concepts in target cultures from a semantic point of view. Other terms, such as *quay*, *dock* and *wharf*, must rely on additional contextual features, since they can all be translated as *muelle*, *embarcadero* and/or *dársena* depending on the size, function and position of the structures. This relation is particularly asymmetric. For instance, *turbera* could hardly ever

be translated as *muskeg*, since unless the communicative situation points to this particular type of Canadian wetland, the canonical translation *bog* would apply in most of the cases.

- Cultural translations apply when cross-cultural differences impair the translation process and affect both concepts and terms. There would be another way of addressing problems 6, 7 and 8 that consists of adapting original culture-bound terms to other culture-bound terms in the target culture. The usual canonical translation of *pier* is *embarcadero*, but piers are often recreational areas that do not fit with the Spanish concept. In these cases, the most suitable translation would be *paseo marítimo* (literally *boardwalk*) or even *malecón* or *costanera* for South American Spanish, since even if these kinds of constructions are slightly different, the cultural component of the concept is preserved.
- Descriptive translations would also address culture-bound problems and make explicit certain semantic features according to user communication needs (problems 7 and 8) or in order to distinguish a concept that has not been termed in the target culture (problems 2, 9). For lay users, the term *muskeg* could be translated as *humedal canadiense muskeg* (Canadian wetland muskeg), adding and highlighting its hypernym and location. In contrast, the term *jetty* can be translated as *espigón*, which is the canonical translation of *groyne* or even *dique*, which would be a functional translation according to its general nature and the wide array of functions it may have. However, if both terms are found in a text (*jetty* and *groyne*), some distinction must be made. In this sense, a descriptive translation could be *espigón de encauzamiento*, which explains the particular function of jetties.
- Non-translations also address culture-bound problems (7, 9) when entities and/or lexicalizations do not exist in the target culture (*pejerrey*) but also in specialized communication. Terms like *muskeg* or *billabong* can be kept in their original form if the receivers are experts who do not need any description or contextualization.
- Metonymic translations would address problem 10 and apply when original terms are expressed in the form of a metonymic variant and target terms are not. *Groyne* could be translated both as *espigón* and *escollera* (metonymic variant), but *escollera*, in its coastal structure sense, can only be translated as *groyne*.

Translation term pairs are thus hardly ever symmetric and can be highly dynamic, since any term can be translated by many others when localization accounts for context and context includes terminological, conceptual and pragmatic factors. Furthermore, as pointed out by Hirst (2014), we cannot assume that translation equivalents have identical meanings.

Localizing ontologies is a powerful way of gaining multilingual resources. In this line, several approaches (McCrae et al. 2011; Assoja et al. 2012) have proposed the semi-automatic or automatic translation of labels and other natural language descriptions contained in ontologies, such as comments or definitions. An example of this is LabelTranslator (Espinoza et al. 2008), a system that localizes ontologies automatically. Its input is an ontology whose labels are expressed in a source natural language and obtains the most probable translation of each label into another target

natural language. It relies on translation web services, such as GoogleTranslate, and lexical resources combined with a ranking method based on the ontological context of each label. Another approach consists of generating multilingual ontologies based on existing multilingual resources (Gromann and Declerck 2014). This is in our view a more reliable way of getting better-quality results, which also highlights the important role of genuine multilingual resources. However, all of these approaches aim at identifying the single most appropriate translation for labels instead of storing the vast array of possibilities that may arise as different contextually based lexicalizations of the same concept. As stated by Gangemi (2012), when we envisage applications that are cross-linguistic, they need to work at the level of cognitive relevance, not at that of decontextualized data or term equivalences. Since ontology localization is usually a decontextualized process, all possible translation relations should be considered.

## 5 Conclusions

In this chapter, we have presented an approach to pragmatic constraints for the description of concepts and terms as the first step to establishing cross-lingual correspondences. For the Semantic Web to be truly multilingual, it is imperative to integrate context and equivalence dynamics in knowledge representation systems. However, both constructs have a myriad of interrelated variables to account for. Both concept and term dynamics are the result of diverse pragmatic factors, such as domain-based and culture-based constraints. Translation relations converge at the intersection of all monolingual variants in every language, and between descriptive and cultural translations, there is a vast array of possibilities.

**Acknowledgements** This research has been funded by project FFI2011-22397, from the Spanish Ministry of Science and Innovation.

## References

- Aguado de Cea, G., & Montiel-Ponsoda, E. (2012). Term variants in ontologies. In *Proceedings of the 30th Conference of AESLA*, Lleida, Spain.
- Assoja, K., Gracia, J., Aggarwal, N., & Gómez-Pérez, A. (2012). Using cross-lingual explicit semantic analysis for improving ontology translation. In *Proceedings of the Second Workshop on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid MT (COLING)*, Mumbai, India.
- Berners-Lee, T. (2006). Linked Data - Design Issues. Online.
- Bowker, L., & Hawkins, S. (2006). Variation in the organization of medical terms. Exploring some motivations for term choice. *Terminology*, 12(1), 79–110.
- Buitelaar, P., Cimiano, P., McCrae, J., Montiel-Ponsoda, E., & Declerck, T. (2011). Ontology lexicalisation: The lemon perspective. In *Proceedings of the 9th International Conference on Terminology and Artificial Intelligence*, Paris, France.

- Cimiano, P., McCrae, J., Buitelaar, P., & Montiel-Ponsoda, E. (2012). On the role of senses in the ontology-lexicon. In A. Oltramari, P. Vossen, L. Qin, & E. Hovy (Eds.), *New trends of research in ontologies and lexical resources*. New York: Springer.
- Cimiano, P., Montiel-Ponsoda, E., Buitelaar, P., Espinoza, M., & Gómez Pérez, A. (2010). A note on ontology localization. *Applied Ontology*, 5(2), 1–10.
- Daille, B. (2005). Variations and application-oriented terminology engineering. *Terminology*, 11(1), 181–197.
- Declerck, T., & Gromann, D. (2012). Combining three ways of conveying knowledge: Modularization of domain, terminological, and linguistic knowledge in ontologies. In CEUR (Ed.), *Proceedings of the 6th International Workshop on Modular Ontologies*, Graz, Austria (Vol. 875, pp. 28–40).
- Dong, H., Hussain, F., & Chang, E. (2010). A context-aware semantic similarity model for ontology environments. *Concurrency and Computation: Practice & Experience*, 23(5), 505–524.
- Espinoza, M., Gómez-Pérez, A., & Mena, E. (2008). Enriching an ontology with multilingual information. In *Proceedings of the 5th European Semantic Web Conference*, Tenerife, Spain.
- Espinoza, M., Montiel-Ponsoda, E., & Gómez-Pérez, A. (2009). Ontology localization. In A. Press (Ed.), *5th International Conference on Knowledge Capture*, Redondo Beach, USA.
- Fernández-Silva, S., Freixa, J., & Cabré, M. (2011). A proposed method for analysing the dynamics of cognition through term variation. *Terminology*, 17(1), 49–73.
- Freixa, J. (2006). Causes of denominative variation in terminology. A typology proposal. *Terminology*, 12(1), 51–77.
- Fu, B., Brennan, R., & O’Sullivan, D. (2010). Cross-lingual ontology mapping and its use on the multilingual semantic web. In *2nd Workshop on the Multilingual Semantic Web*, Bonn, Germany.
- Gangemi, A. (2012). Hybridizing formal and linguistic semantics for the multilingual semantic web. In *Proceedings of the 3rd Workshop on the Multilingual Semantic Web*, Boston, USA.
- Gracia, J., Montiel-Ponsoda, E., & Gómez Pérez, A. (2012). Cross-lingual linking on the multilingual web of data. In *Proceedings of the 3rd Workshop on the Multilingual Semantic Web*, Boston, USA.
- Gromann, D., & Declerck, T. (2014). Cross-lingual correcting and completive patterns for multilingual ontology labels. In P. Buitelaar & P. Cimiano (Eds.), *Towards the multilingual semantic web: Principles, methods and applications*. Berlin: Springer. doi:10.1007/978-3-662-43585-4.
- Hirst, G. (2014). Overcoming linguistics barriers to the multilingual semantic web. In P. Buitelaar & P. Cimiano (Eds.), *Towards the multilingual semantic web: Principles, methods and applications*. Berlin: Springer. doi:10.1007/978-3-662-43585-4.
- House, J. (2006). Text and context in translation. *Journal of Pragmatics*, 38, 338–358.
- León Araúz, P., Gómez-Romero, J., & Bobillo, F. (2012). A fuzzy ontology extension of wordnet and euwordnet for specialized knowledge. In *Proceedings of the 10th Terminology and Knowledge Engineering Conference*, Madrid, Spain (pp. 139–154).
- León Araúz, P., Reimerink, A., & Aragón, A. (2013). Dynamism and context in specialized knowledge. *Terminology*, 19(1), 31–61.
- Leroi, V., & Holland, J. (2010). Guidelines for mapping into SKOS, dealing with translations. Online. Deliverable D.7.2 for the ECP-2005-CULT-038099 project.
- McCrae, J., Aguado-de Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., et al. (2010). The lemon cookbook. Online.
- McCrae, J., Aguado de Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., et al. (2012). Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 46, 701–719.
- McCrae, J., Espinoza, M., Montiel-Ponsoda, E., Aguado-de Cea, G., & Cimiano, P. (2011). Combining statistical and semantic approaches to the translation of ontologies and taxonomies. In *Proceedings of the 5th Workshop on Syntax, Semantics and Structure in Statistical Translation*, Portland, USA (pp. 116–125).

- Montiel-Ponsoda, E., Aguado de Cea, G., & McCrae, J. (2012). Representing term variants in lemon. In *Proceedings of the 9th International Conference on Terminology and Artificial Intelligence*, Paris, France.
- Montiel-Ponsoda, E., Gracia, J., Aguado-de Cea, G., Buitelaar, P., Wunner, T., & Declerk, T. (2010). Multilingual ontologies for networked knowledge. Online. D2.1 Ontology-lexicon model. Final Deliverable for the FP7-ICT-4-248458 project.
- Montiel-Ponsoda, E., Gracia, J., Aguado de Cea, G., & Gómez-Pérez, A. (2011). Representing translations on the semantic web. In *Proceedings of the 2nd International Workshop on the Multilingual Semantic Web*, Bonn, Germany.
- Newmark, P. (1981). *Approaches to translation*. Oxford: Pergamon.
- Nida, E., & Taber, C. (1969). *The theory and practice of translation*. Leiden: EJ. Brill.
- Nord, C. (1997). *Translating as a purposeful activity. Functionalist Approaches explained*. Manchester, UK: St. Jerome.
- Picht, H., & Draskau, J. (1985). *Terminology: An introduction*. Guildford: University of Surrey.
- Rogers, M. (2004). Multidimensionality in concepts systems: A bilingual textual perspective. *Terminology*, 10(2), 215–240.
- Singh, S., & Vajirkar, P. (2003). Context-aware data mining using ontologies. In *Proceedings of the 22nd International Conference on Conceptual Modeling*, Japan (pp. 405–418).
- Vossen, P. (2004). EuroWordNet: A multilingual database of autonomous and language-specific wordnets connected via an interlingual index. *International Journal of Lexicography*, 17(2), 161–173.

# The Multilingual Semantic Web as Virtual Knowledge Commons: The Case of the Under-Resourced South African Languages

Laurette Pretorius

**Abstract** The participation of the under-resourced South African languages in the Multilingual Semantic Web as Virtual Knowledge Commons is imperative in terms of sharing in and contributing to the knowledge commons, in sustaining multilingualism and the technological development of these languages and in preserving cultural diversity and indigenous knowledge systems. This chapter takes a closer look at the challenges that the under-resourced languages of South Africa face in this regard and addresses two of these challenges. It is shown how three different types of high-quality language data, viz. multilingual terminology in English, Afrikaans, Tswana and Zulu; indigenous knowledge on astronomy nomenclature in Tswana; and a parallel corpus of English, Afrikaans, Tswana and Zulu could be exposed as Linked Data in a principled way. The conclusion contains various possibilities for future work.

**Key Words** Linked Data • Multilingual Semantic Web • Under-resourced languages • Virtual knowledge commons

## 1 Introduction

In the philosophy of science, much has been written on the importance of a democracy of knowledge, an emerging concept that addresses the relationships between knowledge production and dissemination, as well as the functions of the media and democratic institutions (In 't Veld 2010; Innerarity 2011; Visvanathan 2009). Indeed,

the confrontation of Science by traditional, indigenous, and local knowledges will result in more comprehensive dialogues on sustainable and peaceful development. The notion of *cognitive justice* offers us the option for a [diverse], pluralist and inclusive knowledge base from which we can draw our plans for building a better world (Van der Velden 2006).

---

L. Pretorius  
University of South Africa, PO Box 392, Pretoria 0003, South Africa  
e-mail: [pretol@unisa.ac.za](mailto:pretol@unisa.ac.za)

In this discourse, *diversity* refers to knowledge found in different cultures and languages, while *plurality* specifically concerns engagement across cultures and languages. Towards this end, we as humankind should, therefore, strive to create those conceptual spaces where it is possible to diminish the boundaries between the wide variety of locations of knowledge creation, forms of knowledge and uses of knowledge—a so-called *knowledge commons*.

We consider two perspectives on such a knowledge commons. Firstly, there is the conceptual level where we distinguish between knowledge that is of a general mainly culture- and language-independent nature and traditional, indigenous and local knowledge that are often culture and language specific. Secondly, there is the linguistic level where we are concerned with the multitude of languages as carriers of knowledge and the reality that significant amounts of knowledge are represented in so-called under-resourced languages. In other words, the conceptual level serves as interlingua for the representation and rendering of knowledge in and across different languages. For the knowledge commons to be comprehensive and inclusive, infrastructure and mechanisms have to be put in place to allow the speakers of all, even under-resourced, languages to participate fully in the knowledge commons.

For the first time in history, technology, such as the World Wide Web, offers the possibility to connect humans from all cultures and languages on a grand scale. It is indeed the vision of the Multilingual Semantic Web (MSW), as proposed by Buitelaar et al. (2012), to create

a level playing field for users with different cultural backgrounds, native languages and originating from different geo-political environments.

The MSW,

in which all languages have the same status, every user can perform searches [and contribute] in their own language, and information can be contrasted, compared and integrated across languages (Buitelaar et al. 2012)

has all the potential to serve as a virtual knowledge commons. In short, the MSW may be seen as a significant step towards cognitive justice.

Africa, and in particular South Africa, seems poised to embrace this cognitive justice: On the one hand, it represents significant cultural and language diversity, and, on the other hand, broadband access to the Internet is growing rapidly.

The vast majority of (the over 2,000) languages on the continent of Africa are under-resourced. By under-resourced languages, we mean languages that have a small or economically disadvantaged user base, that are therefore typically ignored by the commercial world and that are technologically underdeveloped due to limited human, financial and linguistic/language resources (LRs). For example, only three sub-Saharan African languages, Yoruba, Swahili and Afrikaans, have any meaningful representation (but all still less than 30,000 entries) in Wikipedia (Deumert *in press*). Table 1 summarises the distribution of world languages by area of origin (Lewis 2009).



**Table 1** Distribution of languages by area of origin<sup>a</sup>

Area	Living languages		Number of speakers			
	Count	Percentage	Count	Percentage	Mean	Median
Africa	2,146	30.5	789,138,977	12.7	367,726	27,000
Americas	1,060	14.9	51,109,910	0.8	48,217	1,170
Asia	2,304	32.4	3,742,996,641	60.0	1,624,565	12,000
Europe	284	4.0	1,646,624,761	26.4	5,797,975	63,100
Pacific	1,311	18.5	6,551,278	0.1	4,997	950
Totals	7,105	100.0	6,236,421,567	100.0	877,751	7,000

<sup>a</sup> <http://www.ethnologue.com/statistics>

South Africa has eleven official languages enshrined in its constitution, ten of which are indigenous under-resourced languages. The numbers of mother-tongue speakers per language are Zulu (11,587,374), Xhosa (8,154,258), Afrikaans (6,855,082), English (4,892,623), Northern Sotho (4,618,574), Tswana (4,067,248), Southern Sotho (3,849,563), Tsonga (2,277,148), Swati (1,297,046), Venda (1,209,388) and Ndebele (1,090,223) (Statistics South Africa 2012).

In sub-Saharan Africa, only 12 % of the population owns a desktop PC, with laptops at the same level of penetration. However, already 18 % of the population owns a smartphone, while in South Africa 90 % of the population owns a mobile phone, of which 48 % are basic feature phones, 19 % are advanced feature phones and 33 % are smartphones (TNS 2013).

Other drivers for cognitive justice are the education system, which is constantly under siege and deteriorating, with less than 10 % of the population being mother-tongue speakers of English, and the literacy rate, which is 86.4 % (Index Mundi 2012), well below average (number 128 out of 204 countries). The challenges for South Africa to participate in the twenty-first-century knowledge economy, which includes the MSW, have to be addressed with some urgency.

Multilingualism is an essential characteristic of South African society. English is the de facto lingua franca; Afrikaans has an established resource base but is, due to historical realities, under constant pressure; a number of the South African Bantu languages may, in the future, become endangered languages due to the attitudes of their speakers—the number one reason for a language to become *endangered* is so-called language shift (Grimes 2001). This means that, in most general terms, parents are no longer teaching the language to their children and are not using it actively in everyday matters. Nevertheless, a rich cultural diversity and indigenous knowledge systems (IKSs) are encoded in the various languages, and multilingual and cross-lingual information is of strategic importance to the public, private, business, technical, science and educational sectors.

The participation of the South African under-resourced languages in the MSW is, therefore, imperative in terms of sharing in and contributing to the knowledge commons; sustaining multilingualism and the technological development of these

languages, cultural diversity and IKSs; and the MSW's potential to support cross-lingual knowledge production and consumption.

The remainder of this chapter is devoted to three aspects. Firstly, four pertinent challenges facing the under-resourced South African languages in participating in the MSW are briefly discussed. Secondly, the chapter focusses on exposing fragments of selected South African language resources as Linked Data. These fragments are taken from a recently developed terminology (Statistics South Africa 2009) for all eleven official languages, novel Tswana indigenous knowledge on astronomy (Leeuw 2007, 2014) and a parallel corpus in the form of the South African Constitution in four of the eleven languages (Constitution of the Republic of South Africa 1996). Advanced aspects such as consuming, contrasting, comparing, integrating and generating knowledge in the MSW fall outside the scope of this chapter. Thirdly, the chapter concludes with a discussion of possible future work.

## 2 Challenges

The first two challenges, viz. under-resourcedness and supporting indigenous knowledge systems, are about building, and the second two, viz. the MSW as platform for the language technology development and interoperability and ease of use, are about using the MSW as virtual knowledge commons.

### 2.1 *Under-Resourcedness*

In the context of the MSW, the most basic challenges that face Africa, and in particular South Africa, include the following:

- The large number of languages, most of them under-resourced with rich cultural diversity and often extensive IKSs encoded in these languages
- The scarceness of political will to use and develop the South African languages in the MSW—both on the side of the users and government
- Time as resource—the diminishing time window of opportunity as the knowledge economies of the rest of the world are gaining momentum
- Limited financial resources—as seen against the background of the demands of a developing society
- Limited human resources—specifically with respect to interested, highly skilled, specialised expertise, both in linguistics and the computational sciences
- Internet connectivity as resource—broadband Internet access is limited, with fast growing mobile phone usage
- Language and (multilingual) linguistic resources—limited (multilingual) language data, enabling technologies, tools and applications

A 2011 technology audit of the South African Human Language Technology (HLT) landscape created a systematic and detailed inventory of the status of the HLT components across the eleven official languages (Grover et al. 2011). In this audit,

the lack of language resources (LRs), limited availability of and access to existing LRs, quality of LRs, small-scale and uncoordinated HLT development, and the lack of infrastructure for LR management

were identified as common issues faced by the development of LRs in resource-scarce languages. Moreover, it was found that very few basic LRs and applications exist across all eleven languages and that the South African languages

lie fallow in terms of the variety, number and maturity of items, compared to other world languages.

Furthermore, the general unavailability of text sources, such as newspapers, books, periodicals and documents, particularly for the smaller Bantu languages, constitutes a severe limitation to HLT development.

However, in 2011, the Department of Arts and Culture established the National Centre for HLT to develop reusable text and speech resources and the Resource Management Agency (RMA 2013) to manage and distribute these from one central point. None of these resources have as yet been exposed as Linked Data. Another initiative that has been reported on in the scientific literature is the African WordNet project (Griessel and Bosch 2014), but at the time of writing the African WordNets have not yet been published. A number of prototype finite-state morphological analysers and a finite-state tokeniser for Tswana are in advanced stages of completion (Pretorius and Bosch 2010; Pretorius et al. 2010; Bosch et al. 2008) and have been applied, amongst others, to corpus annotation (Bosch and Pretorius 2011). Lastly, the social media, for example, Facebook, offer a source of language data that have not as yet been sufficiently exploited (Deumert *in press*).

Indeed,

Only a very small number (perhaps thirty) of the world's 6000+ languages currently enjoy the benefits of modern language technologies such as speech recognition and machine translation. A slightly larger number (less than 100) have managed to assemble the basic resources needed as a foundation for advanced end-user technologies: monolingual and bilingual corpora, machine-readable dictionaries, thesauri, part-of-speech taggers, morphological analysers, and parsers . . . The remainder (certainly more than 98 per cent of the world's living languages) lack most, and usually all, of these tools, and we therefore refer to these as under-resourced languages (Scannell 2007).

On the basis of this summary, the present under-resourcedness in the language technology sense of the indigenous South African languages remains a reality.

## 2.2 *Indigenous Knowledge Systems*

Indigenous knowledge (IK) refers to the large body of knowledge and skills that has been developed outside the formal educational system. IK is embedded in culture and is unique to a given location or society (UNESCO *s.a.*). Moreover, language

is the most fundamental way that cultural information is communicated and preserved, especially in those that until recently did not use written expressions. Language's important relationship to knowledge and the survival of a culture requires that any discussion of IKSs must include [indigenous] language retention (Settee 2008).

The importance of including IKSs in the MSW as virtual knowledge commons is, therefore, clear.

It may not be feasible to try to compile any form of summary of the IK of the Southern African region and its peoples.<sup>1</sup> A more realistic, tractable and sustainable approach is to attempt to create awareness amongst and training for the people owning the IK and speaking the indigenous languages and also the development professionals that collaborate with them, of what the MSW offers as a virtual commons and how they could start to participate. This is a truly interdisciplinary undertaking and will require continued research collaboration between the NLP, MSW and indigenous language communities, development specialists and domain experts in various disciplines.

Important questions relating to IKS, which, due to their scope and complexity, fall outside the scope of this chapter, are stated as future work in the final section of the chapter.

### ***2.3 The MSW as Platform for the Technological Development of Under-Resourced Languages***

A serious issue in under-resourced languages remains the lack of terminology. The MSW with its standards, guidelines and best practices offers unique opportunities in terms of community-based (crowdsourcing) approaches to, amongst others, terminology development and moderation, and conceptual and linguistic representations of culture-specific and IKSs. The MSW may serve as an incubator for the continued development of increasingly sophisticated natural language processing and lexical resources for under-resourced languages. By careful planning and prioritisation and by collaborating, nationally and internationally, with other parties and communities interested in under-resourced languages and the exposition of IKS, new approaches may emerge due to the availability of rich cross-language support, resources, tools and technologies. The sustainability of under-resourced languages and IKS in the MSW will depend on continued engagement with and training of interested and committed cultural and language communities.

---

<sup>1</sup>The plural form of *people* is used here to refer to groupings of persons sharing, for example, a culture.

## 2.4 *Interoperability and Ease of Use*

Ultimately, semantic and semantic web technologies will provide interoperability across and beyond language boundaries in the MSW at a grand scale, as many of the contributions in this book attest to. The MSW will be characterised by its vastness, inference capabilities and diversity in knowledge, language and formats.

At a more modest scale, for the real uptake of emerging semantic technologies and the MSW, it should also be relatively easy for a single user to produce and consume specialised content; to conceptualise his/her arbitrarily complex interest domains, tasks, and applications; and to use the range of available MSW resources, representation and reasoning tools to his/her competitive advantage. Examples of specific functionalities that may be relevant for a wide range of MSW users include:

- To have access to state-of-the-art support and best practices of knowledge representation
- To do sophisticated intelligent searches of specified scope
- To delimit the search, access, generation and publication of information in languages of choice
- To perform automated reasoning of specified scope and complexity in the MSW
- To obtain semantically accurate translations of the retrieved or generated material and of the reasoning results, on request
- To provide large-scale automated decision-making support in (multiple) natural language(s)
- To have access to approaches and tools to evaluate results obtained

## 3 Towards Linked Data: Examples of Basic Resources

### 3.1 *General Approach*

We address the first two challenges of Sect. 2 by considering three example contributions towards producing resources in the MSW and exposing them as Linked Data, specifically focussing on multilingual aspects.

The first two (lexical-semantic) example contributions concern multilingual terminology and astronomy IK. As basis for this, we use *lemon* (McCrae et al. s.a., 2012), a model for the representation of ontology-lexica as Linked Data that has gained in use in the past years. *lemon* is an extensible model for Linked Data lexica; it was designed to interact with existing technologies and standards; its data categories allow for representation of arbitrary linguistic information; and it supports, amongst others, importing from ISOcat, an ISO Data Category Registry aimed at facilitating interoperability at the level of linguistic encoding (tag sets, metadata elements, etc.) (Windhouwer et al. 2013). *lemon* is particularly suitable

for our purpose since it is concise, descriptive, modular and Resource Description Framework (RDF) native—the emphasis in our work on under-resourced languages being on agility, urgency and parsimony in exposing initially small amounts of high-quality data. Indeed, we will be using little more than the *lemon core* in this chapter. Moreover, *lemon* supports multilingualism by allowing *lemon* lexica in and for different languages and the linking of their individual entries to shared abstract concepts (ontology) via the `lemon:sense` information. Finally, the model provides a principled chain between the semantic representation and its linguistic realisation—semantics by reference. In summary, *lemon* can be considered as an emerging standard in lexical-semantic resources for the MSW (Chiarcos et al. 2013).

The third example contribution concerns the use of Linked Data principles to develop a parallel corpus. Here no clear emerging standards could be identified, but the POWLA ontology (Chiarcos 2012), an OWL/DL-based formalism for presenting interoperable linguistic corpora, such as parallel corpora, offers appropriate and promising possibilities.

Note that the resources, the URIs of which are provided in the footnotes of the following sections, form an integral part of the text.

### 3.2 *Brief Overview of the Relevant Language Specifics*

Afrikaans is a language closely related to Dutch and Flemish, with compounding as productive word-forming process (e.g. see Botha et al. 1989). Tswana (a member of the Sotho language group) and Zulu (a member of the Nguni language group) belong to the Southern Bantu language family (Kosch 2006, p. x). The Bantu languages are morphologically complex, with large numbers of morphemes sequenced together to form words. Syntactically, they are characterised by a nominal classification system with concordial agreement. In particular, the term *class gender* is used to refer to the way in which nouns are grouped together into classes in a grammatically significant way. There are up to 20 different noun classes, occurring in singular/plural pairs. Gender agreement must be observed in all parts of the utterance which are linked to the noun (Kosch 2006, p. 90). Tswana has a so-called disjunctive orthography in which sequences of prefixes (morphemes) in verb constructions are written with whitespace in between (Krüger 2006). Zulu, on the other hand, has a conjunctive writing style (Poulos and Msimang 1998) where morpheme sequences in words are not separated by whitespace, as is the case in Tswana. A more detailed discussion of the general characteristics of Afrikaans, Tswana and Zulu falls outside the scope of this chapter. Specific aspects are mentioned as they pertain to the examples shown.

The choice of the specific Bantu languages, used in this chapter, was also informed by the availability of a finite-state tokeniser and finite-state morphological analyser for Tswana (Pretorius et al. 2010) and a finite-state morphological analyser for Zulu (e.g. see Pretorius and Bosch 2010). Due to their complex morphology, such enabling technologies are essential for any future semiautomated processing or annotation of language resources and resource development.

### 3.3 *Multilingual Terminology in English, Afrikaans, Tswana and Zulu*

Statistics South Africa identified a set of 896 terms across 19 domains that were considered to be of core interest for their reporting to the South African Government (Statistics South Africa 2009). These terms are provided in all 11 official languages and were compiled according to relevant standards and best practices. They, therefore, constitute a valuable high-quality resource worth exposing. We outline a basic procedure for creating *lemon* lexica, one per language, for this resource. We consider four languages and use one example term from the published list. These *lemon* lexica are interlinked via the concepts associated with the entries that they contain. Only the most basic metadata are provided, viz. the source of the information and the language of the lexicon.

For each term, we proceed as follows (McCrae et al. s.a.):

1. Find or build the basic (*lemon core*) English (source language) term, that is the canonical forms and senses for the term, as well as for each part of its composition, if the term consists of more than one part. While the canonical form is the written representation of the term or part thereof in the relevant language, the definition of the different senses requires the identification of cool URIs to link to—one for the entire concept and one for each part. Readily available semantic search engines are employed for this purpose. In this crucial step, the essential notion of Linked Data is established since the terms in different target languages would be linked to the same concept.
2. Build basic *lemon* entries for the terms in the other (target) languages, reusing the sense information in (1), where appropriate.
3. Add basic linguistic information as appropriate for each language, for example, number, class gender and part of speech (POS) by linking to resources such as ISOcat and LexInfo (Cimiano et al. 2011) for linguistic interoperability.

The English term *intangible assets* (Ontology-Lexica W3C Community Group 2013; Statistics South Africa 2009) was chosen as our example, and Afrikaans, Tswana and Zulu *lemon* entries<sup>2</sup> were then handcrafted for it. Although morphological information was not added, the modularity of *lemon* allows for this at any time in the future. We briefly comment on the differences between the English entry and the others (see URI in footnote 2):

- Afrikaans: The distinction between the attributive (*ontasbare*) and predicative (*tasbaar*) forms of the adjective, both included in the *lemon* entry.
- Tswana and Zulu: The main differences are twofold. Firstly, the inclusion of class gender information and the replacement of the adjective with the relative clause. The noun root for *asset* belongs to class 9, both in Tswana (*thoto*) and Zulu (*mpahla*), and takes its plural in class 10. Classes 9/10 often contain foreign

---

<sup>2</sup><http://gama.unisa.ac.za/files/rdf/MSW-chapter-lex>.

**Table 2** Morphological analysis of *dithoto tse di sa tshwareng* (intangible assets)<sup>a</sup>

Token in term	Morphological analysis	English meaning
<i>Dithoto</i>	NPre10+ [thoto]	Goods/possessions
<i>Tse</i>	QualPart10	That
<i>Di sa tshwareng</i>	SC10+NegPre+ [tshwareng] +VerbEnd+ RelSuf	They cannot keep/hold

<sup>a</sup> <http://gama.unisa.ac.za/files/MSW-chapter-morphTags.pdf>

words and also diverse words of Bantu origin, designating, amongst others, various kinds of fruit, names of animals and objects in everyday use. Secondly, the two components in the *lemon* decomposition should exhibit the required class gender agreement, viz. both components should be in class 10 since *intangible assets* is plural in number. When the Tswana or Zulu term is to be used in a sentence, morphosyntactic modifications will be required in accordance with the nominal classification and concordial agreement system of the specific language. For this reason, the usefulness of the Tswana and Zulu lexicons would be greatly enhanced by adding morphological information using, for example, the *lemon* morphology module. In the Tswana example in Table 2, each token exhibits the expected class gender agreement (i.e. class 10).

Work is currently underway to develop *lemon core* entries for each term in Statistics South Africa (2009) for all 11 official languages.

### 3.4 Indigenous Astronomy Knowledge in Tswana

By way of example, the novel Tswana *astronomy lesson* (Leeuw 2014) serves as a small body of indigenous knowledge (Leeuw 2007) to be exposed. It contains fascinating principled information on Tswana astronomy nomenclature, together with explanations as to the origins of these terms in Tswana. Table 3 shows the terms, morphological analysis, literal meanings and the astronomical IK meanings.

In the *lemon* lexicon,<sup>3</sup> a *lemon core* entry is provided for each term, enhanced with a `rdfs:comment`, which provides the English term and a `dublincore:description`, which gives a short English narrative, explaining the origin of IK by which the term was coined. The most time-consuming part was the manual process of identifying suitable URIs for the different concepts, to be linked to in the `lemon:sense` part of the entry. The addition of the class gender information, as given in Table 3, would enhance the usability of these terms in sentences,

<sup>3</sup><http://gama.unisa.ac.za/files/rdf/MSW-chapter-IKS>.



**Table 3** Morphological analyses of the Tswana words in the IKS lexicon

Term	Morphological analysis	Literal meaning	IKS meaning
<i>Dikolojwane</i>	NPre10+ [kolobe] +DimSuf	Piglets	Orion's belt
<i>Dintswa</i>	NPre10+ [ntswa]	Dogs	Orion's sword
<i>Kopadilalelo</i>	NPre9+ [kop] +DeverbSuf+ NPre8+ [lalelo] :Compound	Young seeker of evening meals	Venus, the evening star
<i>Mosese</i>	NPre3+ [sese]	Dress	Moon
<i>Molagodimo</i>	NPre3+ [la]+ Npre5+ [godimo] :Compound	The path above	Milky way
<i>Mphatlalatsane</i>	NPre9+ [mphatlalatsane]	The brilliant one	Venus, the morning star
<i>Ngwedi</i>	NPre9+ [ngwedi]	Female monthly cycle	Moon

as described in Sect. 3.2. This, as well as expanding the lexicon with more terms, forms part of future work.

### 3.5 A Multilingual Parallel Corpus Fragment

The significance of annotated parallel corpora for, amongst others, multilingual natural language processing and machine translation is well documented (e.g. see Ahrenberg et al. 2010). The development of such resources is a challenging task and usually involves linguistic annotation at multiple levels, for example, morphological, POS, named entities, chunks/phrases, other sentence constituents, sentences, etc., but also any number of semantic annotations, depending on what the corpora will be used for. For the South African languages, parallel corpora annotated by any means or for any purpose have, up to the time of writing, not been available. In this section, we follow a pragmatic approach that makes use of what is available in terms of data and enabling technologies and then propose a possible Linked Data approach for the future exposition of such data as a parallel corpus.

As in the previous two sections, the work is exploratory in nature but already forms the basis of an ongoing project. It is explicated by means of a small fragment—one sentence from the Constitution of the Republic of South Africa<sup>4</sup> in English, Afrikaans, Tswana and Zulu—which already presents various challenges and salient features of such an endeavour. The sentence (SE) below is from Paragraph 80(4) of the Constitution of the Republic of South Africa, Constitutional

<sup>4</sup>Apart from the Bible, the Constitution is one of a small number of high-quality parallel corpora, available in all 11 official languages.

**Table 4** The clauses in the respective sentences

	1	2	3
CE	If an application is unsuccessful	(And) did not have a reasonable prospect of success	The Constitutional Court may order the applicants to pay costs
CA	<i>Indien 'n aansoek nie slaag nie</i>	<i>(En) nie 'n redelike vooruitsig gehad het om te slaag nie</i>	<i>Kan die Konstitusionele Hof die aansoekers gelas om die koste te betaal</i>
CT	<i>Fa kopo e sa atlega</i>	<i>Kgotlatshekelo ya Molaotheo e tshwanetse go laela bakopi gore ba duele ditshwenyegelo</i>	<i>Ntle le fa kopo e ne e na le thono e e isegang ya go atlega</i>
CZ	<i>Uma isicelo singaphumeleli</i>	<i>iNkantolo yoMthethosisekelo kufanele inqume ukuthi labo abafake isicelo kufanele bakhokhe izindleko</i>	<i>Ngaphandle uma isicelo besinamathuba anele okuphumelela</i>

Law No. 108 of 1996 (Constitution of the Republic of South Africa 1996), with translations into Afrikaans (SA), Tswana (ST) and Zulu (SZ). The clauses that constitute the sentences, are given in Table 4.

SE: If an application is unsuccessful and did not have a reasonable prospect of success, the Constitutional Court may order the applicants to pay costs.

SA: *Indien 'n aansoek nie slaag nie, en nie 'n redelike vooruitsig gehad het om te slaag nie, kan die Konstitusionele Hof die aansoekers gelas om die koste te betaal.*

ST: *Fa kopo e sa atlega, Kgotlatshekelo ya Molaotheo e tshwanetse go laela bakopi gore ba duele ditshwenyegelo ntle le fa kopo e ne e na le thono e e isegang ya go atlega.*

SZ: *Uma isicelo singaphumeleli, iNkantolo yoMthethosisekelo kufanele inqume ukuthi labo abafake isicelo kufanele bakhokhe izindleko ngaphandle uma isicelo besinamathuba anele okuphumelela.*

We first discuss the data and the approaches available for obtaining the annotations, using available resources, tools and also handcrafting, where necessary. We then consider how this information can be best exposed as Linguistic Linked Data towards creating a state-of-the art parallel corpus, based on Chiarcos (2012) and Chiarcos et al. (2013), and, in particular, on POWLA as a formalism for representing linguistic corpora in RDF (POWLA s.a.).

### 3.5.1 Linguistic Annotation Approaches

The detailed morphological, POS and semantic information<sup>5</sup> for this multilingual one-sentence fragment was obtained as follows:

- *Tokenisation*: English, Afrikaans and Zulu were tokenised mainly on whitespace, while for Tswana, due to its disjunctive orthography, a finite-state tokeniser (Pretorius et al. 2010) was used.
- *Morphological analysis*: For agglutinating languages such as Tswana and Zulu, morphological analysis is essential for the extraction of roots, which carry substantial semantic knowledge in the form of the central meaning of words. For both these languages, finite-state morphological analysers were employed (Pretorius et al. 2010; Pretorius and Bosch 2010). Each language has its own system of morphological analysis (e.g. see Krüger 2006 for Tswana and Poulos and Msimang 1998 for Zulu), which manifests itself in the different, but often closely related, annotations and tag sets that are employed.<sup>6</sup> For English and Afrikaans, no morphological analyses were added since they are usually not required for next levels of processing.
- *Morphological disambiguation*: For the Tswana and Zulu data, this was done manually. Semiautomated morphological disambiguation forms part of future work.
- *POS*: For English and Afrikaans, with their less complex morphology, available POS tagging tools were used—the NCLT LFG-online parser<sup>7</sup> for English and a POS-tagger, based on Pilon (2005), for Afrikaans. For Tswana and Zulu, the POS information was directly derived from the morphological analyses.
- *Semantic information in the form of the English equivalents*: Was manually added to the tables for Afrikaans, Tswana and Zulu.

Although the sentences mean exactly the same and each has been analysed in quite some linguistic detail, there is not yet any way in which this information can be exploited in the context of a parallel corpus. If we assume for the moment that a parallel corpus such as the SA Constitution may be semantically aligned at the sentence level, it remains to exploit these available annotations to enrich the corpus at subsentence level.

### 3.5.2 Towards a Parallel Corpus Using Linked Data Principles with POWLA

PAULA (Chiarcos 2012) is a generic data model for the representation of annotated corpora. It captures the insight that any kind of linguistic annotation can be

---

<sup>5</sup><http://gama.unisa.ac.za/files/MSW-chapter-annotations.pdf>.

<sup>6</sup><http://gama.unisa.ac.za/files/MSW-chapter-morphTags.pdf>.

<sup>7</sup><http://lfg-demo.computing.dcu.ie/lfgparser.html>.

represented by means of directed acyclic graphs. It makes provision for primary data (text); linguistic annotations consisting of three principal components, viz. segments (spans of text, modelled as nodes); relations between segments (modelled as edges); and annotations that describe different types of segments or relations (modelled as labels). PAULA differentiates between two types of edges with respect to their relationship to the primary data. For hierarchical structures, for example, phrase structure trees, a notion of coverage inheritance is required (the text covered by a child node is always covered by the parent node)—such edges are referred to as dominance relations. For other kinds of relational annotation, no constraints on the coverage of the elements connected need to be postulated. A typical and relevant example is alignment in parallel corpora. Such edges are referred to as pointing relations (Chiaros 2012).

POWLA is an RDF/OWL linearisation of PAULA and consists of two basic components: (1) an OWL/DL ontology that defines the valid data types, relations and constraints as classes, properties and axioms (2) an RDF document that represents a corpus as a knowledge base consisting of individuals, instantiated object properties and data values assigned to individuals through data-type properties. POWLA formalises the structure of annotated corpora and linguistic annotations of textual data. For example, it provides data types such as `Node` and `Relation`, as well as more specialised data types that directly reflect the underlying graph-based data model. With OWL/DL axioms, the relationship between these data types can be formalised and automatically verified (Chiaros et al. 2013).

As we have already seen with the variations in morphological annotation of Tswana and Zulu and the POS tags of Afrikaans and English, communities create their own grammatical notations, often developed to represent different terminological traditions and different language systems. OliA (Ontologies for Linguistics Annotation) are a modular set of ontologies that establish linking between such differing systems. The OliA Reference Model specifies the common terminology that different annotation schemes can refer to, and the OliA Annotation Models formalise annotations schemes and tag sets for a wide variety of languages. For every Annotation Model, a Linking Model defines relationships between concepts and properties in the respective Annotation Model and the Reference Model. In combination with POWLA, the OliA ontologies allow the representation of linguistic annotations and their meaning within the Linguistic Linked Open Data cloud in an interoperable way (Chiaros 2012; Chiaros et al. 2013).

Using this model, our one-sentence parallel corpus (in which each entity would have a URI) may be conceptualised as follows. For argument's sake, we differentiate between the following layers in our corpus:

- *Layer 1*: Four sentences (SE, SA, ST and SZ) as segments, with appropriate identity relations (Halpin et al. 2010) as pointing relations between them to realise the notion of parallelism.
- *Layer 2*: Three clauses for each of the four sentences (CE1, CE2, . . . , CZ3) (see Table 4) as segments, appropriately linked to their parent segments in layer 1 by means of dominance relations (e.g. `hasParent`) and to their respective

other language equivalents, again by appropriate identity relations as pointing relations, realising parallelism at the clause level. The different translations require that, amongst others, segment CT2 be linked to CE3, CZ2 to CT2, CT3 to CE2 and CZ3 to CT3 by means of appropriate identity relations as pointing relations to model the difference in the clause order in the respective sentences. The RDF representation will ensure that the other implicit links between CZ2 and CE3 and CZ3 and CE2 are also appropriately made.

- *Layer 3 to n*: Any appropriate number of layers as required by the respective linguistic analyses and annotation requirements<sup>8</sup> of the different languages, with appropriate dominance and pointing relations, ending with the tokens. The model allows for further layers, for example, the morphological layer. These decisions are the prerogative of the builder(s) of the parallel corpus.

The annotations at all levels, modelled as labels on segments and relations, would be done according to the different language systems and tag sets, while the linking between these entities in the different systems and between tags in the different tag sets may be taken care of through OliA. Finally, POWLA also allows linking to other data repositories, such as *lemon* lexica and the ISOcat data registry, which would enhance cross-lingual semantic interoperability of the parallel corpus. The linking possibilities are extensive and offer potential for an increasingly useful multilingual resource. As a next future step, this one-sentence parallel corpus, which was discussed extensively, should be built as a first small prototype since it contains many of the challenges and salient characteristics of a larger parallel corpus.

## 4 Conclusion and Future Work

In this chapter, we argued for the full participation of the under-resourced South African languages in the MSW as a virtual knowledge commons. We identified four challenges that these languages face in achieving this ideal, viz. under-resourcedness, IKSSs, the MSW as platform for the technological development of under-resourced languages, and interoperability and ease of use. We showed how the first two challenges may be mitigated. Procedures were proposed for exposing multilingual terminology in English, Afrikaans, Tswana and Zulu, Tswana indigenous knowledge on astronomy and a fragment of a parallel corpus in English, Afrikaans, Tswana and Zulu as Linked Data. These procedures serve as a platform for growing such Linked Data resources in the near future.

Future work also includes the continued development and refinement of natural language technologies for the under-resourced languages of South Africa, as well as the sustained exposition of available and newly developed resources as Linked

---

<sup>8</sup><http://gama.unisa.ac.za/files/MSW-chapter-annotations.pdf>.

Data in all 11 official languages, using state-of-the-art semantic and semantic web technologies.

Much still needs to be done with respect to IKSs: How should IKS, including new concepts and/or new relationships between concepts, old and new, be documented and exposed in the MSW at an increasing scale? How could the work on Cultural Heritage and the MSW (e.g. see Hyvönen 2012) impact on IKSs and their place in the MSW as virtual knowledge commons? What infrastructure and mechanisms should, for example, be in place for the lexicalisation and verbalisation of the new (indigenous) knowledge in multiple languages? How would the relevant communities be empowered to assume these responsibilities?

Finally, the significance, the character, the scientific and the societal impact of the MSW as a virtual knowledge commons should continue to be cherished, studied and expanded for the greater good.

## References

- Ahrenberg, L., Tiedemann, J., & Volk, M. (Eds.). (2010). *Proceedings of the Workshop on Annotation and Exploitation of Parallel Corpora*. NEALT Proceedings Series (Vol. 10). Tartu, Estonia: University of Tartu.
- Bosch, S.E., & Pretorius, L. (2011). Towards Zulu corpus clean-up, lexicon development and corpus annotation by means of computational morphological analysis. *South African Journal of African Languages*, 31(1), 138–158.
- Bosch, S. E., Pretorius, L., & Fleisch, A. (2008). Experimental bootstrapping of morphological analysers for Nguni languages. *Nordic Journal of African Studies*, 17(2), 66–88.
- Botha, T. J. R., Ponelis, F. A., Combrinck, J. G. H., & Odendal, F. F. (1989). *Inleiding tot die Afrikaanse taalkunde*. Pretoria, South Africa: Academica.
- Buitelaar, P., Choi, K.-S., Cimiano, P., & Hovy, E. D. (Eds.). (2012). The multilingual Semantic Web (Dagstuhl Seminar 12362). *Dagstuhl Reports*, 2(9), 15–94.
- Chiarcos, C. (2012). Interoperability of corpora and annotations. In C. Chiarcos, S. Hellmann, & S. Nordhoff (Eds.), *Linked data in linguistics*. Berlin, Germany: Springer.
- Chiarcos, C., McCrae, J., Cimiano, P., & Fellbaum, C. (2013). Towards open data for linguistics: Linguistic linked data. In A. Oltramari, et al. (Eds.), *New trends of research in ontologies and lexical resources, theory and applications of natural language processing*. Berlin, Germany: Springer.
- Cimiano, P., Buitelaar, P., McCrae, J., & Sintek, M. (2011). LexInfo: A declarative model for the lexicon-ontology interface. *Journal of Web Semantics*, 9(1), 29–51.
- Constitution of the Republic of South Africa (English). (1996). Retrieved from <http://www.info.gov.za/documents/constitution/93cons.htm>.
- Deumert, A. (in press). Sites of struggle and possibility in cyberspace - Wikipedia and Facebook in Africa. In J. Androutsopoulos (Ed.), *The media and sociolinguistic change*. Berlin, Germany: De Gruyter (forthcoming).
- Griessel, M., & Bosch, S. (2014). Taking stock of the African Wordnets project: 5 years of development. In *Proceedings of the 7th Global WordNet Conference (GWC2014)*, Tartu, Estonia.
- Grimes, B. F. (2001). Global language viability. In O. Sakiyama (Ed.), *Endangered languages of the Pacific rim: Lectures on endangered languages 2. ELPR Publication Series C002*. Osaka, Japan: ELPR. Retrieved from <http://www.sil.org/sociolx/ndg-1g-grimes.html>.

- Grover, A. S., Van Huyssteen, G. B., & Pretorius, M. W. (2011). South African human language technology audit. *Language Resources and Evaluation*, 45(3), 271–288.
- Halpin, H., Hayes, P., McCusker, J. P., McGuinness, D. L., & Thompson, H. S. (2010). *When owl: sameAs isn't the same: An analysis of identity in linked data*. *Lecture Notes in Computer Science* (Vol. 6496, pp. 305–320). Berlin/Heidelberg: Springer.
- Hyvönen, E. (2012). Publishing and using cultural heritage linked data on the semantic web. In J. Hendler (Series Ed.), *Synthesis lectures on the semantic web: Theory and technology*. CA, USA: Morgan & Claypool Publishers.
- Index Mundi. (2012). South Africa literacy. Retrieved from [http://www.indexmundi.com/south\\_africa/literacy.html](http://www.indexmundi.com/south_africa/literacy.html).
- Innerness, D. (2011). *The democracy of knowledge. For an intelligent society* (H. D. D'Ambrosio, Trans.). Retrieved from [http://www.essayandscience.com/upload/ficheros/libros/201203/cap\\_innerness.pdf](http://www.essayandscience.com/upload/ficheros/libros/201203/cap_innerness.pdf).
- In 't Veld, R. (Ed.). (2010). *Knowledge democracy: Consequences for science, politics, and media*. Berlin, Germany: Springer.
- Kosch, I. M. (2006). *Topics in morphology in the African language context*. Pretoria, South Africa: University of South Africa.
- Krüger, C. J. H. (2006). *Introduction to the morphology of Setswana*. Munich, Germany: Lincom GmbH.
- Leeuw, L. L. (2007). Setswana astronomical nomenclature. *African Skies*, 11, 17–18.
- Leeuw, L. L. (2014). An exemplary astronomical lesson that could potentially show the benefits of multilingual content and language in higher education. In L. Hibbert & C. van der Walt (Eds.), *Multilingual Universities in South Africa: Reflecting society in higher education*. Bristol, UK: Channel View Publications Ltd.
- Lewis, M. P. (Ed.). (2009). *Ethnologue: Languages of the world* (16th ed.). Dallas, TX: SIL International. Retrieved from <http://www.ethnologue.com/>.
- McCrae, J., Aguado-de-Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gomez Perez, A., et al. (s.a.) *The lemon cookbook*. Retrieved from <http://www.lexinfo.net/sites/default/files/lemon-cookbook.pdf>.
- McCrae, J., Aguado-de-Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gomez Perez, A., et al. (2012). Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 46(4), 701–719.
- Ontology-Lexica W3C Community Group. (2013). Specification of requirements on terminological analysis. Retrieved from <http://www.w3.org/community/ontolex/wiki/>.
- Pilon, S. (2005). *Outomatiese Afrikaanse woordsoortetikettering*. Unpublished master's dissertation, North-West University, Potchefstroom, South Africa.
- Poulos, G., & Msimang, C. T. (1998). *A linguistic analysis of Zulu*. Cape Town, South Africa: Via Africa.
- POWLA. Retrieved from <http://nachhalt.sfb632.uni-potsdam.de/powla/>.
- Pretorius, L., & Bosch, S. E. (2010). *Finite state morphology of the Nguni language cluster: Modelling and implementation issues*. *Lecture Notes in Computer Science* (Vol. 6062, pp. 123–130). Berlin, Heidelberg: Springer.
- Pretorius, L., Viljoen, B., Pretorius, R., & Berg, A. (2010). *A finite-state approach to Setswana verb morphology*. *Lecture Notes in Computer Science* (Vol. 6062, pp. 131–138). Berlin, Heidelberg: Springer.
- RMA. (2013). Language resource management agency. Retrieved from <http://www.rma.nwu.ac.za/>.
- Scannell, K. (2007). The Crúbadán project: Corpus building for under-resourced languages. In C. Fairon, H. Naets, A. Kilgariff & G.-M. de Schryver (Eds.), *Building and Exploring Web Corpora. Proceedings of the 3rd Web as Corpus Workshop*, Louvain-la-Neuve, Belgium.
- Settee, P. (2008). *Native languages supporting indigenous knowledge*. United Nations Department of Economic and Social Affairs, Division for Social Policy and Development, New York, PFII/2008/EGM1/13.

- Statistics South Africa. (2009). *Multilingual statistical guide* (272 pp.). Pretoria, South Africa: Statistics South Africa. ISBN 978-0-621-38513-7.
- Statistics South Africa. (2012). *Census 2011 census in brief* (105 pp.). Report No.: 03-01-41. ISBN 978-0-621-41388-5.
- TNS. (2013). *Navigating growth in Africa*. Retrieved from [http://uk.kantar.com/media/138155/tns\\_navigating\\_growth\\_in\\_africa.pdf](http://uk.kantar.com/media/138155/tns_navigating_growth_in_africa.pdf).
- UNESCO. (s.a.). Retrieved from <http://www.unesco.org/most/bpindi.htm>.
- Van der Velden, M. (2006). *A case for cognitive justice*. Retrieved from <http://www.globalagenda.org/file/6>.
- Visvanathan, S. (2009). *The search for cognitive justice*. Retrieved from [http://www.india-seminar.com/2009/597/597\\_shiv\\_visvanathan.htm](http://www.india-seminar.com/2009/597/597_shiv_visvanathan.htm).
- Windhouwer, M., Schuurman, I., & Wright, S. E. (2013). Collaboratively defining widely accepted linguistic data categories in the ISOcat data category registry. Retrieved from [https://catalog.clarin.eu/isocat/index\\_bestanden/publications.html](https://catalog.clarin.eu/isocat/index_bestanden/publications.html).



# A Three-Dimensional Paradigm for Conceptually Scoped Language Technology

Jeroen van Grondelle and Christina Unger

**Abstract** Language technology is used increasingly for providing speech- and text-based interfaces to existing applications and services. However, a number of characteristics of today's language technology make it hard to be adopted by non-linguistically skilled developers. In this chapter, we propose a paradigm that conceptually scopes the coverage of the language technology that is adopted into existing applications. It is backed by a three-dimensional approach to modularization of resources that decouples the domains, tasks and languages that need to be supported. We present an implementation of this paradigm based on the ontology-lexicon format *lemon* and Grammatical Framework (GF), and show how the proposed modularity facilitates low impact adoption, through sharing and reuse of technology components and lexical resources on the web.

**Key Words** Conceptual scoping • Grammar generation • Modularity • Natural language interfaces • Ontology-lexica

## 1 Introduction

Natural language plays an increasingly important role as interface to existing services and data. Social networks, for instance, present updates and newsfeeds as natural language content, virtual assistants support users by allowing them to query different sources of information and to manipulate them using speech dialogs (Zue and Glass 2000; Kaljurand and Alumäe 2012) and business applications allow domain experts to customize the services by creating rules or manage complex configurations using natural language-based interfaces (Spreeuwenberg and Healy 2010; Spreeuwenberg et al. 2012). The development of language technology that has

---

J. van Grondelle (✉)  
Be Informed, Apeldoorn, The Netherlands  
e-mail: [j.vangrondelle@beinformed.com](mailto:j.vangrondelle@beinformed.com)

C. Unger  
CITEC, Bielefeld University, Bielefeld, Germany  
e-mail: [cunger@cit-ec.uni-bielefeld.de](mailto:cunger@cit-ec.uni-bielefeld.de)

to support these new application scenarios, however, so far has built on objectives and requirements that widely differ from those imposed by their role as interfacing technology, and the consequences still hinder an easy adoption in such scenarios. This can be demonstrated along three main points.

First, an objective of language technology often has been to process unrestricted language. On the one hand, this involves challenges that can be tackled very differently when interfacing with an application. In an unrestricted setting, for example, natural language expressions are highly ambiguous, while in the context of a particular application, they usually have a single, very specific meaning. That is, the application introduces a context that can be exploited for disambiguation. On the other hand, there is a mismatch between the very general, usually surface-oriented meaning representations created in an unrestricted setting and the demand of aligning language with data and services in the context of a particular application. Again, the interpretation of natural language expressions can be restricted and guided by the underlying application, as it inherently introduces a scope that determines the language fragment that is relevant and meaningful when interfacing with it.

Second, language technology tools and techniques have mainly been developed and used by linguistically trained people. Choosing from the range of available approaches and tools and implementing the selected technology in a specific application require linguistic expertise typically not found in the companies that develop applications that a language interface is adopted into. Therefore, the adoption is costly and requires high upfront investments.

And third, language technology tools often trade precision for additional coverage, while for companies adopting those tools, high precision as well as reliability and predictability become critical, as any misinterpretation can lead to immediate errors in the invocation or execution of the underlying service.

To support the adoption of language technology into existing services and applications by companies with little or no linguistic expertise, we propose a new architecture for language technology that

- Is *conceptually scoped* in the sense that it uses the application's conceptualization to scope language technology and as a consequence limits and tailors all interpreted and generated language to the specific application it is meant to interface with
- Modularizes the creation and use of resources by clearly separating three dimensions: domains, tasks, and languages

This shifts the mainstream paradigm of unscoped, monolithic, and therefore costly language technology to a new, strongly modular and inexpensive way of creating and exploiting natural language resources.

## 2 A Three-Dimensional Paradigm for Conceptually Scoped Language Technology

To address the challenges described above, we propose the following three principles for guiding the development of conceptually scoped, modular language technology.

### 2.1 *Scoping Natural Language Through a Conceptualization*

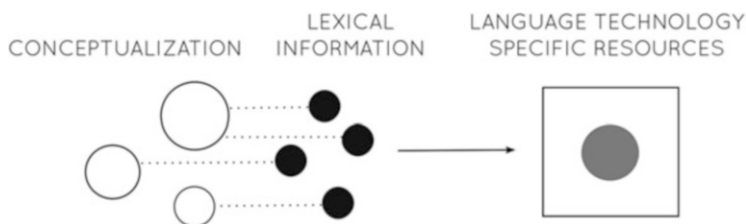
We propose that the scope of the natural language fragment that is to be supported, for instance, through interpretation, generation or translation, should be determined by a conceptualization. Such a conceptualization typically defines the individuals, classes, properties and relations that will be expressed in natural language, independent of the particular representation formalism used, and it should follow from the application or service that language is supposed to interface with. As a result, conceptual scoping grounds any supported natural language utterance in the underlying conceptualization and ensures it to be meaningful within that conceptualization (e.g. as advocated by Gatus and Rodríguez 1996 and Nirenburg and Raskin 2004).

This improves on the mainstream paradigm, where conceptualization and attached language technology are often developed independently from each other and where it is therefore hard to ensure that the conceptual and the linguistic scopes of an application are aligned, especially if the conceptualization changes over time.

### 2.2 *Automatically Generating Resources from Declarative Lexical Information*

The supported conceptualization has its own lexical aspects, which conventionally have been captured in formalisms that are highly dependent on the type of language technology used. In contrast, by building on a declarative format for specifying lexical information, those lexical aspects can be captured in a technology-neutral way. Technology-specific artefacts, such as grammars, can then be generated by means of a mapping from the declarative lexical representation to the technology-specific formalism.

We therefore propose the pipeline in Fig. 1, starting from a conceptualization that is enriched with a declarative specification of the lexical information associated with the given concepts (Reymonet et al. 2007; Montiel-Ponsoda et al. 2008; McCrae et al. 2012; Wróblewska et al. 2012). The resulting lexical representations are input to an automatic mapping to a language technology-specific resource, such as grammars, phrase tables, or semantic annotations.



**Fig. 1** Pipeline from conceptualization and lexical information to specific resources

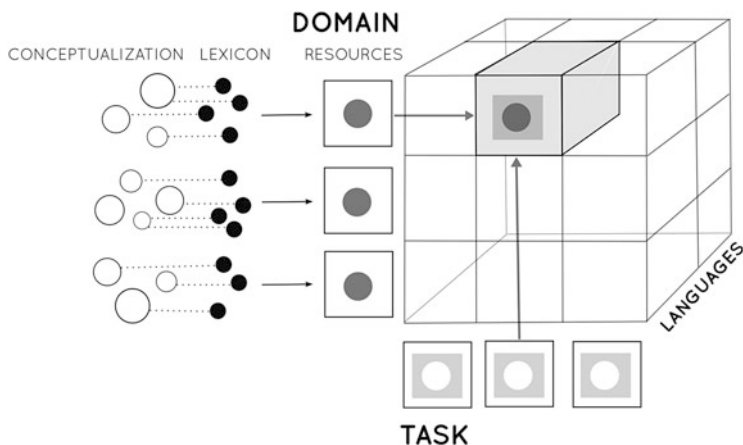
By automatically generating resources from an intermediate declarative lexical level, the investment into the lexical resources is protected, and consistency across the technology-specific resources is guaranteed. Furthermore, the developers of the lexical resources do not need to have expertise in language technology implementations, such as specific grammar formalisms. This lowers the investment required for natural language-based applications and furthermore removes the dependency on particular third-party tools. Also, when using declarative lexical formalisms, developers are less likely to make implicit choices concerning particular linguistic theories, which would hamper the reuse of the resources when adopting new technologies that do not agree with these choices.

### 2.3 *Decoupling Domain and Task Aspects*

Conventionally, a lot of emphasis has been on domain aspects (Martins and de Almeida Falbo 2008). But when providing natural language support in an application, the type of tasks supported by that application typically also has linguistic implications in terms of the natural language fragment that needs to be supported. For instance, the task of customer service dialog introduces its own words and sentence structures, independent of the domain. Similarly, task-specific linguistic aspects exist for tasks such as validation of domain ontologies, documentation, etc.

In order to allow for a reuse of task aspects across different domains, we propose the model depicted in Fig. 2 to decompose the scope introduced by the underlying applications that language technology interfaces with into three dimensions: the *domain* of the application, specified by some conceptualization; the *task* that the application offers, such as verbalizing domain data for explanation or documentation purposes or providing online services that include transactions and web forms in terms of the domain; and the *languages* in which the natural language capabilities are offered. The language fragment supported by an application is now defined by a subspace of the resulting cube, involving one or more domains together with one or more tasks and spanning one or more languages.

This orthogonal modularization of domain and task aspects supports specification of the conceptualization and lexical information per dimension, that



**Fig. 2** Three-dimensional model for conceptually scoped language technology

is specifying domains independent from tasks and vice versa. The dimensions can then be freely combined by choosing the particular domains, tasks and languages supported for a specific application. This not only allows for the reuse of already existing conceptualizations, such as adding new tasks to an existing domain or reusing task conceptualizations across different domains, but steadily increases the return on investment, since the more of these building blocks already exist, the easier and faster it is to plug them together to build new applications.

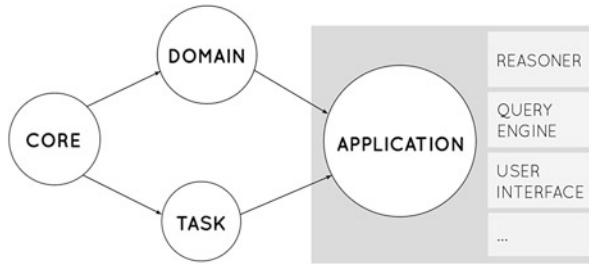
The importance of separating domain and task conceptualizations has also been noted, for example, by Guarino (1997) and Mizoguchi et al. (1995).

### 3 Proof of Concept in the Context of the Multilingual Semantic Web

As a proof of concept of the three-dimensional paradigm proposed, we implement a dialog-oriented natural language interface based on these principles, using a stack of technologies suitable in the context of the multilingual Semantic Web, and show that it supports typical scenarios in the incremental adoption of language technology.

#### 3.1 Implementation

For capturing conceptualizations and lexicalizations, we build on existing Semantic Web standards, in particular Web Ontology Language (OWL) and Resource Description Framework (RDF). Since these standards by their very nature enable linking and sharing of data, this supports the reuse of modules and facilitates an ecosystem of language technology resources, as discussed in Sect. 4.



**Fig. 3** Grammar modularity. *Arrows indicate grammar inheritance*

The domain conceptualization is captured as an ontology in the standard ontology format OWL (McGuinness and van Harmelen 2004). In order to be able to associate linguistic information with concepts in an ontology, the lexical component is implemented using *lemon* (McCrae et al. 2012), a model for the declarative specification of multilingual ontology-lexica in RDF. It allows lexical data to be published, shared and interlinked on the web and thus fits very well with our approach’s strong emphasis on modularity. Furthermore, it is independent of a particular linguistic theory or grammar formalism. In the following, we use Grammatical Framework (Ranta 2011) as target grammar formalism, benefiting from its inherent modularity and its support for more than 30 languages, which allows for very fast and effortless porting across those supported languages.

The instantiation of the pipeline from conceptualization and lexicalization to a specific grammar thus starts from an OWL ontology; then requires the creation (or reuse) of an ontology-lexicon for the target languages, specifying lexicalization of the ontology elements in those particular languages; and then relies on the automatic generation of multilingual grammars from that lexicon.

In order to percolate modularity up to the grammar level, we implement application grammars as being composed of three modules, as depicted in Fig. 3: a domain- and task-independent *core* grammar, an automatically generated *domain* grammar and an accompanying *task* grammar.

The core grammar comprises domain- and task-independent expressions, especially closed class expressions such as determiners, pronouns, auxiliary verbs, coordination expressions and negation. It is created manually and can be reused for every domain and task. It provides an independent basis on which both domain and task grammars build, therefore acting as a decoupler between them.

The domain grammar extends the core with expressions that are automatically generated from a given ontology-lexicon, following the pipeline in Fig. 1. The task grammar, on the other hand, extends the core with task-relevant expressions. As of now it is created manually, but carrying over the grammar generation pipeline from the domain to the task dimension and thereby also allowing for the automatic generation of task grammars constitute future work (see Sect. 5). The fact that domain and task grammars are constructed independently from each other, together with the fact that they share the core grammar as their basis, allows for a free and smooth combination of any domain with any task.

The application grammar finally combines core, domain and task grammars and furthermore allows for application-specific extensions or fine-tuning. The final application grammar is used for the specific purpose of the application, which possibly interfaces it with additional modules such as a reasoner, a query engine or a user interface.

## 3.2 Typical Adoption Scenarios

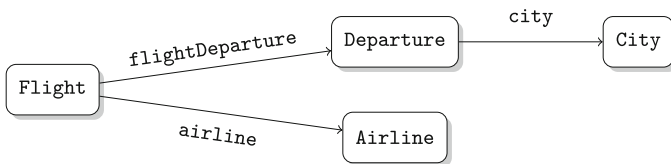
In the following, we illustrate typical adoption scenarios, in particular decoupling domain and task aspects, incorporating new domains and tasks and adding further languages. The mentioned resources can be accessed at <http://purl.org/3dlt/home>.

### 3.2.1 Decoupling Domain and Task Aspects

We start by building an application grammar for customer service dialog in the flight travel domain in English. That is, given a conceptualization of flight travel, we want to construct a grammar that captures utterances such as the following ones:

- Show me all flights from Boston to Detroit.
- Which airlines fly to San Francisco?
- I want to travel to New York tomorrow.
- When do you want to depart?
- Do you need a hotel in New York?

The domain conceptualization is modelled as an OWL ontology that was built in the context of the PortDial project,<sup>1</sup> based on terms used in a corresponding *Airline Travel Information System (ATIS)* domain grammar (PortDial Consortium 2013). It is organized around the concept of a trip, which consists of flights, hotel stays and car rentals. Flights in turn are composed of flight legs and are connected to their arrival and departure as well as the operating airline. As an example, Fig. 4 shows



**Fig. 4** Conceptualization of flights with their departure city and operating airline

<sup>1</sup><https://sites.google.com/site/portdial2/>.

```

ClassNoun("flight", :Flight) with plural "flights"
ClassNoun("city", :City) with plural "cities"

StateVerb("operate", :airline,
          propSubj=DirectObject,
          propObj=Subject)

StateVerb("depart", :flightDeparture ◦ :city,
          propSubj=Subject,
          propObj=PrepositionalObject("from"))

```

**Fig. 5** Lexical patterns for the nouns “flight” and “city” as well as the verbs “to depart from” and “to operate”

a small part of the ontology, capturing flights, the city of their departure and their operating airline.

Connected to the ontology is an ontology-lexicon that specifies how the classes, properties and individuals are verbalized in a specific language. The classes `Flight` and `City`, for example, are expressed in English using the nouns “flight” and “city”, while `Departure` is an auxiliary construct that a user would probably not address directly. The latter also holds for both the properties `flightDeparture` and `city`: On their own, they are not relevant to the user, but what is relevant is their composition, connecting flights to the city of their departure. The property chain `flightDeparture ◦ city` can be verbalized as the verb chunk “to depart from” and the verb “to leave” or as the noun chunk “flight from”. A natural verbalization of the property `airline` is by means of the verb “to operate”. Examples for lexical patterns specifying those verbalizations are listed in Fig. 5, using a catalogue of *lemon* design patterns (McCrae and Unger 2014) that was created in order to relieve lexicon engineers from the need to understand and write RDF as well as to support them in the construction of lexical entries. All patterns specify a canonical form (possibly together with additional inflectional forms) as well as a reference to the particular ontology concept they denote. The verb patterns moreover give a mapping from semantic to syntactic arguments: In the case of “to operate”, the subject of the denoted property (a flight) corresponds to the direct object in the syntactic structure, and the object of the denoted property (an airline) corresponds to the syntactic subject, like in “Pan Am operates flight 27B-6”, while in the case of “to depart” the subject of the denoted property chain (a flight) corresponds to the syntactic subject, and the object of the denoted property chain (a city) corresponds to a prepositional object in the syntactic structure, marked with the preposition “from”.

In a similar way, the lexicon specifies alternative verbalizations of the same concepts, such as “to leave from” or “flight from”, as well as all relevant verbalizations of other ontology concepts. In exactly the same way, lexicalizations of instances can be given, for example, verbalizing the individual `Boston` by its name “Boston” and the individual `John_F_Kennedy_International_Airport` as “JFK”.



Once an ontology-lexicon is constructed, it is used for the automatic generation of a domain grammar. For this, we build on *lemongrass*,<sup>2</sup> a Python script for mapping *lemon* lexica to different grammar formats, including Grammatical Framework (GF). GF distinguishes abstract and concrete syntax. The abstract syntax is a type-theoretical framework for specifying abstract concepts in a language-independent manner. These concepts are usually semantic ones, which makes it possible to, for example, use the abstract syntax to represent ontologies (Angelov and Enache 2012). A concrete syntax is a mapping from abstract syntax concepts to linearizations of those concepts in a particular language. Based on the concepts in the ontology, *lemongrass* constructs an abstract syntax, and from the morphosyntactic information specified in the lexicon, *lemongrass* instantiates templates for constructing a corresponding concrete syntax. The result is a domain grammar that, together with the domain-independent expressions from the core grammar, captures phrases like “all flights to Boston” and “the flight is operated by an airline which serves JFK”.

Since the domain conceptualization does not cover any task-relevant concepts, neither the lexicon nor the resulting grammar comprises expressions specific for customer service dialogs. Providing such expressions is the job of the task grammar, for example, specifying constructions for requesting and offering information, as well as dialog constructions such as greetings and expressions for agreement or disagreement, possibly taking into account parameters like formal vs. informal speech.

The final application grammar is then composed of the core grammar, an automatically generated domain grammar for the flight travel domain, and a (for now manually constructed) task grammar for user service dialogs. Combining these three grammar modules, the covered language fragment includes utterances like “give me all flights to Boston” and “which airlines operate flights from Boston to Denver”.

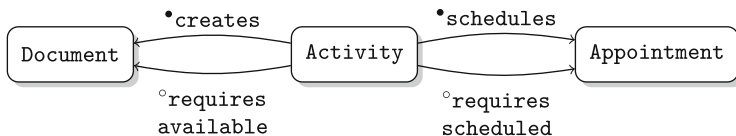
### 3.2.2 Porting to a New Domain

Porting the above dialog application to another domain requires a conceptualization of that new domain, together with lexical information from which a new domain grammar can be generated. Depending on the size and complexity, lexicon creation can be very labour intensive and thus would greatly benefit from semi-automatic methods (Walter et al. 2013) and an ecosystem of resources as described in Sect. 4. Because of the independence of grammar modules, core and task grammars remain unchanged.

We illustrate the domain porting by an example from the business processes domain, centered around activities and their preconditions and postconditions (van Grondelle and Gülpers 2011). Figure 6 shows an instantiation for the particular case

---

<sup>2</sup><https://bitbucket.org/chru/lemongrass>.



**Fig. 6** Conceptualization of activities and their preconditions and postconditions, where *open circle* marks precondition relations and *filled circle* marks postcondition relations

```

: Available          a owl:Class .

: precondition      a owl:ObjectProperty .

: requires          a owl:ObjectProperty ;
  rdfs:subPropertyOf :precondition .

: requires_available a owl:ObjectProperty ;
  rdfs:subPropertyOf :requires ;
  rdfs:domain :Activity ;
  rdfs:range [ owl:intersectionOf (:Document :Available) ].
  
```

**Fig. 7** Definition of the precondition relation *requires available*, based on the general precondition property *requires* and the class *Available*

of housing benefit requests, where relevant activities are, for instance, assessing a request, planning a meeting, or publishing a decision. Preconditions of such activities comprise the availability of some document or a scheduled appointment, while postconditions include the creation of a document, for example, a confirmation or rejection letter.

In the ontology, both preconditions and postconditions are modelled as object properties, with *creates* and *schedules* as subproperties of the postcondition property and *requires* as subproperty of the precondition property. States like *available* and *scheduled* are modelled as classes. The composed precondition relations *requires available* and *requires scheduled* are then defined as properties with a range comprising individuals from the union of, for example, *Document* and *Available*. An example of such a definition is given in Fig. 7.

Similar to the flight travel domain, a corresponding ontology-lexicon specifies how the classes, relations and instances are verbalized. The precondition *requires*, for example, can straightforwardly be expressed using the verb “to require”, as in the following example:

- The assessment of the request requires that the customer visit is scheduled.

An example of a lexicon pattern for this verbalization as well as one for the class *Available* is given in Fig. 8.

Coupling the housing benefit domain grammar with the customer service dialog task used above then allows for the generation of questions and requests such

```

StateVerb("require", :requires)

IntersectiveAdjective("available", :Available)
IntersectiveAdjective("unavailable", :Available-1)

```

**Fig. 8** Lexical patterns for the verb “requires” as well as the intersective adjectives “available” and “unavailable”, where  $Available^{-1}$  denotes the complement class of  $Available$

as “Which documents are required for assessing the request?” and “We need the confirmation letter.”

### 3.2.3 Incorporating New Tasks

Analogously to replacing one domain by another, we can also replace one task by another. For instance, for the purpose of creating explanatory texts, a task grammar could contain constructions for combining fact verbalizations using “because”, “therefore”, “but” and other conjunctions, as well as expressions for putting emphasis on particular aspects. Combining such a documentation task grammar with the housing benefit domain grammar can cover expressions such as the following ones:

- Especially the customer visit is required.
- A confirmation letter was created. Therefore, the activity of assessing the request is completed.

The new task can of course also be combined with the flight travel domain, covering expressions such as the following ones:

- Especially JFK is served by most airlines.
- A flight from Los Angeles to San Francisco takes 1 h. Therefore, there is no meal.

### 3.2.4 Adding Further Languages

Extending an application to other languages requires porting both the lexicalizations and the lexicon-to-grammar mapping.

First, the domain lexicon needs to be ported to the target language. This process can exploit automatic methods for ontology lexicalization (Walter et al. 2013), label translation methods (Mejía et al. 2009; McCrae et al. 2011) and linguistic resources such as BabelNet (Navigli and Ponzetto 2012). Figure 9 shows Dutch versions of the flight travel lexicalizations given in Fig. 5. Since Dutch is very close to English, the lexicalizations only differ in their form and in the specification of gender in the case of nouns.

Second, the lexicon-to-grammar mapping and the core grammar module needs to be ported to the target language. The involved effort strongly depends on the

```

ClassNoun("vlucht", :Flight) commonGender
                               with plural "vluchten"
ClassNoun("stad",   :City)   commonGender
                               with plural "steden"

StateVerb("opereren", :airline,
          propSubj = DirectObject,
          propObj  = Subject)

StateVerb("vertrekken", :flightDeparture o :city,
          propSubj = Subject,
          propObj  = PrepositionalObject("van"))

```

**Fig. 9** Dutch lexical patterns for flight travel concepts

grammar formalism and the multilingual resources available in that formalism. In the case of our implementation using GF, porting a grammar to another language is almost trivial for all languages for which GF provides resource grammars, that is, implementations of low-level morphosyntactic operations. This is the case for about 30 languages from a variety of language families. Mapping the core grammar module to Dutch and German required about ten lines of GF code each, and extending *lemongrass* with templates for additional concrete syntaxes for those languages required a similarly low amount of effort.

The grammar constructed from the Dutch flight travel lexicon, together with the Dutch core and task grammar modules, then covers utterances such as the following ones:

- Toon alle vluchten vanaf Detroit naar Boston.
- Welke luchtvaartmaatschappijen vliegen naar San Francisco?
- Ik wil morgen naar New York reizen.

## 4 An Ecosystem for Language Technology

The architecture presented is extremely modular, both in terms of technologies and resources. This provides new ways of sharing and marketing language technology, as granular components can be developed independently and can then be shared, reused and composed into language technology-based solutions, thereby facilitating an ecosystem of cooperating language technology producers and consumers.

In addition to language resources like declarative lexical resources for domains and tasks, a number of different kinds of components could be shared:

- Generic domain and task conceptualizations
- Technologies to mine and extend lexical resources

- Technology mappings from declarative lexical resources to technology-specific formalisms, such as different grammar formalisms, phrase tables, semantic annotations (Davis et al. 2011), etc.

Being able to reuse technology and lexical resources at a granular level provides nonlinguistically trained developers with a low impact adoption path of language technologies into existing applications and solutions. Initial support for natural language can be added at low cost, as default lexical and grammar resources are available for reuse, as are tools to create and enrich those resources. Optimization and customization can then be performed as expertise grows.

The open standards of the Semantic Web provide a very suitable way to implement such an ecosystem, as it supports the publishing and sharing of resources and services on the web, based on Semantic Web formalisms and tools. Examples are the Linguistic Linked Data (Chiaros et al. 2012) cloud,<sup>3</sup> which forms a growing ecosystem of interlinked language resources such as dictionaries and lexica, and the Language Grid (Murakami et al. 2014) which offers an architecture for sharing and composing language services.

A different way to exploit the modularity of the resources is creating extensible, novel end-user services, as shown in Fig. 10. Imagine a virtual assistant, presumably on a smartphone, that could easily be extended by app developers with new capabilities and that allows consumers to create their own personal virtual assistant supporting services of interest to them and, as a consequence, covering exactly the range of dialog needed for those selected services. The architecture we presented in this chapter could be the basis of a software development kit that allows app

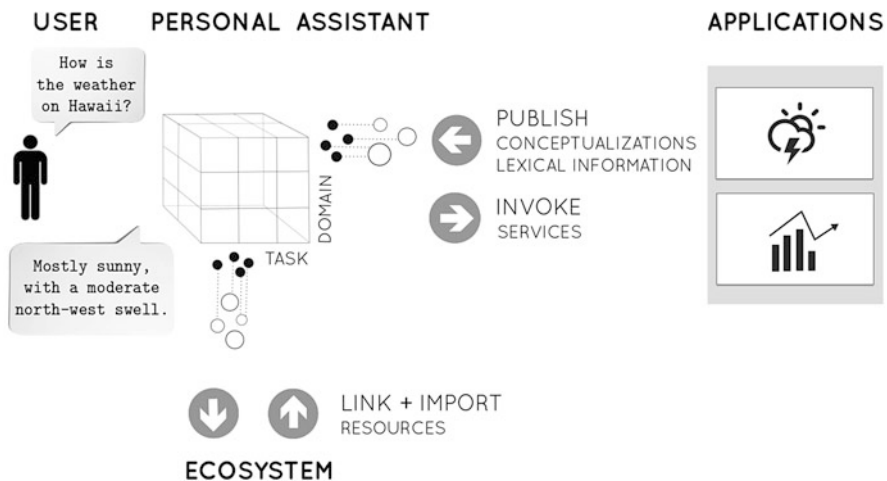


Fig. 10 A virtual assistant as natural language interface to applications

<sup>3</sup><http://linguistics.okfn.org/resources/llod/>.

developers to associate their apps with domain and task conceptualizations and lexicalizations to allow for instance the phone's standard virtual assistant to disclose the app's services using voice dialog. For instance, a weather app could come with a conceptualization and lexicalization of the weather domain, allowing the consumer to query the phone's virtual assistant for the weather situation, possibly using a standard vocabulary for querying.

## 5 Conclusion and Future Work

In order to support the adoption of language technology into existing services and applications, especially by companies with little or no linguistic expertise, we proposed a new paradigm for the creation and use of language technology resources. Starting from a conceptualization that scopes the supported language fragment to exactly those expressions and constructions relevant for the application in question, we exploited declarative lexical information for specifying verbalizations of concepts. On both levels, conceptual and lexical, we clearly separated domain and task aspects. Further, lexical representations served as input for the automatic generation of language technology resources, thereby removing both the need for expertise in specific formats and the dependence on particular implementations of them.

As proof of concept, we provided an implementation based on Semantic Web standards, creating GF grammars from *lemon* lexicalizations attached to an underlying OWL conceptualization of a domain, showing that it supports typical adoption scenarios.

A limitation to be addressed in future work is that in the given implementation, task grammars were still constructed manually. This mirrors the fact that the conceptualizations and lexica already present on the web so far mainly focus on domains, whereas the task that is supported is often implicitly assumed to be querying. Conceptualization of other tasks as well as multilingual lexical information for verbalizing them are still widely lacking. We therefore aim at a general conceptualization of different tasks and, if necessary, an extension of the *lemon* model for task verbalizations.

Furthermore, we plan to explore how the proposed paradigm can be applied to other areas of language technology, for example, generating phrase tables for machine translation, possibly building on the same Semantic Web standards for conceptualizations and lexicalizations.

This will lift the proposed three-dimensional architecture to its full potential, enabling the reuse of multilingual lexical resources for domains and tasks across the web and allowing the application of these resources in a wide range of language technologies, moving towards an ecosystem for language technology.

**Acknowledgements** This work was partially funded in the EU projects PortDial (FP7-296170), Monnet (FP7-248458) and MOLTO (FP7-247914). We also want to thank the organizers of the Dagstuhl seminar, where many of the ideas in this chapter took form, especially Philipp Cimiano for numerous invaluable discussions. We are also indebted to Aarne Ranta, Jouri Fledderman and Frank Smit.

## References

- Angelov, K., & Enache, R. (2012). Typeful ontologies with direct multilingual verbalization. In M. Rosner & N. E. Fuchs (Eds.), *Controlled natural language. Lecture Notes in Computer Science* (Vol. 7175, pp. 1–20). New York: Springer.
- Chiarcos, C., Nordhoff, S., & Hellmann, S. (Eds.). (2012). *Linked data in linguistics: Representing and connecting language data and language metadata*. New York: Springer.
- Davis, B., Badra, F., Buitelaar, P., Wunner, T., & Handschuh, S. (2011). Squeezing lemon with GATE. In *Proceedings of the 2nd International Workshop on the Multilingual Semantic Web (MSW 2011), Workshop at the 10th International Semantic Web Conference (ISWC 2011)*.
- Gatius, M., & Rodríguez, H. (1996). A domain-restricted task-guided natural language interface generator. In *Proceedings of the Second Edition of the Workshop Flexible Query Answering Systems (FQAS'96)*.
- Guarino, N. (1997). Understanding, building and using ontologies. *International Journal of Human-Computer Studies*, 46(2–3), 293–310.
- Kaljurand, K., & Alumäe, T. (2012). Controlled natural language in speech recognition based user interfaces. In *Controlled natural language. Lecture Notes in Computer Science* (Vol. 7427, pp. 79–94). New York: Springer.
- Martins, A. F., & de Almeida Falbo, R. (2008). Models for representing task ontologies. In F. Freitas, H. Stuckenschmidt, S. Pinto, A. Malucelli, & O. Corcho (Eds.), *Proceedings of the 3rd Workshop on Ontologies and Their Applications (WONTO 2008)*.
- McCrae, J., de Cea, G. A., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., et al. (2012). Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 46(4), 701–719.
- McCrae, J., Espinoza, M., Montiel-Ponsoda, E., Aguado-de Cea, G., & Cimiano, P. (2011). Combining statistical and semantic approaches to the translation of ontologies and taxonomies. In *Proceedings of the Fifth Workshop on Syntax, Structure and Semantics in Statistical Translation (SSST-5)* (pp. 116–125).
- McCrae, J., & Unger, C. (2014). Design patterns for engineering the ontology-lexicon interface. In P. Buitelaar & P. Cimiano (Eds.), *Towards the multilingual semantic web: Principles, methods and applications*. Heidelberg: Springer. doi:10.1007/978-3-662-43585-4.
- McGuinness, D., & Van Harmelen, F. (2004). OWL web ontology language overview. *W3C Recommendation*, 10, 2004–03. <http://www.w3.org/TR/owl-features/>
- Mejía, M. E., Montiel-Ponsoda, E., & Gómez-Pérez, A. (2009). Ontology localization. In *Proceedings of the Fifth International Conference on Knowledge Capture (K-CAP 2009)* (pp. 33–40).
- Mizoguchi, R., Tijerino, Y., & Ikeda, M. (1995). Task analysis interview based on task ontology. *Expert Systems with Applications*, 9(1), 15–25.
- Montiel-Ponsoda, E., Aguado de Cea, G., Gómez-Pérez, A., & Peters, W. (2008). Modelling multilinguality in ontologies. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)* (pp. 67–70).
- Murakami, Y., Lin, D., & Ishida, T. (2014). Service oriented architecture for interoperability of multi-language services. In P. Buitelaar & P. Cimiano (Eds.), *Towards the multilingual semantic web: Principles, methods and applications*. Heidelberg: Springer. doi:10.1007/978-3-662-43585-4.

- Navigli, R., & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193, 217–250.
- Nirenburg, S., & Raskin, V. (2004). *Ontological semantics*. Cambridge: MIT Press.
- PortDial Consortium. (2013). D2.1 Free Data Deliverable. <https://sites.google.com/site/portdial2/deliverables-publications/free-data-deliverable>.
- Ranta, A. (2011). *Grammatical framework: Programming with multilingual grammars*. Stanford: CSLI Publications.
- Reymonet, A., Thomas, J., & Aussenac-Gilles, N. (2007). Modelling ontological and terminological resources in OWL DL. In *Proceedings of the OntoLex07 Workshop at the 6th International Semantic Web Conference (ISWC 2007)*.
- Spreeuwenberg, S., & Healy, K. A. (2010). SBVR's approach to controlled natural language. In *Proceedings of the Workshop on Controlled Natural Language (CNL 2009)* (pp. 155–169).
- Spreeuwenberg, S., van Grondelle, J., Heller, R., & Grijzen, G. (2012). Using CNL techniques and pattern sentences to involve domain experts in modeling. In *Proceedings of the Workshop on Controlled Natural Language (CNL 2010)* (pp. 175–193).
- van Grondelle, J., & Gülpers, M. (2011). Specifying flexible business processes using pre and post conditions. In *Practice of enterprise modeling. Lecture Notes in Business Information Processing* (Vol. 92, pp. 38–51). Heidelberg: Springer.
- Walter, S., Unger, C., & Cimiano, P. (2013). A corpus-based approach for the induction of ontology lexica. In *Proceedings of the 18th International Conference on Application of Natural Language to Information Systems (NLDB 2013)*.
- Wróblewska, A., Protaziuk, G., Bembenik, R., & Podsiadły-Marczykowska, T. (2012). LEXO: A lexical layer for ontologies—design and building scenarios. *Studia Informatica*, 33(2B), 173–186. <http://studiainformatica.polsl.pl/index.php/SI/article/view/183>.
- Zue, V. W., & Glass, J. R. (2000). Conversational interfaces: Advances and challenges. *IEEE Special Issue on Spoken Language Processing*, 88(8), 1166–1180.



# Towards Verbalizing Multilingual N-Ary Relations

Yan Tang Demey and Clifford Heath

**Abstract** The idea of a Multilingual Semantic Web is to provide access to knowledge available on the Semantic Web (SW) to speakers of different languages. In this chapter, we concentrate on a particular aspect of the Multilingual Semantic Web vision and discuss the challenge of *multilingual verbalization* for conceptual models. Natural verbalizations require n-ary fact types, whereas popular Description Logic (DL) dialects [such as those used by the Web Ontology Language (OWL)] only support binary fact types. We use a Fact-Based Modelling (FBM) approach because it supports n-ary verbalization. We also discuss the use of natural language taggers in the creation of these models, which preserves the natural form of those verbalization patterns. Patterns that are *typical* and *representative* are studied in English and Chinese. In order to publish such models to the SW, we use *objectification* (i.e. model reification) to transform n-ary fact types to a binary form.

**Key Words** Conceptual modelling • Fact-based modelling • Object role modelling • Ontology verbalization

## 1 Introduction and Motivation

Multilingualism is a common phenomenon on the Internet and the web of data, where almost all websites of international companies provide information in at least their local language (e.g. Chinese, French, German, etc.) and English. This is especially appealing for a country like Belgium, the official languages of which are French, Dutch and German. It is also needed in knowledge intensive websites, such as Wikipedia. Like the Web, the Semantic Web (SW) also needs to address *multilingualism*.

---

Y.T. Demey (✉)

Department of Computer Science, Vrije Universiteit Brussel, Brussels 1050, Belgium  
e-mail: [yan.tang.demey@gmail.com](mailto:yan.tang.demey@gmail.com)

C. Heath

Data Constellation, 2/69 Hill Street, Roseville, Sydney, NSW 2069, Australia  
e-mail: [cjh@dataconstellation.com](mailto:cjh@dataconstellation.com)

There are four challenges for multilingual web of data (Gracia et al. 2012), namely, (1) *ontology localization*, which is the process of adapting an ontology to the needs of a particular (linguistic and cultural) community (e.g. in León-Araúz and Faber 2014, this volume); (2) *cross-lingual mapping*, which is the process of establishing the link between the levels of conceptual, linguistic and instance (e.g. in Gromann and Declerck 2014, this volume); (3) *representation of multilingual lexical information*, which consists in the reification of labels with different degrees of expressivity; and (4) *cross-lingual access and querying of Linked Data*, which is realized by a querying mechanism based on mappings between vocabulary elements in different languages.

In this chapter, we discuss a fifth challenge called *multilingual verbalization* for conceptual models. In particular, Fact-Based Modelling (FBM) languages<sup>1</sup> and the Web Ontology Language (OWL) will be used as the modelling means. FBM usually applies the Conceptual Schema Development Procedure (CSDP, Halpin and Morgan 2008) to create models, by discussing natural sentences that express concrete situations in a domain and generalizing these to form a conceptual model. In this chapter, we explore instead the use of natural language processing (NLP) taggers to create the conceptual models and translate them into sentences in a natural language.

The chapter is organized as follows. Section 2 is the chapter background and the related work. We will illustrate verbalization patterns in English and Chinese in Sect. 3 and use objectification to map n-ary fact types into binary fact types to allow publishing n-ary fact types with OWL. In Sect. 4, we will discuss the issues of implementation and present the discussion. Section 5 concludes the chapter.

## 2 Background and Related Work

For about four decades, FBM dialects, such as Object Rule Modelling language (ORM, Halpin and Morgan 2008), Developing Ontology-Grounded Methods and Applications (DOGMA, Spyns et al. 2008), Natural language Information Analysis Method (NIAM, Nijssen and Halpin 1989), Cognition-enhanced NIAM (CogNIAM, Nijssen and Lemmens 2008) and Fully Communication-Oriented Information Modelling (FCO-IM, Bakema et al. 2002), have been intensively studied for modelling information and knowledge. A methodological principle of an FBM language is to extract information from *plausible facts* in a given domain by adhering to the *Conceptualization and 100 % Principles* of ISO TR9007 (Jardine 1984).

FBM methodological principles have been used since 1999 for modelling ontologies and to support *verbalization* of ontology models. Verbalization is an unambiguous mapping between a conceptual model and a finite set of sentences

---

<sup>1</sup>FBM initiative: <http://www.factbasedmodeling.org/>.

in a controlled language<sup>2</sup> (such as a pseudo natural language). Jarrar (2005) and Demey et al. (2002) have illustrated how a particular FBM dialect—ORM—can be used for modelling ontologies and ontology verbalization. More recently, ORM has been extended for modelling ontology-based rules. One extension called Semantic Decision Rule Language (Tang and Meersman 2008) is used for modelling semantically rich decision support rules and business rules. Its markup language—SDRule-ML—has been designed to store and exchange ontology-based business rules.

A few researchers have studied the multilingual aspect in FBM. Jarrar presented a flexible, extensible and maintainable engineering solution to verbalize multilingual ORM schemas (see Jarrar 2005). Constellation Query Language (CQL, Heath 2009) offers an alternative solution that can be used to represent almost any ORM model in plain text using a natural language, with the goal of supporting direct involvement by all business stakeholders.

In this chapter, we use ORM<sup>3</sup> to model domain ontologies and to verbalize them. Unlike most of the work in the FBM community, which only deals with verbalization using individual sentences, we extend verbalization by identifying a set of verbalization patterns that recur in a particular domain. It is common practice in FBM to generate the same verbalization pattern (type of sentence in a controlled language) every time a given conceptual pattern occurs. However, in our proposed approach to extracting conceptualizations from source text, it is necessary to recognize multiple possible ways of describing the same state of affairs. By doing so, we give the freedom to the end user for processing the documents using their preferred manner of expression.

Another important point regards parsing informal texts as provided by domain experts using verbalization patterns to generate a conceptual model. Such texts often contain general statements. FBM usually applies the Conceptual Schema Development Procedure (CSDP, Halpin and Morgan 2008) to create conceptual models by generalizing from concrete statements and conceptualizing the general statements. The concrete statements are either produced and validated by a face-to-face communication between a domain expert and a modeller or by asking a modeller to propose statements through manual analysis of informal texts and subsequently to validate them by negotiation. The basis of analysis may change during negotiation, which requires iteration. Our approach reduces the iteration and tedious manual work of the modeller by implementing a semiautomated process for analysing source texts.

---

<sup>2</sup>A controlled natural language is a language that a computer can process without any extra information.

<sup>3</sup>The ORM family contains ORM (the initial ORM) and ORM2 (the second generation of ORM). Including the update of graphical notations, ORM2 also contains a few extensions to the initial ORM, such as constraint modality and derivation rules. In this chapter, we use the term “ORM” for the whole ORM family.

Gangemi and Presutti (2010) have defined a *pattern* as *invariances across observed data or objects*. In this sense, we observe that a verbalization pattern is comparable to a *design pattern* (Alexander et al. 1977), which is a reusable solution to a recurrent modelling problem. Ontology design patterns (ODPs, Gangemi and Presutti 2009) are a kind of design patterns used to capture an arrangement of conceptual design elements that is recurrent within a domain. The *ontologydesign-patterns.org* initiative maintains a repository of ODPs. The verbalization patterns that we illustrate in this chapter can be considered as ODPs.

We have noted the fact that ORM models can contain ternary, quaternary and other  $n$ -ary ( $n \geq 3$ ) fact types. Some ORM extension languages, such as SDRule-L, support only unary and binary fact types, because (1) most ontology modelling languages (e.g. RDF and OWL) support only unary and binary fact types, (2) there is still a debate on the stability of  $n$ -ary fact types in a model and (3) it simplifies mechanisms of traversing conceptual models. With binary fact types, we can easily roll back to previous model views. While traversing a model containing lots of  $n$ -ary fact types, it is easy to lose the track. Nevertheless, due to prevalence of  $n$ -ary fact types in natural verbalization, we wish to retain them, but to do that while utilizing ontology tools which do not support them raises the following research questions:

- How to generate conceptual models from multilingual  $n$ -ary facts?
- How to formalize  $n$ -ary fact types in DL (Baader et al. 2010)?
- How to map constrained  $n$ -ary fact types to constrained binary fact types so that they can be stored and published in OWL?

In order to answer the first question, we address the multilingualism principle of FBM for *ontology modelling* and *verbalization* by presenting a few verbalization patterns that are *typical* and *representative* in a *multilingual* business domain.

The second question may be answered by proposing a DL dialect. In particular, an extension to *ALCRP* ( $\mathcal{D}$ ) (Haarslev et al. 1999) may be useful. Not much work has been done concerning formalizing  $n$ -ary models using DL. A thorough discussion could be lengthy, so we will not focus on this topic in this chapter.

We use *objectification* (i.e. model reification) for transforming patterns from  $n$ -ary fact types to binary fact types in order to answer the third research question.

### 3 Verbalization Patterns

#### 3.1 Verbalization Patterns and Linguistic Patterns

Verbalization is a linguistic grounding process, which covers the following two tasks:

- Task of *interpretation*: Textual information in a controlled language, which is easily understood by a non-technical domain expert, is used to interpret a model designed by a modeller. The goal of interpretation is to achieve an easy

communication during the phases of ontology creation, validation and evolution. A related work is NORMA (Halpin and Curland 2006).

- Task of *conceptualization*: It is a process of extracting conceptual models from sentences in a controlled language for discovering constrained facts. Lexico-Syntactic Patterns (LSPs, Hearst 1992) and LSPs for conveying the conceptual relations formalized in ODPs (de Cea et al. 2008) are the related work.

We propose extra steps before conceptualization and after interpretation. The extra step before conceptualization is to extract basic verbalization patterns from the informal texts provided by a domain expert to a modeller. The extra step after interpretation is to regenerate natural text from the conceptual model aiming to use a reading style consistent with the domain expert's preferences.

We do not expect to automate the whole modelling task, because that is almost the same unsolved problem as general artificial intelligence. Instead, we take a semiautomatic approach, in which we create ontological models based on a set of plausible facts proposed following automated analysis of texts.

If necessary, we first use "classical" brainstorming and knowledge elicitation techniques (Schreiber et al. 1999) to manually get facts in a domain. Such facts are, for example, "Yan booked seat28". Then, we group instances into types. For instance, the type of "Yan" is "Visitor" and "seat28" is "Seat". In the meantime, verbalization patterns are manually discovered either from documents or from a large number of informal textual data. For instance, common nouns like "cinema" or "people" often refer to object types; labels like "Yan" or "VUB" often refer to objects. The discovered patterns are iteratively applied to the sentences of the source texts to help find patterns which had not yet been discovered.

Discovery of patterns can be assisted using NLP techniques. One of such techniques is called Part-Of-Speech (POS) tagging and is often used for syntactical analysis. The Stanford Tagger<sup>4</sup> (Toutanova and Manning 2000; Toutanova et al. 2003) is a Maximum Entropy Part-of-Speech Tagger. It uses feature-rich models with extensive lexicalization, bidirectional inference and effective regularization algorithms. In this chapter, we adopt the Stanford Tagger seeing that it provides many advantages: (1) Its templates are expressive enough for our purpose; (2) compared to the ones in (Marshall 1987; Collins 2002), the Stanford Tagger has the best per-position-tag accuracy and highest whole-sentence correct rate; and (3) it is loosely coupled with tagger models; it is possible to retrain it for other languages.

It is not our aim to show the exact syntax and semantics of our controlled English or Chinese, but rather how particular verbalization patterns are designed. Figure 1 contains the English patterns using the Penn Treebank set (Marcus et al. 1993). We use OWL DL for indexing them; for example, `owl:Class` is used to index the patterns for the verbalization pattern called EO. Each set of patterns is presented in three parts, being pattern names, patterns and examples. The symbol "|" is used for separation.

---

<sup>4</sup>The tool is available at <http://nlp.stanford.edu/downloads/tagger.shtml> (last retrieval date: January 9, 2014).

<b>Pattern 1: EO</b> (elementary object) <b>owl:Class</b>
Common noun ( <i>NN, NNS</i> ) Proper noun ( <i>NNP, NNPS</i> ) Pronoun ( <i>PRP, WP</i> )
<i>Examples:</i> cinema ( <i>NN</i> ) films ( <i>NNS</i> ) Brussels ( <i>NNP</i> ) it ( <i>PRP</i> ) what ( <i>WP</i> )
<b>Pattern 2: DTP</b> (data type property, fact type) <b>rdf:Property</b>
<b>EO</b> + possessive ( <i>POS</i> ) + <b>EO</b>   Possessive pronoun ( <i>PRP\$</i> ) + <b>EO</b>   <b>EO</b> + <u>of</u> <sub>6</sub> + <b>EO</b>   <b>EO</b> + <u>from</u> + common noun   <b>EO</b> contains / belongs to / has / has part / is part of + <b>EO</b>
<i>Examples:</i> visitor's address ( <i>NN+POS+NN</i> ), its colour ( <i>PRP\$+NN</i> ), number of seat ( <i>NN</i> + of + <i>NN</i> ), row contains chair ( <i>EO</i> + contains + <i>EO</i> )
<b>Pattern 3: OP</b> (object property, binary fact type) <b>owl:ObjectProperty</b>
<b>EO</b> +(verb( <i>VB, VBD, VBG, VBN, VBP, VBZ</i> )/preposition or subordinating conjunct.( <i>IN</i> ))*+ <b>EO</b> <sub>7</sub>
<i>Examples:</i> person booked seat ( <i>NN</i> + <i>VBN</i> + <i>NN</i> ), seat is in row ( <i>NN</i> + <i>VBZ</i> + <i>IN</i> + <i>NN</i> )
<b>Pattern 4: S</b> (subtype) <b>rdfs:subClassOf</b>
<b>EO</b> + <u>is</u> + a / <u>an</u> / subtype of / subclass of + <b>EO</b>   <b>EO</b> + <b>EO</b> + ... + coordinating conjunction ( <i>CC</i> ) + <b>EO</b> + <u>are types of</u> + <b>EO</b>   <b>EO</b> + <u>is classified into</u> + <b>EO</b> + ... + coordinating conjunction + <b>EO</b>   <u>each</u> + <b>EO</b> + <u>is an instance of</u> + <b>EO</b>
<i>Examples:</i> Event is classified into concert, film, exposition, theatre or festival. ( <i>EO</i> + <u>is classified into</u> + <i>EO</i> + <i>EO</i> + <i>EO</i> + <i>EO</i> + <i>CC</i> + <i>EO</i> )
<b>Pattern 5: U</b> (uniqueness constraint/ property) <b>owl:maxCardinality</b>
<u>each</u> + <b>EO</b> + (verb/ preposition or subordinating conjunction)* + <u>at most one</u> + <b>EO</b>
<i>Examples:</i> Each booking is (for) at most one seat (each + <i>NN</i> + <i>VBZ</i> + at most one + <i>NN</i> ).
<b>Pattern 6: M</b> (mandatory constraint) <b>owl:minCardinality</b>
<u>each</u> + <b>EO</b> + (verb/ preposition or subordinating conjunction)* + <u>at least one</u> + <b>EO</b>   <b>EO</b> + <u>must</u> + (verb/ preposition or subordinating conjunction)* + <b>EO</b>   <u>it is necessary/mandatory that</u> + <b>OP</b>
<i>Examples:</i> Each booking is (for) at least one seat (each + <i>NN</i> + <i>VBZ</i> + at most one + <i>NN</i> )   Booking must have Seat ( <i>NN</i> + <u>must</u> + <i>VB</i> + <i>NN</i> ).
<b>Pattern 7: SS</b> (subset constraint) <b>owl:someValuesFrom</b> + <b>rdfs:subPropertyOf</b>
<u>if</u> + <b>OP</b> + <u>then</u> + <b>OP</b>   <u>if</u> + determiner ( <i>DT</i> ) + <b>EO</b> <sub>1</sub> + verb + determiner + <b>EO</b> <sub>2</sub> + <u>then</u> + determiner + <b>EO</b> <sub>1</sub> + verb + determiner + <b>EO</b> <sub>3</sub> <sup>8</sup>   <b>OP</b> + <u>implies</u> + <b>OP</b>
<i>Examples:</i> If a booking has allocated a seat, then that booking involves some showing (if + <i>DT</i> + <i>NN</i> + <i>VBZ</i> + <i>VBN</i> + <i>DT</i> + <i>NN</i> + then + <i>DT</i> + <i>NN</i> + <i>VBZ</i> + <i>DT</i> + <i>NN</i> ).
<b>Pattern 8: E</b> (equality constraint) <b>EquivalentObjectProperties</b>
<u>for each</u> + <b>EO</b> <sub>1</sub> + <u>this/that</u> + <b>EO</b> <sub>1</sub> + (verb/ preposition or subordinating conjunction)* + determiner + <b>EO</b> <sub>2</sub> + <u>if and only if</u> + <u>this/that</u> + <b>EO</b> <sub>1</sub> + (verb/ preposition or subordinating conjunction)* + determiner + <b>EO</b> <sub>3</sub>   <b>OP</b> <sub>1</sub> + <u>implies</u> + <b>OP</b> <sub>2</sub> + <u>and</u> + <b>OP</b> <sub>2</sub> <u>implies</u> + <b>OP</b> <sub>1</sub>
<i>Examples:</i> For each seat, that seat has some number if and only if that seat is in some row (for each + <i>NN</i> + that + <i>NN</i> + <i>VBZ</i> + <i>DT</i> + <i>NN</i> + if and only if + <i>DT</i> + <i>NN</i> + <i>VBZ</i> + <i>IN</i> + <i>DT</i> + <i>NN</i> ).
<b>Pattern 9: OP3</b> (ternary fact type) <b>no direct support from OWL DL</b>
<b>EO</b> + verb* + preposition or subordinating conjunction+ <b>EO</b> + preposition or subordinating conjunction + <b>EO</b>   <b>EO</b> + verb* + <b>EO</b> + preposition or subordinating conjunction + <b>EO</b>
<i>Examples:</i> Film is showing on Time at Cinema ( <i>NN</i> + <i>VBZ</i> + <i>VBG</i> + <i>IN</i> + <i>NN</i> + <i>IN</i> + <i>NN</i> ).
<b>Pattern 10: OPN</b> (nary, $n \geq 3$ ) <b>no direct support from OWL DL</b>
<b>EO</b> + ((and/verb/preposition or subordinating conjunction)* + <b>EO</b> ) * + (and/verb/subordinating conjunction)* + <b>EO</b>   <b>EO</b> + verb* + ( <b>EO</b> + preposition or subordinating conjunction)* + <b>EO</b>
<i>Examples:</i> Film is showing on Time at Cinema using Projector and watched by Visitor ( <i>NN</i> + <i>VBZ</i> + <i>VBG</i> + <i>IN</i> + <i>NN</i> + <i>IN</i> + <i>NN</i> + <i>IN</i> + <i>NN</i> + <i>VBG</i> + <i>NN</i> + <u>and</u> + <i>VBN</i> + <i>IN</i> + <i>NN</i> ).

**Fig. 1** English verbalisation patterns (ordered by pattern names, patterns and examples). Note that due to the limit of the paper length, only the patterns that are used in our examples are illustrated. An interesting related work is OntoLT (<http://olp.dfki.de/OntoLT/OntoLT.htm>). <sup>a</sup>The underlined words are the reserved keywords. <sup>b</sup>The symbol \* means that the particular part might be repeated and/is a function of selection. <sup>c</sup>We use subscripts to indicate how a particular part is repeated in a pattern

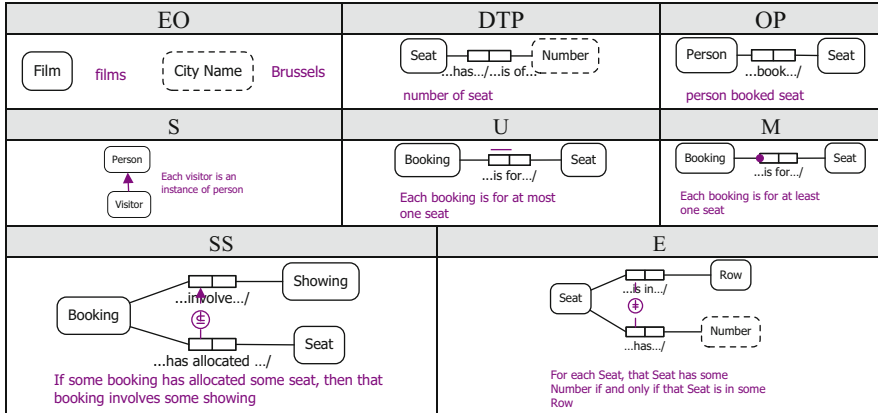
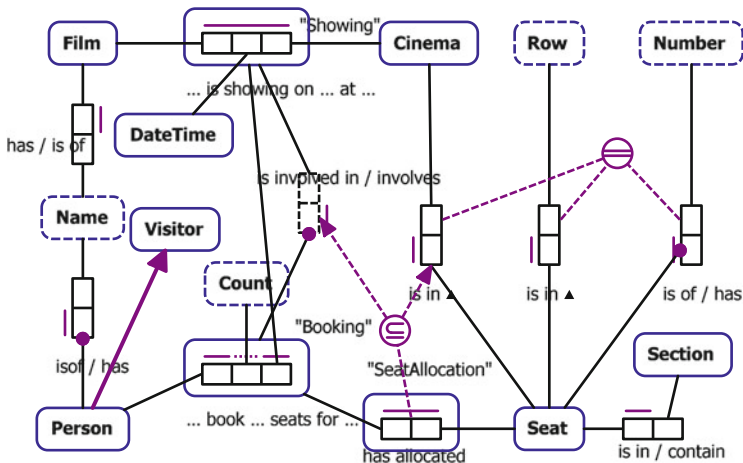


Fig. 2 ORM examples for the patterns from Fig. 1

Once the sentences are unified into those in controlled English, we can use the following method to create conceptual models:

- **EO**: Transform plural nouns to singular and use them to indicate nonlexical object types (NOLOT, Halpin and Morgan 2008); map proper nouns into lexical object types (LOT) and define them as instances of these types; determine types of pronouns; when transforming plural nouns to singular in other patterns, the associated statements must be reworded, for example, “all people have names” must be refined and singularized as “each person has a name”.
- **DTP**: Determine types of possessive pronoun and add properties.
- **OP**: Use the base form of verbs as object properties.
- **S**: Find a subclass relation between two elementary object types and use a subtype relation to connect them.
- **U**: Find a unique role that an object type can play; add a fully spanned uniqueness constraint as the default uniqueness if no unique role is specified.
- **M**: Find a mandatory role that an object type must play; add both uniqueness and mandatory when specifying an identifier (a more complex approach is required for multirole identifiers).
- **SS**: Discover an object type of which the instances that play one role (or group of roles) are a subset of the instances which play another role (or another compatible group of roles). A subset constraint connects at least two fact types.
- **E**: Discover an object type of which the instances that play one role (or group of roles) are equivalent to the instances of those which play another role (or group of roles). It connects at least two fact types, which must share one object type. Note that an equality constraint must involve two non-mandatory roles. Otherwise, though the equality is implicitly present, it is always preferred to conceptualize this as mandatory.

Figure 2 shows the examples using the method discussed above.



**Fig. 3** An ORM model containing ternary fact types (*left*: ORM2 model; *right*: combined verbalization patterns)

An entity type (e.g. “Film” from **EO** in Fig. 2) is graphically represented as a named, solid rectangle with rounded corners. A value type (e.g. “City Name” from **EO** in Fig. 2) is represented as a named, broken rectangle with rounded corners. Relations between object types (also called *roles*) are shown as boxes (see “... has ... / ... is of ...” from **DTP** in Fig. 2). An arrow-tipped bar indicates the “is-a” relationship (see pattern **S** in Fig. 2). A bar above a role suggests a uniqueness constraint (see pattern **U** in Fig. 2). A mandatory constraint is indicated using a dot on the line that connects an object type and a role (see pattern **M** in Fig. 2). An arrow-tipped bar with a circled symbol  $\subseteq$  indicates a subset relationship between two roles (see pattern **SS** in Fig. 2). A circled symbol  $=$  indicates an equivalence relationship between the connected two roles (see pattern **E** in Fig. 2).

The basic patterns of n-ary fact types are illustrated as OP3 and OPN in Fig. 1. They can be combined with the patterns of constraint (U, M, SS and E). Figure 3 shows an example that contains ternary fact types and other patterns from Fig. 1.

Note that in Fig. 3, the round cornered rectangles “Showing”, “Booking” and “Seat Allocation” are used to indicate *objectification*, which is a means of turning a role pair into an object type and which we will discuss in detail later in this section. Figure 3 also contains examples of combined verbalization patterns in the model.

With regard to the Chinese verbalization patterns, we have extended the LDC Chinese Treebank POS tag set (Xia 2000) for Chinese taggers as illustrated in Table 1.

On top of the English patterns, we add a new pattern called single role (SR) that can be implemented using **owl:ObjectProperty**. Unlike the English patterns, in which roles can be abstracted from verbs and preposition or subordinating conjunction (e.g. “drive” and “look for”), roles in Chinese patterns can be:



**Table 1** Chinese verbalization patterns

<b>EO</b>	Proper noun ( <i>NR</i> )/[personal/demonstrative/possessive pronoun or anaphora ( <i>PN</i> )]/other noun ( <i>NN</i> )/verb* ( <i>VC, VE, VV</i> ) + maker ( <i>DEC</i> )  <b>EO</b> * <i>Examples:</i> 影院( <i>NN</i> ), 我( <i>PN</i> ), 布鲁塞尔( <i>NR</i> ), 吃的( <i>food</i> ) ( <i>VV</i> + <i>DEC</i> )
<b>SR</b> (single role)	verb* + aspect particle ( <i>AS</i> )  verb* + 到/予/于  Localizer ( <i>LC</i> ) <i>Examples:</i> 预订( <i>VV</i> ), 是( <i>VC</i> ), 有( <i>VE</i> ), 红了( <i>VA+AS</i> ),涉及到( <i>VV</i> + 到)
<b>DTP</b>	<b>EO</b> + genitive or associative maker ( <i>DEG</i> ) + <b>EO</b> <i>Examples:</i> 访客的地址( <i>NN</i> + <i>DEG</i> + <i>NN</i> )
<b>OP</b>	<b>EO</b> + <b>SR</b> + <b>EO</b> <i>Examples:</i> 访客预订座位( <i>NN</i> + <i>VV</i> + <i>NN</i> )
<b>S</b>	<b>EO</b> + 的子类是 + <b>EO</b> /( <b>EO</b> + coordinating conjunction ( <i>CC</i> ) + <b>EO</b> )/( <b>EO</b> + ... + <b>EO</b> + coordinating conjunction + <b>EO</b> ) ( <b>EO</b> + coordinating conjunction + <b>EO</b> )/( <b>EO</b> + ... + <b>EO</b> + coordinating conjunction + <b>EO</b> ) + 归类于 + <b>EO</b>   <b>EO</b> + 分成 + cardinal number ( <i>CD</i> )/determiner ( <i>DT</i> ) + 类; + <b>EO</b> + ... + coordinating conjunction ( <i>CC</i> ) + <b>EO</b>   <b>EO</b> /( <b>EO</b> + coordinating conjunction + <b>EO</b> )/( <b>EO</b> + ... + <b>EO</b> + coordinating conjunction + <b>EO</b> ) + 是 + <b>EO</b> + 的子类 <i>Examples:</i> 访客是人的子类( <i>NN</i> + 是 + <i>NN</i> + 的子类)   文化活动分成五类:音乐会,电影,展览,舞台剧和节庆( <i>NN</i> + 分成 + <i>CD</i> + 类; + <i>NN</i> + <i>NN</i> + <i>NN</i> + <i>NN</i> + <i>CC</i> + <i>NN</i> )
<b>U</b>	每一个 + <b>EO</b> + 最多/至多 + <b>SR</b> + 一个 + <b>EO</b> <i>Examples:</i> 每一个预订最多订一个位子(每一个 + <i>NN</i> + 最多 + <i>VV</i> + 一个 + <i>NN</i> )
<b>M</b>	每一个 + <b>EO</b> + 需最少/需至少/最少/至少 + <b>SR</b> + 一个 + <b>EO</b>   每一个 + <b>EO</b> + 必须 + <b>SR</b> + <b>EO</b> <i>Examples:</i> 每一个预订最少订一个位子(每一个 + <i>NN</i> + 最少 + <i>VV</i> + <i>NN</i> )
<b>SS</b>	假如 + <b>OP</b> + 那么 + <b>OP</b>   假如 + cardinal number/determiner ( <i>DT</i> ) + measure word ( <i>M</i> ) + <b>EO</b> <sub>1</sub> + <b>SR</b> + cardinal number/determiner + measure word + <b>EO</b> <sub>2</sub> + 那么 + determiner + measure word + <b>EO</b> + <b>SR</b> + measure word + <b>EO</b>   <b>OP</b> 意味着 <b>OP</b> <i>Examples:</i> 假如一个预订配备了一个位子,那么这个预订涉及到某个表演(假如 + <i>CD</i> + <i>M</i> + <i>NN</i> + <i>VV</i> + <i>X</i> + <i>CD</i> + <i>M</i> + <i>NN</i> + 那么 + <i>DT</i> + <i>NN</i> + <i>VV</i> + <i>CC</i> + <i>DT</i> + <i>NN</i> )
<b>E</b>	对每个 + <b>EO</b> <sub>1</sub> + 来说 + determiner + measure word + <b>EO</b> <sub>1</sub> + <b>SR</b> + determiner + measure word + <b>EO</b> <sub>2</sub> + 当且仅当 + determiner + measure word + <b>EO</b> <sub>1</sub> + <b>SR</b> + determiner + measure word + <b>EO</b> <sub>3</sub> <i>Examples:</i> 对每个位子来说,那个位子有某个号码,当且仅当那个位子坐落于某个排.(对每个 + <i>NN</i> + 来说 + <i>DT</i> + <i>M</i> + <i>NN</i> + <i>VE</i> + <i>DT</i> + <i>M</i> + <i>NN</i> + 当且仅当 + <i>DT</i> + <i>M</i> + <i>NN</i> + <i>VV</i> + 于 + <i>DT</i> + <i>M</i> + <i>NN</i> )
<b>OP3</b>	<b>EO</b> + localizer ( <i>LC</i> )/verb* + <b>EO</b> + localizer/verb* + <b>EO</b> + verb*  <b>EO</b> + localizer/verb* + <b>EO</b> + verb* + <b>EO</b> <i>Examples:</i> 影片于时间在影院放映.( <i>NN</i> + <i>LC</i> + <i>NN</i> + <i>LC</i> + <i>NN</i> + <i>VV</i> )
<b>OPN</b>	<b>EO</b> + (localizer/verb* + <b>EO</b> )* + verb*  <b>EO</b> + (localizer/verb* + <b>EO</b> )* + verb* + <b>EO</b> <i>Examples:</i> 人于时间在影院看电影( <i>NN</i> + <i>LC</i> + <i>NN</i> + <i>LC</i> + <i>NN</i> + <i>VV</i> + <i>NN</i> )

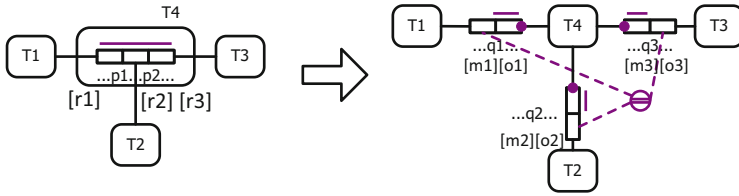
- (1) Verbs (including “is”, 是/*VC*; “have” as a main verb, 有/*VE*; other verb, 预订/*VV*). They are directly used as role names.
- (2) “Duplicated” verbs. An interesting linguistic phenomenon in Chinese is that we often use several verbs, which can be synonyms or contain similar meanings, to describe one verb. For example, 走(walk)动(move)=walk, 开(begin)始(begin)=begin, 活(live)动(move)=act and 分(split)离(leave)=separate. Those examples are used as role names.
- (3) Verbs with aspect particles (including perfective aspect 了 and durative aspect 着). A special kind of verbs in Chinese is called predicative adjunctive verbs (*VA*), which are often used with 了 to map an adjunctive into a verb. For instance, 花儿红了. If we translate it literally, then we get “花儿(flowers)红(red)了”. If it is translated figuratively, then it is “flowers become red”. In this example, we use 红了 as a role name. In Chinese, the meaning of a verb with or without durative aspect marker is often the same. An example is 意味着. 意味) means “imply”, and 意味着) also means “imply”. We can use 意味着) as a role name.
- (4) Verbs with the coordinating conjunctions 到, 予, for example, 涉及到 (involve), 找到 (find), 给予 (give) and 赋予 (endow). Such a coordinating conjunction indicates the follower as the object that receives the actions. When we translate such sentences from Chinese to English, they are translated as if they were without the coordinating conjunctions. For example, both 预订涉及到表演) and 预订涉及表演) can be translated into “Booking involves Showing”. In this example, we use 涉及到) as its role name.
- (5) Localizers (including monosyllabic localizers, e.g. 里) (inside), 外) (outside), 前) (before), 后) (after) and bi-syllabic localizers, e.g. 之间) (between), 周围) (around), 期间) (during)). They are used as role names for n-ary fact types.

The verbalization patterns *OP3* and *OPN* are also quite different. When many stems appear in one sentence [e.g. 影片(Film), 于时间) (on Time), 在影院) (at Cinema), 人(People), 看) (to see) and 放映) (to show)], the verb and the object that receives this verb are normally shifted to the end. For example, “Film shows on Time at Cinema” should be translated into 影片于时间在影院放映. “People see Film on Time at Cinema” should be translated into 人于时间在影院看电影.

In this section, we have presented patterns in English and Chinese, which can be further composed into constrained n-ary fact types. In the next section, we will discuss how n-ary fact types can be mapped into binary fact types.

### 3.2 Mapping Constrained N-Ary Fact Types to Constrained Binary Fact Types

Semantic Web triples are related to binary fact types, and though the binary form can result in verbalisations that are less natural than n-ary ones, it is preferable for automated reasoning and other computation. *Objectification* is a schema mapping (or schema transformation), which treats a relationship between objects as an



**Fig. 4** Method of mapping a ternary fact type into binary fact types

object itself (Halpin and Morgan 2008) and with which we can produce the binary form from the n-ary one. In the logic community, such process is called “model reification”.

It contains two steps. The first step is to label the new object, for example, we can label the ternary relation “... is showing on ... at ...” in Fig. 3 as “showing”. The second step is to treat this relation as an object and verbalize the relevant fact types by referring to the original relation. A verbalization example for Fig. 3 is shown as follows:

a **Film** is showing on a **DateTime** at a **Cinema** .  
 a **Person** books **Count** seat (s) for a **Showing** .

The method of mapping a constrained ternary fact type into constrained binary fact types is illustrated in Fig. 4.

In the mapped binary fact types in Fig. 4, a spanned uniqueness constraint, which is graphically represented as a bar above the roles and indicates a many-to-many relation, is applied. If no such uniqueness constraint is shown, we consider an implicit constraint to span all roles of the fact type. This is because any duplicate is merely a restatement of a known fact, not a new fact.

The objectification labelled with “T4” becomes a new entity type. The predicate text “... p1 ... p2 ...” of the ternary is replaced by three new binary predicates with text q1, q2 and q3 (in Fig. 3, one such binary predicate is shown as “Booking involve(s) Showing”). The original roles r1, r2 and r3 map to new roles—the mirror roles m1, m2 and m3. The population of the mirror roles is the same as the population of the original roles. The objectification T4 plays the objectification roles o1, o2 and o3. For each instance of the earlier relationship fact, an instance of T4 plays all three objectification roles exactly once. Thus, these objectification roles must each be covered by both a mandatory and a uniqueness constraint. The mandatory constraints match the requirement for each role of the original ternary to be populated.

The (default) uniqueness constraint in Fig. 4 must be mapped to a new external uniqueness constraint that covers the corresponding mirror roles.

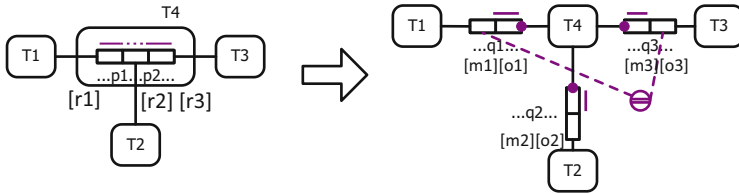


Fig. 5 Method of mapping a constrained ternary fact type into constrained binary fact types

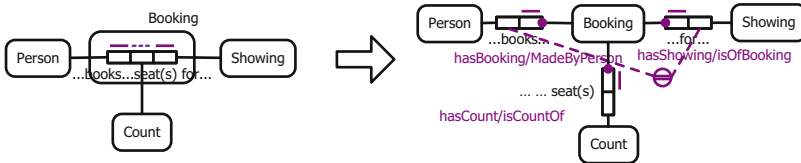


Fig. 6 An example of mapping

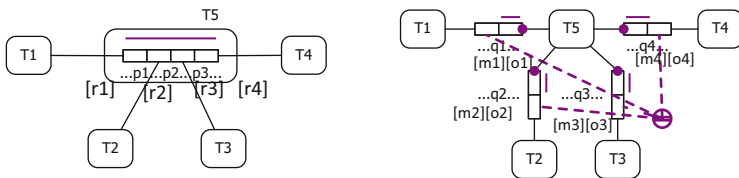


Fig. 7 Method of mapping a constrained quaternary fact type into constrained binary fact types

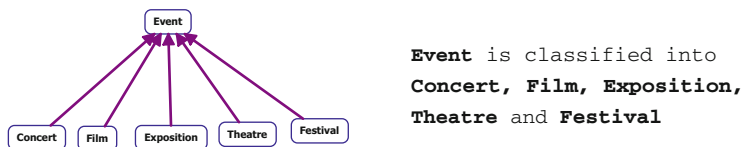
How to map a uniqueness constraint on the role pair  $\langle r1, r3 \rangle$  is illustrated in Fig. 5.<sup>5</sup> An example is shown in Fig. 6.

The method in Fig. 4 can be further extended for quaternary fact types as illustrated in Fig. 7. To present *all the possible mappings* between constrained n-ary fact types and binary fact types is not the focus; we refer the reader to the relevant research from the FBM community, such as (Franconi and Mosca 2013).

## 4 Implementation and Discussion

In the process of *conceptualization*, patterns of fact types provided by the domain experts are automatically identified when free texts are used. Our tagging engine uses the English Penn Treebank tag set from the Stanford Tagger. A combination of regular expressions and the Stanford Tagger has been adopted in the implementation. With regular expressions, keywords like “is a” and “is classified into” are

<sup>5</sup>Readers may imagine the mapping situations when  $\langle r1, r2 \rangle$  or  $\langle r2, r3 \rangle$  is unique.



**Fig. 8** An automatically generated ORM model

identified. Afterwards, it provides the category of each word. At the end, we can identify its pattern according to Fig. 1. For example, given a free text input “event is classified into concert, film, exposition, theatre or festival”, we get “is classified into” as the keyword. Then, we get the following categories for the rest: event (EO), film (EO), exposition (EO), theatre (EO) or (CC) and festival (EO). It matches pattern 4 from Fig. 1. At the end, an ORM model is generated as illustrated in Fig. 8.

We have used the LDC Chinese Treebank POS tag set, which claimed to have 93.99 % accuracy on known Chinese words and 84.60 % accuracy on unknown Chinese words by using distributional similarity clusters for the given data when the training set and test set are both drawn from the same corpus. A surprising observation we had is that it is impossible to analyse any complete sentence with this setting. Given the example in our chapter—预订涉及表演—the whole sentence is tagged as VV (other common verbs). It provides relatively satisfactory results on elementary items, such as 预订, 涉及 and 表演. The problem was caused by the implementation of the Stanford Tagger. We are engaged in ongoing work to re-implement the tagger to fix the Chinese verbalization patterns.

Note that the tagging engine is needed only when domain experts use free texts. If they carefully follow the verbalization patterns, the verbalization process becomes easier (this is the main usage for which CQL is designed to assist). In this case, the Stanford Tagger is only needed to discover basic verbalization patterns like EO.

It is possible to use other solutions for identifying the patterns, such as the one proposed by Bond et al. (2014) (this volume), which uses a multilingual lexicon from multilingual WordNet. We did not choose this solution for the following two reasons. (1) The current Chinese WordNet covers only 28 % of the core, which means that it is still in a primary research/implementation phase. (2) The types of semantic relations need to be extended. It is difficult to measure such effort because of the complexity of Chinese lexical items. Perhaps in the future, we can adopt this approach when it gets mature.

English verbalization has been implemented in NORMA. For example, the subset constraint shown in Fig. 3 requires that any Seat Allocation must be in the same Cinema as the Showing for which the Booking is made. NORMA verbalizes this constraint as follows:

**If some Booking has allocated some Seat then that Booking involves some Showing that involves some Cinema where that Seat is in that Cinema.**

Note that it uses the binary-mapped fact types (as discussed with Fig. 4) where the implied binary predicate text is “. . . involves . . .” CQL—in which this example has been implemented—supports this constraint in the form as illustrated in Fig. 8, which does not require the implied binary predicates:

Some Booking (in which some Person booked some Count seats for some Showing  
 in which some Film is showing on some DateTime at some Cinema))  
 has some allocated Seat  
 only if that Seat is in that Cinema

## 5 Conclusion and Future Work

This chapter introduces the concept of multilingual verbalization and a method of how to use verbalization patterns to formalize informal texts in a natural language. The objective is to raise the interest of the community of conceptual modelling on multilingual verbalization patterns for n-ary fact types and reuse. English and Chinese are chosen to demonstrate the patterns. A preliminary work on how a constrained n-ary fact type can be mapped into constrained binary fact types is demonstrated. In the future, we also want to design patterns for other languages, such as French and Dutch.

It is anticipated that future implementations of CQL, given multilingual object type names and predicate text (readings), will be able to accept complex statements in one language and accurately translate them into another language. Cross-language verbalization may require the automated predicate arity mappings discussed here.

## References

- Alexander, C., Ishikawa, S., & Silverstein, M. (1977). *A pattern language*. New York: Oxford University Press.
- Baader, F., Calvanese, D., McGuinness, D. L., Nardi, D., & Patel-Schneider, P. F. (2010). *The description logic handbook: Theory, implementation and applications*. Cambridge: Cambridge University Press.
- Bakema, G., Zwart Pieter, J., & van der Lek, H. (2002). *Volledig communicatiegeoriënteerde informatiemodellering*. Netherlands: Academic Service.
- Bond, F., Fellbaum, C., Hsieh, S.-K., Huang, C.-R., Pease, A., & Vossen, P. (2014). A multilingual lexico-semantic database and ontology. In P. Buitelaar & P. Cimiano (Eds.), *The multilingual semantic web*. Heidelberg: Springer.
- Collins, M. (2002). Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing 10* (pp. 1–8). Stroudsburg: Association for Computational Linguistics. doi:[10.3115/1118693.1118694](https://doi.org/10.3115/1118693.1118694).

- de Cea, G., Gómez-Pérez, A., Montiel-Ponsoda, E., & Suárez-Figueroa, M. (2008). Natural language-based approach for helping in the reuse of ontology design patterns. In A. Gangemi & J. Euzenat (Eds.), *EKAW 2008*. 5268 (pp. 32–47). Acitrezza: Springer.
- Demey, J., Jarrar, M., & Meersman, R. (2002). A conceptual markup language that supports interoperability between business rule modeling systems. *Proceedings of OTM 2002: COOPIS, DOA, AND ODBASE*. 2519, (pp. 19–35). California: Springer.
- Franconi, E., & Mosca, A. (2013). Towards a core ORM2 language (research note). In Y. Demey & H. Panetto (Eds.), *On the Move to Meaningful Internet Systems: OTM 2013 Workshops*. 8186 (pp. 448–456). Graz: Springer.
- Gangemi, A., & Presutti, V. (2009). Ontology design patterns. In S. Staab & R. Studer (Eds.), *Handbook of ontologies* (2nd ed., pp. 221–243). Heidelberg: Springer.
- Gangemi, A., & Presutti, V. (2010). Towards a pattern science for the semantic web. *Semantic Web*, 1(1–2), 61–68.
- Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Gómez-Pérez, A., Buitelaar, P., & McCrae, J. (2012). Challenges for the multilingual web of data. *Web Semantics*, 11, 63–71.
- Gromann, D., & Declerck, T. (2014). A cross-lingual correcting and complete method for multilingual ontology labels. In P. Buitelaar & P. Cimiano (Eds.), *The multilingual semantic web*. Heidelberg: Springer.
- Haarslev, V., Lutz, C., & Ralf, M. (1999). A description logic with concrete domains and a role-forming predicate operator. *Journal of Logic and Computation*, 9(3), 351–384. doi:10.1093/logcom/9.3.351.
- Halpin, T., & Curland, M. (2006). Automated verbalization for ORM 2. In R. Meersman, Z. Tari, & P. Herreto (Eds.), *On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops*. 4278 (pp. 1181–1190). Montpellier: Springer.
- Halpin, T., & Morgan, T. (2008). *Information modeling and relational databases* (2nd ed.). San Francisco: Morgan Kaufmann.
- Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. *14th International Conference on Computational Linguistics* (pp. 539–545). Stroudsburg: Association for Computational Linguistics.
- Heath, C. (2009). The constellation query language. In R. Meersman, P. Herrero, & T. Dillon (Eds.), *On the Move to Meaningful Internet Systems: OTM 2009 Workshops*. LNCS5872 (pp. 682–691). Vilamoura: Springer.
- Jardine, D. (1984). Concepts and terminology for the conceptual schema and the information base. *Computers and Standards*, 3, 3–17.
- Jarrar, M. (2005). *Towards methodological principles for ontology engineering* (Ph.D. Thesis). Vrije Universiteit Brussel, Brussel.
- León-Araúz, P., & Faber, P. (2014). Context and terminology in the multilingual semantic web. In P. Buitelaar & P. Cimiano (Eds.), *The multilingual semantic web*. Heidelberg: Springer.
- Marcus, M. P., Marcinkiewicz, M., & Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2), 313–330.
- Marshall, I. (1987). Tag selection using probabilistic methods. In R. Garside, G. Sampson, & G. Leech (Eds.), *The computational analysis of English: A corpus-based approach* (pp. 42–67). London: Longman.
- Nijssen, S. G., & Halpin, T. A. (1989). *Conceptual schema and relational database design: A fact oriented approach* (1st ed.). Upper Saddle River: Prentice Hall.
- Nijssen, M., & Lemmens, I. (2008). Verbalization for business rules and two flavors of verbalization for fact examples. In R. Meersman, Z. Tari, & P. Herrero (Eds.), *On the Move to Meaningful Internet Systems: OTM 2008 Workshops*. LNCS Vol. 5333 (pp. 760–769). Mexico: Springer.
- Schreiber, G., Akkermans, H., Anjewierden, A., De Hoog, R., Shadbolt, N. R., Van de Velde, W., et al. (1999). *Knowledge engineering and management — the CommonKADS methodology*. Cambridge: The MIT Press.
- Spyns, P., Tang, Y., & Meersman, R. (2008). An ontology engineering methodology for DOGMA. *Journal of Applied Ontology*, 3(1–2), 13–39 (G. Guizzardi, & T. Halpin, Eds.).

- Tang, Y., & Meersman, R. (2008). SDRule markup language: Towards modeling and interchanging ontological commitments for semantic decision making. In A. Giurca, D. Gasevic, K. Taveter, A. Giurca, D. Gasevic, & K. Taveter (Eds.), *Handbook of research on emerging rule-based languages and technologies: Open solutions and approaches* (Vol. Sec. I, pp. 99–123). USA: IGI Publishing. Chapter V.
- Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. *Proceedings of HLT-NAACL 2003. 1* (pp. 252–259). Stroudsburg: Association for Computational Linguistics. doi:[10.3115/1073445.1073478](https://doi.org/10.3115/1073445.1073478).
- Toutanova, K., & Manning, C. D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000). 13* (pp. 63–70). Stroudsburg: Association for Computational Linguistics. doi:[10.3115/1117794.1117802](https://doi.org/10.3115/1117794.1117802).
- Xia, F. (2000). *The part-of-speech tagging guidelines for the Penn Chinese Treebank (3.0)* (Technical report). Philadelphia: University of Pennsylvania.



# **Part II**

## **Methods**

# Publishing Linked Data on the Web: The Multilingual Dimension

Daniel Vila-Suero, Asunción Gómez-Pérez, Elena Montiel-Ponsoda,  
Jorge Gracia, and Guadalupe Aguado-de-Cea

**Abstract** Linked Data technologies and methods are enabling the creation of a data network where pieces of data are interconnected on the Web using machine-readable formats such as Resource Description Framework (RDF). This paradigm offers great opportunities to connect and make available knowledge in different languages. However, in order to make this vision a reality, there is a need for guidelines, techniques, and methods that allow publishers of data to overcome language and technological barriers. In this chapter, we review existing methodologies from the point of view of multilingualism and propose a series of guidelines to help publishers when publishing Linked Data in several languages.

**Key Words** Linked data • Multilingual linked data • Semantic web

## 1 Introduction

The Linked Data (LD) initiative (Berners-Lee 2006; Bizer et al. 2009) is building a data network where datasets in machine-readable formats are interconnected using web technologies. The growth of this data network has gone hand in hand with important advances in both the methodological (Heath and Bizer 2011; Villazón-Terrazas et al. 2011) and technological support (Auer et al. 2012) to facilitate the publication and consumption of linked datasets. These advances have been proven to be relevant in (1) the production of Resource Description Framework (RDF) datasets out of different types of data sources (e.g., relational databases (Das et al. 2012), spreadsheets (Maali et al. 2012), etc.), (2) the discovery of links between RDF datasets (Ferrara et al. 2011), or (3) the publication of metadata describing linked datasets (Alexander et al. 2011; Maali et al. 2013). However, among the challenges and obstacles that still need to be overcome to truly exploit this data

---

D. Vila-Suero (✉) • A. Gómez-Pérez • E. Montiel-Ponsoda • J. Gracia • G. Aguado-de-Cea  
Ontology Engineering Group, Universidad Politécnica de Madrid, Madrid, Spain  
e-mail: [dvila@fi.upm.es](mailto:dvila@fi.upm.es); [asun@fi.upm.es](mailto:asun@fi.upm.es); [emontiel@fi.upm.es](mailto:emontiel@fi.upm.es); [jgracia@fi.upm.es](mailto:jgracia@fi.upm.es);  
[lupe@fi.upm.es](mailto:lupe@fi.upm.es)

network, containing billions of RDF triples,<sup>1</sup> is the idea of multilingualism as a pervasive aspect of this Web of Data (WoD).

As claimed by Gracia et al. (2011), the growing demand of semantic technologies sets the WoD as an excellent platform to seek for solutions that can manage multilingualism. However, some initial steps have to be taken, so as to ease the publication of high-quality multilingual linked data, as well as to assist organizations in generating new value from such multilingual linked data.

In the last 5 years, some methodological guidelines for publishing LD (Villazón-Terrazas et al. 2011; Hyland et al. 2013) have proved to be successful in several knowledge domains, such as mass media (Kobilarov et al. 2009), geography (Auer et al. 2009; Vilches-Blázquez et al. 2013), or cultural heritage (Isaac and Haslhofer 2013; Vila-Suero and Gómez-Pérez 2013). These guidelines are meant to provide high-quality LD and follow a principle-based practice when publishing and consuming LD. Yet, they have overlooked the language dimension, and therefore no recommendations have been given for publishing LD in one or several natural languages. Our aim, in this chapter, is to reflect on the intricacies of adding language-related features during the LD publication process. For illustration purposes, we will present a real use case *geo.linkeddata.es*<sup>2</sup> (Vilches-Blázquez et al. 2013) the result of the publication of LD out of the databases from the Spanish National Institute of Geography (IGN, Instituto Geográfico Nacional).

The *geo.linkeddata.es* dataset contains metadata in several languages describing geographical and spatial information such as administrative units and hydrography. For instance, names are registered not only in Castilian but also in the several co-official languages of Spain (i.e., Aranese, Basque, Catalan/Valencian, and Galician). Additionally, standards for producing catalogue metadata (like the norm “Núcleo Español de Metadatos,” a profile of the ISO 19115 Norm for Geographic Information) are available in Spanish. Finally, linking to datasets like DBpedia or other national geography institutes means dealing with language heterogeneity as the data can be in English, French, etc. In other words, *geo.linkeddata.es* represents a good use case of data multilingualism and exemplifies three major issues related to language features in the publication of LD:

- Data sources may contain information in several natural languages: multilingual data including landforms or geographical feature names or monolingual data like some river or city names.
- Vocabularies for describing the data may be also in several languages (multilingual vocabularies) or only in one language (monolingual vocabularies) that can be different from the language required by the publisher or some potential data consumers.

---

<sup>1</sup>See, for example, <http://lod-cloud.net>, <http://datahub.io>, or <http://datacatalogs.org> (retrieved March 28, 2014).

<sup>2</sup><http://geo.linkeddata.es> (retrieved March 28, 2014).

- Target datasets for linking and enriching the original data sources can be, in their turn, in several natural languages.

Considering these factors, two questions arise: (1) *Are current guidelines, best practices, and tools suitable to cope with these and other language-related issues?* (2) *Do they provide valuable guidance and mechanisms for producing high-quality multilingual data in such scenarios?*

According to our experience in publishing the *geo.linkeddata.es* dataset, guidance on these aspects is still very limited. So, our purpose here is to review, discuss, and elaborate on the current guidelines for publishing LD (1) by focusing on those methods, techniques, and tools that can help RDF publishers to cope with language barriers and (2) by identifying existing gaps, as well as remaining research and technical challenges. We will discuss each of these guidelines, methods, and tools and illustrate them with examples from the *geo.linkeddata.es* dataset.

As a first step, we build our analysis on the method proposed by Villazón-Terrazas et al. (2011), which adopts an iterative incremental model covering the following activities: (1) *specification*, to analyze and select data sources; (2) *modeling*, to develop the model that represents the information domain of the data sources; (3) *generation*, to transform the data sources into RDF datasets, (4) *linking*, to create links between different RDF datasets; (5) *publication*, to publish on the Web the model, RDF datasets, metadata describing these datasets, and the links to other datasets; and (6) *exploitation*, to develop applications that make use of the dataset at stake. In turn, each activity contains one or more tasks.

The rest of the chapter is organized as follows. In Sects. 2–6, we explain each of the activities included in the guidelines, that is, specification, modeling, generation, interlinking, and publication, respectively. Section 7 provides our conclusions and lessons learnt.

## 2 Specification

In this section, we explain how to deal with language issues during the specification phase. Basically, we need to analyze whether there is available documentation in different natural languages describing the original data sources. A further aspect to take into account at this stage is the design of unique resource identifiers and in particular uniform resource identifiers (URIs) and internationalized resource identifiers (IRIs). For a better understanding of the following sections, we will first define a number of concepts that will be used throughout the chapter.

By *monolingual* datasets, we understand those resources that contain data descriptions only in one language. In the current LOD (Linked Open Data) cloud,<sup>3</sup> we find an example of this in the *geolinkeddata.es* dataset.<sup>4</sup>

---

<sup>3</sup><http://lod-cloud.net/> (retrieved March 28, 2014).

<sup>4</sup><http://geolinkeddata.es> (retrieved March 28, 2014).

*Multilingual* datasets are those that contain data descriptions in several languages. In the LOD cloud, a multilingual dataset can be one that contains values for its RDF properties (`rdfs:label` or `skos:prefLabel`) in several languages, such as the well-known AGROVOC dataset (Caracciolo et al. 2013) (for other ways of providing lexical and linguistic descriptions to RDF datasets, see Sect. 3.3).

Finally, we define as *cross-lingual* those datasets that are linked to other resources with which they do not have any language in common. An example of this type of resource would be EuroWordNet (Vossen 2004), in which the various monolingual Wordnets are mapped to or linked to each other through a central hub of core concepts (known as Interlingual Index). For further details on the linking of Wordnets, see Bond et al. (this volume).

## 2.1 Analysis of the Data Sources and Their Model

The first activity of the LD publication process is to analyze the sources that will be used for publishing LD, as well as the data model(s) used within those sources. We review how language-related features affect the process of specification and how publishers can approach this task in a sensitive way with regard to natural language.

To illustrate these ideas, we take as running example the data sources of IGN, which consist of database records with information such as administrative units, spatial information, landforms, etc.

In this task, we have to take into account two layers: (1) the *data model*, in this case defined by the schema of the database and the associated standards used for describing the data and (2) the *content of the sources* (i.e., the data itself), in this case metadata describing administrative units, bodies of water, etc.

In Fig. 1, we show the metadata descriptions about “Madrid” municipality and “Madrid” province. As shown in the figure, the data model corresponds to the types of entity<sup>5</sup> (i.e., municipality and province) and the different attributes and relationships<sup>6</sup> (e.g., is capital of, alternative name, latitude, etc.), whereas the content corresponds to the value of each attribute (e.g., “Municipio de Madrid,” “40.4178488946733,” etc.)

In order to specify these two layers (i.e., data model and content) in a multilingual setting, we have to consider several issues that we analyze in the following paragraphs.

### 2.1.1 Data Model

The data model (including related standards, terminology, etc.) used for the description of entities, attributes, and relationships can be found in the language of

---

<sup>5</sup>Classes in RDF terminology.

<sup>6</sup>Properties in RDF terminology.

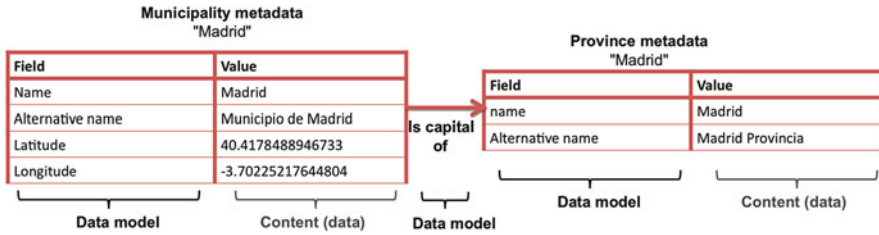


Fig. 1 “Madrid” municipality and “Madrid” province database entries

the dataset publisher or in other languages. For instance, the database schema of IGN and associated standards like the “Núcleo Español de Metadatos”<sup>7</sup> (Spanish Core Metadata norm) are based on standards available in English that have been translated into several languages (e.g., Spanish). As we will see in the modeling activity (Sect. 3), having a specification of the data models, terminology, and standards in several languages, or at least in the language required by the RDF publisher, can save effort and cost and produce higher-quality vocabularies from a multilingual perspective. Thus, in this task we recommend compiling all available information about the data model used in the sources and identifying the natural languages that will be used for designing the vocabulary.

### 2.1.2 Content

Content can be *language independent* or *language dependent*. Some properties such as identifiers, numbers, and some date formats are usually language neutral, whereas names, titles, textual descriptions, and some date formats are normally language dependent as they are bound to a specific language. For instance, latitude and longitude are language neutral. Further, language-dependent properties do not always make explicit the language of the content they carry. For instance, some database schemas do not provide any information about the language of the content of different fields although some content can be in different languages. Given this situation, we recommend specifying and classifying attributes in the following way: (1) language independent or language dependent, based on the content they carry; (2) for language dependent attributes, the language can be *explicit* (e.g., using a metadata annotation, a pointer to the language description, label or code, etc.) or *unspecified* (e.g., “name” shown in Fig. 1 is a language-dependent attribute, with unspecified language). For the former case (i.e., *explicit language*), the mechanisms

<sup>7</sup><http://metadatos.ign.es/metadatos/Normativa/nucleo-espanol-de-metadatos-para-datos-y-servicios> (retrieved March 28, 2014).

that are used to indicate the language should be documented. For the latter case (i.e., *unspecified language*), the dataset publisher should apply language identification techniques in the *generation* activity, as we will discuss in Sect. 4.

## 2.2 URIs and IRIs Design

The goal of this task is to design the structure of the identifiers used to name RDF resources, either those for the TBox (classes and properties) or those for the ABox (instances). There are basically two options: to use meaningful or descriptive resource identifiers, that is, the use of natural language descriptions in the local name of URIs and IRIs (e.g., the URI<sup>8</sup> of the municipality of Madrid), or rather the employment of opaque resource identifiers, i.e., non-human-readable local names (e.g., the URI<sup>9</sup> for the municipality of Madrid in geonames). Both approaches have well-known advantages and disadvantages that we summarize on the light of the multilingual dimension.

As mentioned by Montiel-Ponsoda et al. (2011a, b), the main benefits of using meaningful URIs (also called descriptive by Labra et al. 2014) are that they help developers to faster understand the underlying model, are easy to remember, favor interoperability, and are better displayed by many ontology editing tools. From a technical point of view, in a multilingual scenario, we have several options:

- The use of *meaningful URIs*, in which the local name is normally in English or any other Latin-based language making use of ASCII characters (e.g., the URI<sup>10</sup> of the municipality of Madrid).
- The use of *full IRIs* (Labra et al. 2014), created with the aim of allowing the use of Unicode characters for languages that do not follow the Latin alphabet. This enables the use of Unicode characters not only for local names but also in the domain part (e.g., the IRI<sup>11</sup> for the autonomous community of Madrid).
- The use of *internationalized local names*, which are IRIs in which the domain part is restricted to ASCII characters while the local name can use Unicode characters (Labra Gayo et al. 2014), for instance, the IRI of the municipality of Alcorcón.<sup>12</sup>

Additionally, if our starting point is a multilingual resource in which TBox and ABox contain information in several languages, more fundamental questions should be brought up: *Which language should we use for the local names in meaningful URIs or IRIs? Should English be the default language? In which language was*

---

<sup>8</sup><http://geo.linkeddata.es/resource/Municipio/Madrid>.

<sup>9</sup><http://www.geonames.org/6355233>.

<sup>10</sup><http://geo.linkeddata.es/resource/Municipio/Navacerrada>.

<sup>11</sup>[http://geo.linkeddata.es/resource/ComunidadAutónoma/Comunidad\\_de\\_Madrid](http://geo.linkeddata.es/resource/ComunidadAutónoma/Comunidad_de_Madrid).

<sup>12</sup><http://geo.linkeddata.es/resource/Municipio/Alcorcón>.

*the dataset originally created? Does it contain preferred labels in that language (e.g., by means of the `skos:prefLabel` annotation property)? Or should we opt for opaque URIs to avoid any language bias? Moreover, if we decide to use meaningful URIs or unrestricted IRIs, which format should we follow in the local name (CamelCase strategy, use of space or underscores as word separators)? (e.g., the URI for the property `esCapitalDe`<sup>13</sup> vs. the IRI<sup>14</sup> of the autonomous community of Madrid). These are some questions that should be addressed beforehand in order to choose the naming format.*

There are some arguments that support the use of opaque URIs or IRIs (Montiel-Ponsoda et al. 2011a, b). For example, in a Semantic Web context, resource identifiers are intended for machine consumption, so that there is no need for them to be human readable. It is also well accepted that opaque URIs make ontologies more stable, so once the ontology has been published and adopted by a community of users, local names should not change even if the natural language descriptions associated to them are modified (unless the actual meaning of concepts has changed). Furthermore, opaque URIs may also be a good choice if we want to avoid any language bias.

### 3 Modeling

After the specification activity, it is time to design the model for the selected data sources. The first and most important recommendation at this stage is to reuse available vocabularies as much as possible. Current methodological guidelines divide this activity into two main tasks: (1) analysis and selection of domain vocabularies to maximize reuse of widely deployed vocabularies and (2) development of the domain vocabulary reusing as many terms as possible and creating those concepts that are not covered by the vocabularies analyzed in the previous task. From a multilingual perspective, however, such guidelines may not cover the linguistic and cultural needs. It is frequently the case that LD publishers want to provide descriptions to vocabulary classes and properties in their own language or even in several languages to improve vocabulary usability, data visualization, and so on. For these reasons, we propose an optional task, namely, (3) “ontology localization.” We also review tasks (1) and (2) to account for the multilingual dimension.

#### 3.1 Analysis and Selection of Domain Vocabularies

The goal of this task is to analyze and select already available domain vocabularies. Some catalogues and services are currently available for searching vocabularies on

---

<sup>13</sup><http://geo.linkeddata.es/ontology/esCapitalDe>.

<sup>14</sup>[http://geo.linkeddata.es/resource/ComunidadAutónoma/Comunidad\\_de\\_Madrid](http://geo.linkeddata.es/resource/ComunidadAutónoma/Comunidad_de_Madrid).



the Web such as the Semantic Web Search Engine<sup>15</sup> (SWSE), Sindice,<sup>16</sup> Datahub,<sup>17</sup> Falcons,<sup>18</sup> or LOV<sup>19</sup> (Linked Open Vocabularies). These catalogues and services allow users to (1) *search for similar data* within similar domains (SWSE, Sindice, and Datahub) and (2) *search for vocabularies or specific terms* (Falcons and LOV). As for the multilingual dimension of vocabularies, we question ourselves: *do existing catalogues and services take this dimension into account, facilitating discovery of terms no matter the language they are described in?*

Gómez-Pérez et al. (2013) have described the experiments conducted with the aforementioned services and observed that current multilingual support is still limited. The LOV service is the one that provides a better performance for the following reasons: (1) It is able to index multilingual labels, (2) it provides the best UI support for languages, and (3) it is a well-established repository maintained by the Open Knowledge Foundation (OKF) and with a clear curation strategy.

### 3.2 Development of the Domain Vocabulary

The goal of this task is to develop a vocabulary for modeling the data contained in the data sources. As mentioned before, this task consists in (1) reusing the vocabularies or terms selected in the previous task and/or (2) creating those terms that are not covered by the analyzed vocabularies:

1. As for reusing existing vocabularies or terms found in widely used vocabulary catalogues, the publisher might be interested in adding term descriptions in a language other than the one(s) initially used by the vocabulary publisher.
2. If new vocabulary terms need to be created, it is a good practice to reuse available non ontological resources like standards, glossaries, lexica, or thesauri describing existing knowledge of the domain and transform them into RDF ontologies (Villazón-Terrazas et al. 2011). If the resources to be transformed provide labels in different languages, it might be useful to include them in the ontology. If the resources are not available in the language desired by the publisher, she/he may decide to include labels in those languages.
3. Finally, we foresee a further possibility in which the publisher decides to use his/her own vocabulary in a specific language and establish links to an existing RDF vocabulary, in the same or a different natural language. Should this be the case, the multilingual dimension involves the discovery of cross-lingual links.

---

<sup>15</sup><http://swse.deri.org/> (retrieved March 28, 2014).

<sup>16</sup><http://sindice.com> (retrieved March 28, 2014).

<sup>17</sup><http://datahub.io> (retrieved March 28, 2014).

<sup>18</sup><http://ws.nju.edu.cn/falcons> (retrieved March 28, 2014).

<sup>19</sup><http://lov.okfn.org> (retrieved March 28, 2014).

In the context of the IGN example, the Spanish publishers decided to reuse several existing standards and terminologies, a number of them in English. In order to facilitate the use of their ontology, they manually localized or translated these vocabularies into Spanish. The translation was partly the result of a direct translation of the English labels and partly took into account the Spanish source documentation. Additional labels in Spanish were included for classes and properties making use of the RDF and SKOS (Simple Knowledge Organization System) label annotation property.

### 3.3 *Ontology Localization*

“Ontology localization” has been defined as the process of adapting an ontology to the needs of a particular (linguistic and cultural) community (Espinoza et al. 2009). A localized ontology can be understood as an ontology adapted to the target community and language and used independently of the original ontology or, most commonly, as an ontology in which the vocabulary or TBox has been translated into one or several natural languages, so that it contains terms in several languages for describing classes and properties (Gracia et al. 2011). In the LD context, thus, publishers should decide which representational model to follow according to their multilingual and linguistic needs. For an interesting discussion on translation equivalents, see Hirst (this volume). Three main alternatives have been identified to account for linguistic information in ontologies or vocabularies (Montiel-Ponsoda et al. 2011a, b): (1) *multilingual labeling approach*, (2) *association of the vocabulary to an external lexicon model*, and (2) *cross-lingual linking or matching approach*.

To illustrate these approaches, we will use as example the localization of the ISBD (International Standard Bibliographic Description) standard, a standard for describing bibliographic resources such as books or maps. The ISBD vocabulary contains English labels for classes and properties and their translations into Spanish, resulting in a multilingual vocabulary in English and Spanish. For this specific case, the publishers decided to rely on the SKOS annotation property for preferred labels (`skos:prefLabel`) and agreed on the use of only one preferred label per language (see Sect. 2.2). However, the Spanish translation revealed a problem which was not apparent in the English version, namely, that some labels were adjectives (cartographic in English), which in Spanish require a form change depending on whether the word they modify is masculine (cartográfico) or feminine (cartográfica). Because of the agreed restriction, the formula “cartográfico/a” was suggested (`skos:prefLabel "cartográfico/a"@es`), in which the forward slash is used to indicate the choice between masculine and feminine. This solution also has some problems, since adjectives will not naturally appear in free texts in this way.

### 3.3.1 Multilingual Labeling Approach

The first alternative relies on a single conceptual or data structure to which alternative labeling information is provided in the form of plain literals represented as properties of concepts. This is supported by RDFS, SKOS, or the SKOS extension, SKOS-XL, which allows labels to be treated as RDF classes. See examples below.

```

isbd:T1001    rdfs:label      "cartográfico"@es; # RDFS
              rdfs:label      "cartográfica"@es.
isbd:T1001    skos:prefLabel  "cartográfico/a"@es. # SKOS
isbd:T1001    skosxl:prefLabel :cartografico. # SKOS-XL
:cartografico a skosxl:Label;
              skosxl:literalForm "cartográfico"@es.
isbd:T1001    skosxl:prefLabel :cartografica.
:cartografica a skosxl:Label;
              skosxl:literalForm "cartográfica"@es.

```

**Listing 1** Examples in RDFS, SKOS, and SKOS-XL

The main disadvantage of the labeling facility of RDFS and SKOS is that the labels that can be related with one vocabulary term result in a set of unrelated labels whose motivation cannot be asserted and for which further properties cannot be specified. This is, in a sense, solved by the SKOS-XL description, although it does not provide a principled way for specifying linguistic properties of labels, nor is it conceived to linguistically enrich vocabulary terms (for instance, specifying that the plural forms of *cartográfico* and *cartográfica* are obtained by adding an indicating that *cartográfico/a* is an adjective, etc.). For these reasons, linguistic models have been specifically proposed to enrich ontologies.

### 3.3.2 Association of the Vocabulary to an External Lexicon Model

The second alternative consists in associating the vocabulary to a lexicon model that contains the linguistic information relative to that vocabulary (in one or several languages). An example of a resource that follows this approach is the Open Multilingual Wordnet, which consists of the linking of 22 wordnets to the Princeton Wordnet and to the SUMO ontology. See a detailed description in Bond et al. (this volume). Examples of these ontology-lexicon models are LexInfo (Cimiano et al. 2010), LIR (Montiel-Ponsoda et al. 2011a, b), or *lemon*<sup>20</sup> (McCrae et al. 2012). In fact, the *lemon* model has an RDF implementation that allows publishing linguistic information in the LD format. On the limitations of such models, see Hirst (this volume).

<sup>20</sup>Here it is also worth mentioning the OntoLex W3C community effort that aims at proposing a standard model of linguistic descriptions relative to ontologies and vocabularies.

In order to illustrate the potential of such models, we present how *lemon* allows for the inclusion of the two adjectival forms of the cartographic adjective in Spanish, the masculine and the feminine, by linking them to that property in the ontology by means of a `LexicalEntry` with two `LexicalForm`'s (masculine and feminine). The model is also able to represent that these are form variants of the same lexical entry. ISOcat categories are used in the example to represent the grammatical gender.

```
isbd:T1001 lemon:isReferenceOf [lemon:isSenseOf :cartographic].
:cartographic a lemon:LexicalEntry;
    lemon:form [lemon:writtenRep "cartográfico"@es;
                isocat:grammaticalGender isocat:masculine];
    lemon:form [lemon:writtenRep "cartográfica"@es;
                isocat:grammaticalGender isocat:feminine].
isocat:grammaticalGender rdfs:subPropertyOf lemon:property.
```

**Listing 2** Example in *lemon*. For readability, we have substituted the identifier of the `isocat` categories by descriptive names (e.g., `isocat:grammaticalGender` for **isocat:DC-1297**)

### 3.3.3 Cross-Lingual Linking or Matching Approach

This third possibility can be followed whenever there are two or several vocabularies defined in different natural languages, covering the same or similar subject domains. In this approach, links are established between the vocabulary terms that describe the two vocabularies. This scenario also involves the automatic discovery of links, another crucial issue in the Multilingual Semantic Web. On the difficulties for establishing cross-lingual mappings, see Hirst (this volume).

A number of recently developed cross-lingual ontology alignment tools can be used to that end. For a survey on the topic, see Trojahn et al. (this volume). Currently, equivalent links can be represented by means of properties of current Semantic Web languages such as OWL (`owl:sameAs` to link individuals in ontologies or `owl:equivalentClass` and `owl:equivalentProperty` to link classes and properties in ontologies that have the same extension), as well as with other commonly used vocabularies such as SKOS (`skos:closeMatch` to link two concepts that are sufficiently similar and `skos:exactMatch`, when the similarity degree is even higher). It could be argued that such links can be reused for the purpose of establishing links between classes, properties, and individuals expressed in different natural languages in the LOD cloud. However, we claim that some of these cross-lingual equivalences need to be analyzed carefully within the multilingual dimension, since we may want to establish cross-lingual and cross-cultural equivalences that may not admit the strong ontological commitments that current links make. For more on this, see Montiel-Ponsoda et al. (2011b). See also León-Araúz and Faber (this volume) for an extensive discussion on cross-linguistic problems.

Continuing with our example, we show a simple example of a cross-lingual link between the entity “province of Madrid,” as it is represented in the geolinkeddata dataset (“provincia de Madrid”) and the geonames dataset (“Province of Madrid”):

```
@prefix geoes: <http://geo.linkeddata.es/ontology/> .
@prefix geonames: <http://geonames.org/> .

geonames:6355233 a geonames:Place;
rdfs:label "Province of Madrid"@en.
<http://geo.linkeddata.es/resource/Provincia/Madrid>
a geoes:Municipio;
  rdfs:label "Provincia de Madrid"@es;
  owl:sameAs <http://www.geonames.org/6355233> .
```

**Listing 3** Example of cross-lingual mapping

### 3.3.4 Discussion

The main difference between the first two approaches is that the first option considerably restricts the amount and type of linguistic information that can be related to vocabulary elements, whereas the second one leaves open the inclusion of as much linguistic information as needed by the final application. The choice between one and the other model will depend on the linguistic requirements of each use. As for the third approach, it depends on the availability of similar vocabularies in different natural languages.

## 4 Generation

This activity deals with the transformation of the data sources selected in the *specification* activity (Section 2) using the model developed in the *modeling* activity (Section 3). This is a crucial activity in the process of publication and is, of course, influenced by language-related features. In this chapter, we focus on two core aspects of the RDF generation and point the reader to other relevant works.

### 4.1 Language Identification

As reported by Gómez-Pérez et al. (2013), the current usage of language tags in RDF datasets is still limited (only 21.5 % of all recorded literals on average), and there is a need of adequate guidelines and techniques for tagging the language. Also, as discussed in Sect. 2.2, *language-dependent* properties can (1) explicitly specify the language of the content that they carry (via language codes, external

information, etc.) or (2) leave the language *unspecified*. For the former case, the generation activity should include mechanisms to leverage the specified language and to properly tag the language of the generated RDF literals. For the latter scenario, it might be necessary to automatically “guess” or “identify” the language, using so-called “language identification” techniques (Dunning 1994). For this, we find literature that can be useful for the case of RDF properties, where literals are usually short (Gottron and Lipka 2010), as well as some available tools.<sup>21</sup>

## 4.2 Encoding Issues

An important aspect when working with languages whose scripts make use of characters not included in ASCII is the appropriate handling of the encoding of such characters. The *generation* activity is probably the most important activity in order to assure proper encoding, thus producing quality RDF data. When generating LD, encoding issues affect several levels: (1) URI and IRI handling, (2) different RDF serialization formats (e.g., RDF/XML, NTriples, etc.), and (3) libraries and tools for RDF (e.g., triple-stores,<sup>22</sup> APIs,<sup>23</sup> RDF generation tools,<sup>24</sup> etc.). Taking informed decisions in the selection of technologies, serialization formats, and unique identifiers (IRIs or URIs) will lead to better quality RDF data and avoid problems for consumers. In this sense, we point the reader to Auer et al. (2010), which provides an in-depth survey of known issues that might help publishers to make suitable choices.

## 5 Interlinking

In a multilingual WoD, semantic data with lexical representations in one natural language are mapped to equivalent or related information in other languages, thus allowing navigation across multilingual information by software agents (Gracia et al. 2011). Several activities have to be carried out for cross-lingual interlinking: (1) the selection of relevant and authoritative mono/multilingual datasets to link, (2) the automatic discovery of equivalent and/or related entities between the dataset and the selected external resources, and finally (3) the representation and storage

---

<sup>21</sup>See, for example, <http://tika.apache.org/>, <http://code.google.com/p/language-detection> and <http://nutch.apache.org> (retrieved March 28, 2014).

<sup>22</sup>For example, Virtuoso (<http://openlinksw.com>), 4Store (<http://4store.org>), and Allegrograph (<http://www.franz.com/agraph/allegrograph/>) (retrieved March 28, 2014).

<sup>23</sup>For example, Apache Jena (<http://jena.apache.org/>), Sesame (<http://www.openrdf.org/>), and ARC2 (<https://github.com/semsol/arc2>).

<sup>24</sup>For example, RDF refine (<http://refine.deri.ie/>) and Apache Any23 (<http://any23.apache.org/>).

of the discovered links. In particular, cross-lingual link discovery involves the automatic discovery of relationships between data items to increase the external connectivity of the RDF dataset in a multilingual scenario. This poses an added challenge because of data sources being available in different natural languages. There are many tools and techniques for discovering links between data items of different RDF datasets (see Ferrara et al. 2011 for a survey). Nevertheless, none of these techniques consider multilingualism as an explicit feature and do not include specific techniques to deal with language diversity during the process of link discovery. Therefore, more research is also needed on *automatic* methods for cross-lingual instance matching.

## 6 Publication

The publication of multilingual resources would involve the same tasks as in a monolingual process: (1) dataset publication, (2) metadata publication, and (3) enabling effective discovery. In the context of this chapter, we limit the scope to the second task. In recent years, there have been two major initiatives for providing vocabularies for publishing metadata describing datasets and catalogues: VoID<sup>25</sup> (Vocabulary of Interlinked Datasets) (Alexander et al. 2011) and DCAT<sup>26</sup> (Data Catalog Vocabulary) (Maali et al. 2013), both published by W3C. In this section, we show through examples how to account for the language dimension of datasets using these two vocabularies.

Although there might be other areas where language could be involved (for instance, when the dataset contains cross-lingual links), the most basic aspect to describe is *the language or languages used in the dataset*. Surprisingly, the language dimension in VoID is not included in its specification. DCAT, on the other hand, includes a property to indicate language by means of the `dcterms:language` property and defines the range of the property in the following way: (1) use resources defined by the Library of Congress,<sup>27</sup> and (2) if an ISO 639-1 (two-letter) code is defined for language, then its corresponding IRI *should* be used; otherwise, (3) if no ISO 639-1 code is defined, then the IRI corresponding to the ISO 639-2 (three-letter) code *should* be used. As both VoID and DCAT reuse the *Dublin Core Metadata Terms*<sup>28</sup> vocabulary for providing basic metadata (e.g., `dcterms:publisher`, `dcterms:title`, etc.), it seems natural to recommend publishers to follow the recommendation found in DCAT, also for building VoID descriptions. Therefore, in

---

<sup>25</sup><http://www.w3.org/TR/void/>.

<sup>26</sup><http://www.w3.org/TR/vocab-dcat/>.

<sup>27</sup><http://id.loc.gov/vocabulary/iso639-1.html> (retrieved March 28, 2014).

<sup>28</sup><http://dublincore.org/documents/2010/10/11/dcmi-terms/> (retrieved March 28, 2014).

Listing 4, we provide an example of the recommended mechanism to indicate the dataset language for VoID and DCAT.

```
# VoID description
:geoes a void:Dataset;
  dcterms:language <http://id.loc.gov/vocabulary/iso639-1/es> .
# DCAT description
:geoes a dcat:Dataset;
  dcterms:language <http://id.loc.gov/vocabulary/iso639-1/es>;>;
```

**Listing 4** VoID and DCAT descriptions indicating the language of the dataset

## 7 Conclusions

In this chapter, we revisit available methodological guidelines for the publication of data sources according to the LD paradigm from a multilingual perspective. Our aim has been to identify which methods, technologies, and tools, currently used for publishing and consuming LD, can be directly applied to multilingual resources and which need to be enhanced to account for multilingualism. As has been shown, the five activities identified in the methodological guidelines (specification, modeling, generation, linking, and publication) all involve a certain degree of revision.

As for the first activity, specification, a careful analysis of the data sources has to be performed, whenever they contain data descriptions in several natural languages. This will have a decisive influence on subsequent activities, especially on the modeling one. Related with this activity is an adequate identification or tagging of the RDF literals, as well as the decision on the use of meaningful vs. opaque URIs/IRIs.

The next activity, modeling, also involves some additional tasks. In the first place, publishers have to search for existing vocabularies documented in a certain natural language or in several languages. As has been reported, only a limited number of services provide this functionality. Secondly, the creation of new vocabulary classes and properties to meet linguistic and cultural needs has to be considered. Finally, it has to be decided which modeling strategy better suits the publisher's requirements according to the linguistic needs (multilingual labeling, external lexicon, or cross-lingual linking) and the availability of similar vocabularies in other languages.

In the generation activity, technologies for producing RDF will have to be customized to deal with multilingual data sources. Moreover, when mapping the data sources and the domain model, publishers ought to be sensitive to language-dependent properties and use RDF language tags.

Once the datasets have been generated, publishers will proceed with the linking activity. The linking possibilities grow considerably in a multilingual scenario but also involve greater difficulties. Vocabularies and datasets can be linked to other



datasets in the same language or in different languages. However, the discovery of cross-lingual links involves dealing with cultural divergences and cannot count on sound technological support. In any case, publishers have to consider the ontology commitments of the different types of links in a multilingual scenario.

The last activity, the publication activity, should also be enhanced to take into account the specification of the natural languages used in the dataset when publishing the metadata descriptions (VoID and/or DCAT dataset descriptions).

**Acknowledgments** This work has been supported by the BabelData (TIN2010-17550) and myBigData (TIN2010-17060) Spanish projects and by the FP7 Monnet (FP7-ICT-4-248458) and FP7 Lider (FP7-ICT-2013.4.1) European projects. The authors would also like to thank the editors and the anonymous reviewers for their valuable suggestions and Luis Vilches-Blázquez for providing the examples of geo.linkeddata.es.

## References

- Alexander, K., Cyganiak, R., Hausenblas, M., & Zhao, J. (2011). Describing linked datasets with the VoID vocabulary. W3C interest group note, W3C. <http://www.w3.org/TR/void/>
- Auer, S., Bühmann, L., Dirschl, C., Erling, O., Hausenblas, M., Isele, R., & Williams, H. (2012). Managing the life-cycle of linked data with the LOD2 Stack. *The Semantic Web-ISWC 2012* (pp. 1–16). Berlin: Springer.
- Auer, S., Lehmann, J., & Hellmann, S. (2009). Linkedgeodata: Adding a spatial dimension to the web of data. *The Semantic Web-ISWC 2009* (pp. 731–746). Berlin: Springer.
- Auer, S., Weidl, M., Lehmann, J., Zaveri, A. J., & Choi, K. S. (2010). I18n of semantic web applications. *The Semantic Web-ISWC 2010* (pp. 1–16). Berlin: Springer.
- Berners-Lee, T. (2006). Design issues: Linked Data (Last viewed 2013, April). Online resource available at <http://www.w3.org/DesignIssues/LinkedData>
- Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data-the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3), 1–22.
- Caracciolo, C., Stellato, A., Morshed, A., Johannsen, G., Rajbhandari, S., Jaques, Y., et al. (2013). The AGROVOC linked dataset. *Semantic Web Journal*, 4, 341–348.
- Cimiano, P., Buitelaar, P., McCrae, J., & Sintek, M. (2010). LexInfo: A declarative model for the lexicon-ontology interface. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(1), 29–51.
- Das, S., Sundara, S., & Cyganiak, R. (2012). R2RML: RDB to RDF mapping language. W3C Recommendation, World wide web consortium.
- Dunning, T. (1994). *Statistical identification of language. (Memoranda in computer and cognitive science)*. Computing Research Laboratory, New Mexico State University.
- Espinoza, M., Montiel-Ponsoda, E., & Gómez-Pérez, A. (2009). Ontology localization. *Proceedings of the 5th International Conference on Knowledge Capture (KCAP09)* (pp. 33–40).
- Ferrara, A., Nikolov, A., & Scharffe, F. (2011). Data linking for the semantic web. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 7(3), 46–76.
- Gómez-Pérez, A., Vila-Suero, D., Montiel-Ponsoda, E., Gracia J., & Aguado-de-Cea, G. (2013). Guidelines for multilingual linked data. *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics (WIMS '13)*. New York: ACM, Article 3, 12 pages.
- Gottron, T., & Lipka, N. (2010). A comparison of language identification approaches on short, query-style texts. In C. Gurrin, Y. He, G. Kazai, U. Kruschwitz, S. Little, T. Roelleke, S.M. Rüger, & K. van Rijsbergen (Eds.), *ECIR. Volume 5993 of Lecture Notes in Computer Science* (pp. 611–614). Berlin: Springer.

- Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Gómez-Pérez, A., Buitelaar, P., & McCrae, J. (2011). Challenges for the multilingual web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 11, 63–71.
- Heath, T., & Bizer, C. (2011). Linked Data: Evolving the Web into a Global Data Space (1st edition). *Synthesis Lectures on the Semantic Web: Theory and Technology*, 1:1, 1–136. Morgan & Claypool.
- Hyland, B., Villazón-Terrazas, B., & Atemezic, G. (2013). Best practices for publishing linked data. W3C Note 18 April 2013. <http://www.w3.org/TR/gld-bp/>.
- Isaac, A., & Haslhofer, B. (2013). European linked open data – data.europeana.eu. *Semantic Web Journal*, to appear. Available from <http://www.semantic-web-journal.net/>
- Kobilarov, G., Scott, T., Raimond, Y., Oliver, S., Sizemore, C., Smethurst, M., Bizer, C. & Lee, R. (2009). Media meets semantic web – how the BBC uses DBpedia and linked data to make connections. *Proceedings of the 6th European Semantic Web Conference on The Semantic Web: Research and Applications (ESWC 2009 Heraklion)* (pp. 723–737). Berlin: Springer.
- Labra Gayo, J. E., Kontokostas, D., & Auer, S. (2014) Multilingual linked open data patterns. *Semantic Web Journal* (to appear). <http://www.semantic-web-journal.net/>.
- Maali, F., Cyganiak, R., & Peristeras, V. (2012). A publishing pipeline for linked government data. In *The semantic web: Research and applications* (pp. 778–792). Berlin: Springer.
- Maali, F., Erickson, J., & Archer, P. (2013) Data catalog vocabulary (DCAT) W3C working draft 12 March 2013. <http://www.w3.org/TR/vocab-dcat/>
- McCrae, J., Aguado de Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., et al. (2012). Interchanging lexical resources in the Semantic Web. *Language Resources and Evaluation*, 46(4), 701–719.
- Montiel-Ponsoda, E., Gracia, J., Aguado de Cea, G., & Gómez-Pérez, A. (2011a). Representing translations on the Semantic Web. *Proceedings of the Workshop on the Multilingual Semantic Web, CEUR-Proceedings Vol. 775* (pp. 25–37).
- Montiel-Ponsoda, E., Vila-Suero, D., Villazón-Terrazas, B., Dunsire, G., Escolano Rodríguez, E., & Gómez-Pérez, A. (2011b). Style guidelines for naming and labeling ontologies in the multilingual Web. *Proceedings of the 2011 International Conference on Dublin Core and Metadata Applications, DCMI '11*, Dublin Core Metadata Initiative, 2011.
- Vila-Suero, D., & Gómez-Pérez, A. (2013). Datos.bne.es and MARiMbA: an insight into library linked data. *Library Hi Tech*, 31(4), 575–601.
- Vilches-Blázquez, L. M., Villazón-Terrazas, B., Corcho, O., & Gómez-Pérez, A. (2013). Integrating geographical information in the linked digital earth. *International Journal of Digital Earth*, 7(7), 554–575.
- Villazón-Terrazas, B., Vilches-Blázquez, L., Corcho, O., & Gómez-Pérez, A. (2011). Methodological guidelines for publishing government linked data. In D. Wood (Ed.), *Linking government data* (pp. 27–49). New York: Springer.
- Vossen, P. (2004). EuroWordNet: A multilingual database of autonomous and language-specific wordnets connected via an inter-lingual-index. *International Journal of Lexicography*, 17(2), 161–173.

# State-of-the-Art in Multilingual and Cross-Lingual Ontology Matching

Cássia Trojahn, Bo Fu, Ondřej Zamazal, and Dominique Ritze

**Abstract** Ontology matching is one of the key solutions for solving the heterogeneity problem in the Semantic Web. Nowadays, the increasing amount of multilingual data on the Web and the consequent development of ontologies in different natural languages have pushed the need for multilingual and cross-lingual ontology matching. This chapter provides an overview of multilingual and cross-lingual ontology matching. We formally define the problem of matching multilingual and cross-lingual ontologies and provide a classification of different techniques and approaches. Systematic evaluations of these techniques are discussed with an emphasis on standard and freely available data sets and systems.

**Key Words** Cross-lingual matching • Evaluation • Multilingual matching • Ontology matching

## 1 Introduction

As the amount of multilingual content on the Semantic Web and thus the number of vocabularies/ontologies in multiple languages continue to grow, methods for matching vocabularies across languages become more and more important in order to allow access to data in multiple languages to end users. As a motivating scenario involving querying data with vocabularies in different languages, let us consider

---

C. Trojahn (✉)

University of Toulouse 2 & IRIT, Toulouse, France

e-mail: [cassia.trojahn@irit.fr](mailto:cassia.trojahn@irit.fr)

B. Fu

University of Victoria, Victoria, BC, Canada

e-mail: [bofu@uvic.ca](mailto:bofu@uvic.ca)

O. Zamazal

University of Economics, Prague, Czech Republic

e-mail: [ondrej.zamazal@vse.cz](mailto:ondrej.zamazal@vse.cz)

D. Ritze

University of Mannheim, Mannheim, Germany

e-mail: [dominique@informatik.uni-mannheim.de](mailto:dominique@informatik.uni-mannheim.de)

the case of an Italian user that wants to retrieve parliament members and their political party membership from the DBpedia data set. But she/he is only familiar with “Ontologia Camera dei Deputati” (occd)<sup>1</sup> and has already prepared a SPARQL query for querying relevant data as defined by the occd<sup>2</sup>:

```
?persona rdf:type occd:deputato.
?persona occd:aderisce ?gruppo.
```

In order to query the DBpedia data set, another ontology must be used. For multilingual querying over these data, matching of the occd vocabulary and the DBpedia ontology can help to generate the following two correspondences:

- occd:deputato is equivalent to dbpedia-owl:MemberOfParliament
- occd:aderisce is equivalent to dbpedia-owl:party

The reformulated query<sup>3</sup> would look as follows:

```
?persona rdf:type dbpedia-owl:MemberOfParliament.
?persona dbpedia-owl:party ?party.
```

One approach to overcome the challenge of accessing distributed data and semantics across natural language barriers is by means of multilingual and cross-lingual ontology matching. Ontology matching (Euzenat and Shvaiko 2007) is a well-established research area aiming at developing methods for finding correspondences between ontological entities. Existing monolingual matching techniques typically rely on lexical comparisons made between names of entities (i.e. labels, descriptions, URI fragments, etc.), which limits their deployment to ontologies in the same natural language or at least in comparable natural languages<sup>4</sup> as demonstrated by Fu et al. (2009). This limitation coupled with the emergence of ontologies labelled in different natural languages results in a pressing need for the development of novel matching techniques for multilingual environments.

Systematic evaluation of these new techniques is thus another important aspect for the field of multilingual and cross-lingual ontology matching, as it may help designers and developers of such methods to improve the underlying techniques as well as help users to evaluate the suitability of the proposed methods to their specific needs. While systematic evaluation of monolingual ontology matching systems has

<sup>1</sup>The ontology is available at [http://dati.camera.it/occd/reference\\_document/](http://dati.camera.it/occd/reference_document/).

<sup>2</sup>occd: standing for <http://dati.camera.it/occd/> and dbpedia-owl: for <http://dbpedia.org/ontology/>.

<sup>3</sup>In order to get parliament members and their political party membership from the DBpedia.

<sup>4</sup>An example of comparable natural languages is English and German, both belonging to the Germanic language family. Comparable natural languages can also be languages that are not from the same language family. For example, Italian belonging to the Romance language family and German belonging to the Germanic language family can still be compared using string comparison techniques such as edit distance, as they are both alphabetic letter based with comparable graphemes. An example of natural languages that are not comparable in this context can be Chinese and English, where the former is logogram based and the latter is alphabetic letter based. In this chapter, we consider natural languages to be comparable when they contain graphemes that can be analysed using automated string comparison techniques.

been carried out extensively in the context of the Ontology Alignment Evaluation Initiative (OAEI) campaigns (Euzenat et al. 2011), evaluation of cross-lingual and multilingual matching systems has not received much attention until more recently (Aguirre et al. 2012; Meilicke et al. 2012).

This chapter provides a survey about multilingual and cross-lingual ontology matching. In particular, the different aspects of designing multilingual ontologies are presented (Sect. 2). The multilingual and cross-lingual ontology matching problems are defined (Sect. 3). Moreover, an overview and a classification of different techniques and approaches are presented (Sect. 4). In addition, systematic evaluations of these techniques (Sect. 5) are discussed with an emphasis on data sets which are open, freely available and frequently used, especially in evaluation campaigns. The chapter concludes with a discussion of the limitations and challenges in multilingual and cross-lingual ontology matching (Sect. 6).

## 2 Multilingualism on the Semantic Web

Most of the ontologies on the Semantic Web are in English, but ontologies in other languages have been appearing.<sup>5</sup> According to Gracia et al. (2012), the realisation of the multilingual Semantic Web is accelerated by ontology matching as well as by techniques to generate multilingual ontologies from monolingual ones.

One approach to generate multilingual ontologies for the Semantic Web is to translate or localise existing monolingual ontologies. While translation into another language is natural first choice for getting high-quality multilingual ontology variants, it should be further equipped with the localisation of an ontology (Espinosa et al. 2008). From the localisation perspective, an ontology consists of a *lexical* layer and a *conceptualisation* layer (Cimiano et al. 2010). The lexical layer contains labels and names of entities in a natural language, which is affected by the localisation process. On the contrary, the conceptualisation layer can remain the same after the localisation process although it may be adapted given the specific cultural or geopolitical context (e.g. law, organisation of countries, universities, etc.).

On the one hand, there are some efforts to linguistically enrich ontologies (Pazienza and Stellato 2006; Buitelaar et al. 2009; Montiel-Ponsoda et al. 2011) and on the other hand to provide richer models for associating linguistic (and also potentially multilingual) information with ontology entities. McCrae et al. (2011) proposed the *lemon* model (Lexicon Model for Ontologies) for promoting sharing of terminological and lexicon resources on the Semantic Web (more details are given by León-Araúz and Faber [this volume] and McCrae et al. [this volume]). It allows lexical information to be represented relative to an ontology. In order to be more

---

<sup>5</sup>For example, in the case of 329 “Linked Open Vocabularies” at <http://lov.okfn.org/>, there are 271 vocabularies in English, 23 in French, 17 in German, 17 in Spanish, etc. Growing trend of multilinguality is documented by Vila-Suero et al. (this volume).

flexible and provide arbitrarily complex linguistic information, it is possible to link to linguistic concepts described elsewhere, e.g. in *LexInfo* ontology. *LexInfo* has been proposed by Cimiano et al. (2011) as a joint model for linguistic grounding of ontologies.

### 3 Monolingual, Multilingual and Cross-Lingual Ontology Matching Definitions

Ontology matching is defined as “a function  $f$  which, from a pair of ontologies to match  $o$  and  $o'$ , an input alignment  $A$ , a set of parameters  $p$  and a set of oracles and resources  $r$ , returns an alignment  $A'$  between these ontologies:  $A' = f(o, o', A, p, r)$ ” (Euzenat and Shvaiko 2007). Extending Euzenat and Shvaiko’s definition, natural languages used in the ontologies can be defined as:

- $L\langle o \rangle$  is the set of natural languages used for entities in  $o$ ,
- $L'\langle o' \rangle$  is the set of natural languages used for entities in  $o'$ .

Inspired by related definitions in established fields such as multilingual and cross-lingual information retrieval (Peters et al. 2012), this section defines *monolingual*, *multilingual* and *cross-lingual ontology matching* as follows.

**Definition 1 (Monolingual Ontology Matching).** *In monolingual ontology matching, the ontologies involved use a single shared natural language to name the entities or, more formally,  $(L = L') \wedge (|L| = |L'| = 1)$ .*

The terms *multilingual* and *cross-lingual ontology matching* are often ambiguously defined and interchangeably used in the literature, although others have attempted to define the problem (Spohr et al. 2011). In this chapter, *multilingual ontology matching* techniques refer to those that are concerned with establishing relationships among ontological entities labelled in multiple natural languages, where the matching process can encompass both monolingual matching and those carried out across natural languages. *Cross-lingual ontology matching* is a special case of multilingual ontology matching, which refers specifically to those techniques that concern the matching of source ontologies in one natural language (or one set of natural languages) to target ontologies in other natural languages (or other sets of natural languages).

**Definition 2 (Multilingual Ontology Matching).** *In multilingual ontology matching, the ontologies involved can either share no common natural language, or they can share common natural language(s), but at least one ontology contains two or more natural languages within itself or thus, more formally,  $(L \cap L' = \emptyset)$  or  $(L \cap L' \neq \emptyset) \wedge (|L| > 1 \vee |L'| > 1)$ .*

**Definition 3 (Cross-Lingual Ontology Matching).** *In cross-lingual ontology matching, each ontology uses a different natural language (or a different set of natural languages) or, more formally,  $(L \cap L' = \emptyset)$ . In other words, the ontologies involved do not have any natural language in common. It is thus a special case of multilingual ontology matching (as defined in Definition 2), and in theory, cross-lingual matching techniques can be reused for multilingual matching scenarios.*

## 4 Multilingual and Cross-Lingual Matching Approaches

Several monolingual matching approaches have been introduced in recent years. Extensive reviews and classifications of these approaches have been proposed (Rahm and Bernstein 2001; Kalfoglou and Schorlemmer 2003; Shvaiko and Euzenat 2005; Euzenat and Shvaiko 2007). Broadly speaking, these approaches can be classified based on the many features that can be found in ontologies (labels, structures, instances, semantics) or with regard to the disciplines they belong to (e.g. statistics, combinatorics, semantics, linguistics, machine learning or data analysis). Despite the variety of approaches, most of them typically rely on string-based lexical comparisons of entity names whereby an initial estimate of the likelihood that two elements refer to the same real-world phenomenon is provided. However, this lexicon comparison restricts these matching techniques to ontologies that are labelled in the same or comparable natural languages. Consequently, there is a pressing need for matching techniques that are designed to work with ontologies in multilingual environments. Existing multilingual matching techniques can be broadly grouped into the following categories, extending Fu et al.'s classification (Fu et al. 2012),<sup>6</sup> as *manual processing*, *corpus-based approach*, *linguistic enrichment*, *instance-based approach*, *translation-based approach*, *machine learning-based*, *indirect alignment composition*, and *image similarity-based approach*.

A *manual* cross-lingual matching process is presented by Liang and Sini (2006), where the English version of the AGROVOC thesaurus is manually aligned to the Chinese Agriculture Thesaurus. A similar approach has been used for creating MultiFarm (Meilicke et al. 2012), a benchmark for cross-lingual matching that results from the manual translations of a set of ontologies from the conference domain (Sváb-Zamazal et al. 2005) into eight natural languages. Landry (2009) reports the manual construction of an extensible set of correspondences between cross-lingual subject heading lists (SHL) used for indexing book collections. Although such approaches may guarantee high-quality matches, it can be infeasible and unscalable when dealing with large and complex ontologies.

---

<sup>6</sup>These approaches are not exclusive and might overlap. This classification takes into account the kind of technique used (manual, translation, learning, etc.) and resources involved (dictionaries, corpora, etc.).

Using external background resources to assist the matching process, a *corpus-based* approach has been proposed by Ngai et al. (2002), where the English thesaurus WordNet is aligned to the Chinese thesaurus HowNet, using a Chinese-English bilingual corpus. The proposal of Cheng et al. (2008) is to use a corpus from a domain similar to the domain of the ontologies to be aligned, where co-occurrence frequency of two concepts in the corpus acts as a means to compute the relatedness between them. Although applied in a monolingual context, this proposal may be potentially exploited in a cross-lingual context. Eger and Sejane (2010) calculate monolingual and bilingual semantic similarities exploiting the information from bilingual dictionaries. A similar approach is adopted by Mohammad et al. (2007). However, a limitation of such approaches is that bilingual corpora may not be available for domain-specific ontologies.

More recently, the Web has been used as a corpus of background knowledge. Lin and Krizhanovsky (2011) use Wiktionary<sup>7</sup> as a source of lexical background knowledge used to match English and French ontologies. Bouma (2010) exploits the cross-lingual links in Wikipedia as an intermediate resource for linking the thesaurus of the Netherlands Institute for Sound and Vision to English WordNet and DBpedia ontologies (i.e. the strategy used for the Gg2www system Sect. 5.2). Wikipedia is also used by Beisswanger (2010) as well as by Hertling and Paulheim (2012). Beisswanger (2010) uses Wikipedia as a large-scale text corpus for extracting different types of semantic relations between concepts, requiring rich NLP tools for text processing. Hertling and Paulheim (2012) propose the WikiMatch system, which exploits Wikipedia's interlanguage links for finding corresponding ontology elements. A similar strategy is adopted by Hassan and Mihalcea (2009).

A *linguistically* motivated mapping method is proposed by Paziienza and Stellato (2005), who advocate a linguistic-driven approach within the ontology development process for generating enriched ontologies with human-readable linguistic resources. Linguistically enriched ontologies may offer strong evidence when generating matching correspondences.

While the approaches above are mainly based on the TBox (the terminological component describing the concept hierarchy) of the ontologies, *instance-based* approaches exploit the ABox (the assertional component containing the knowledge about instances) level. Wang et al. (2009) apply instance similarity metrics in order to determine correspondences between SHL in different languages. The method makes use of book collections that are annotated with subjects having joint instances (shared books), which can be determined by common ISBN numbers, for instance. Despite the fact that this kind of approach does not depend neither on terminological similarities of concept labels nor on rich semantic structure, it requires rich sets of instances embedded in ontologies, which is a condition that may not always be satisfied in the ontology development process. On the other hand, as instances are not limited to numeric descriptions and may be described by labels in different languages, matching them requires multilingual or cross-lingual strategies.

---

<sup>7</sup>[www.wiktionary.org](http://www.wiktionary.org).



Translation techniques are typically used to overcome the natural language barriers presented in cross-lingual and multilingual matching scenarios (Trojahn et al. 2008; Aguirre et al. 2012; Fu et al. 2012). Trojahn et al. (2008) translate the ontologies using machine translation before applying monolingual matching methods. This translation-based matching approach has been largely adopted by participants of OAEI 2012 (Ase, Automsv2, Gomma, Wesee and Yam++) which use English as pivot language. A limitation of this approach is that inadequate translations can introduce “noise” into the subsequent matching step, where matches may be neglected by matching techniques that (solely) rely on the discovery of lexical similarities. This is examined by Fu et al. (2009), where strong evidence indicates that to enhance the performance of existing monolingual matching techniques in cross-lingual scenarios, appropriate ontology entity name translation is key to the generation of high-quality matching results. Selecting appropriate entity name translations in the specific mapping context is the focus of the SOCOM (Semantic-Oriented Cross-lingual Ontology Mapping) (Fu et al. 2010) and SOCOM++ frameworkS (Fu et al. 2012), where users can adjust and manipulate the translation outcome in an effort to improve the matching results. In particular, the benefits of pseudo feedback (similar to relevance feedback used in the field of information retrieval) to improve the quality of cross-lingual matching results are explored in (Fu et al. 2011).

A *machine learning* approach to multilingual and cross-lingual matching is proposed by Spohr et al. (2011). The proposal relies on machine learning techniques (i.e. essentially support vector machines) to learning a matching function between two ontologies. A requirement of this approach is to have manually aligned concepts as training sets, as well as features representing the characteristics of each possible correspondence. These manually aligned concepts might not always be available, and their creation can be very time-consuming. The *indirect composition* approaches are based on the existence of alignments that can be composed. Jung et al. (2009) propose an indirect composition approach that uses existing intermediary alignments between ontologies to compose new alignments. For instance, an alignment between French and Portuguese ontologies can be generated if intermediary alignments between these two ontologies and a third one (i.e. English) are available. However, it depends on the availability of good-quality alignments, which can be difficult to come by at times.

Finally, an *image similarity-based* approach has been proposed by Mihic and Ivetic (2012), where cross-lingual ontology matching is based on a similarity measure between images associated to the entities to be matched. Following this approach, two ontology entities, labelled in different natural languages (i.e. “river” and “rio”(pt)), are similar if the images within documents containing these labels are similar. It is however challenging to find images that accurately illustrate ontological entities in various domains, particularly for properties and instances.

## 5 Evaluation of Multilingual and Cross-Lingual Matching Approaches

Attempts aiming at evaluating the ability of systems to deal with multilingual and cross-lingual matching have been carried out since 2006 in the context of OAEI. This section provides a comprehensive overview of the data sets, systems and strategies used in those campaigns, but not limiting the overview to OAEI resources. The aim is to present the different available resources that can be used for multilingual and cross-lingual evaluation purposes. This will help to better understand their advantages and drawbacks with respect to the concrete tasks for which they are applicable.

### 5.1 Data Sets for Multilingual and Cross-Lingual Matching

Evaluation of multilingual and cross-lingual matching has been mainly based on subsets of real-world resources, not particularly designed for evaluation purposes. Only recently, benchmarks for evaluating multilingual and cross-lingual matching have been proposed. A brief description and a comparison between these data sets are presented below. This overview also includes a discussion of the data sets proposed outside the context of OAEI.

*Food and Agriculture Organization.* The FAO data set is about matching an SKOS version of part of the United Nations Food and Agriculture Organization (FAO) AGROVOC multilingual thesaurus with the United States National Agricultural Library (NAL) Agricultural thesaurus (monolingual). This data set has been firstly used in the context of the OAEI 2006 campaign.<sup>8</sup>

*Environment Data Set.* This data set<sup>9</sup> contains a thesaurus alignment task that requires to align three SKOS thesauri using relations from the SKOS mapping vocabulary. The thesauri are versions of the European Environment Agency (EEA), GEMET (General Multilingual Environmental Thesaurus) multilingual thesaurus, the United Nations (FAO), AGROVOC thesaurus and the United States (NAL) Agricultural thesaurus. It has been proposed once in OAEI 2007.

---

<sup>8</sup><http://oaei.ontologymatching.org/2006/food/>.

<sup>9</sup><http://oaei.ontologymatching.org/2007/environment/>.

*National Library of the Netherlands Thesaurus (KB)*. The KB data set involves matching two Dutch thesauri used to index books from two collections held by the National Library of the Netherlands (KB). The scientific collection is described using the GTT, a huge vocabulary containing general concepts, while the books contained in the deposit collection are indexed against the Brinkman thesaurus, with a large set of headings that are expected to serve as global subjects of books. The language of both thesauri is Dutch, with a substantial part of Brinkman concepts (around 60 %) having English labels. This data set has been proposed in OAEI 2007<sup>10</sup> and 2008.

*Multilingual Directory Data Set (MLdirectory)*. The MLdirectory (proposed in OAEI 2008)<sup>11</sup> data set contains different Internet directories (having classes and instances) such as *Google* (open directory project), *Yahoo!*, *Lycos Japan*, and *Yahoo! Japan*. It covers five domains: cars, movie, outdoor, photo and software.

*Very Large Crosslingual Resources (vlcr)*. Proposed in the OAEI 2008 campaign, this data set contains very large resources available on the Web: DBpedia, WordNet and the Thesaurus of the Netherlands Institute for Sound and Vision (GTAA).<sup>12</sup> The GTAA is in Dutch, while WordNet is in English. DBpedia contains labels in both languages. This data set has been further evaluated in 2009.

*Subject Heading Lists*. In 2009, the library task proposed to align three large SHL: LCSH, the Library of Congress Subject Headings; RAMEAU, the heading list used at the French National Library; and the SWD, the heading list used at the German National Library.<sup>13</sup>

*MultiFarm Data Set*. The lack of benchmarks for automatic evaluation of cross-lingual<sup>14</sup> matching systems has motivated the creation of the MultiFarm data set (Meilicke et al. 2012). This data set results from the manual translation of seven ontologies of the conference domain (Sváb-Zamazal et al. 2005) into eight languages (cn, cz, nl, fr, de, pt, ru, es). It has been used in OAEI campaigns since 2012.<sup>15</sup>

---

<sup>10</sup><http://oaei.ontologymatching.org/2007/library/>.

<sup>11</sup><http://oaei.ontologymatching.org/2008/mldirectory/>.

<sup>12</sup><http://oaei.ontologymatching.org/2008/vlcr/>.

<sup>13</sup><http://oaei.ontologymatching.org/2009/library/>.

<sup>14</sup>In this chapter, we have revised the definitions of multilingual and cross-lingual terms. Contrary to what is reported in Meilicke et al. (2012) MultiFarm is a benchmark for cross-lingual ontology matching.

<sup>15</sup><http://oaei.ontologymatching.org/2012/multifarm/>.

*Library Data Set.* For the OAEI 2012 library data set<sup>16</sup>, the STW (Standard Thesaurus Wirtschaft) Thesaurus for Economics and the Thesaurus for the Social Sciences (TheSoz) were taken since they are of a comparable size, cover overlapping domains, are often used by libraries and a (incomplete) reference alignment already exists (Mayr and Petras 2008). STW (6,573 classes with  $\varnothing$  3,35 German and  $\varnothing$  1,62 English labels) provides a vocabulary on any economic subject, while TheSoz (8,378 classes with  $\varnothing$  1,73 German and  $\varnothing$  1,56 English labels) covers all topics related to social sciences. Originally, they are available as SKOS, but we transformed them into OWL because ontology matching systems are not specialised to match SKOS (Aguirre et al. 2012).

*Financial Accounting Standards.* Spohr et al. (2011) used a data set dealing with financial accounting standards (FAS). They are annotated in more than one language and thus represent the problem of multilingual interoperability of financial information. FAS are captured as *taxonomies*. This data set has not been used in OAEI. A similar financial data set was used by Thomas et al. [this volume].

Based on Table 1 summarising the data set features, we can conclude that although there are already many diverse (w.r.t. domain, language and size coverage) data sets for experimenting with multilingual tasks, only less than half have a reference-alignment and only one of them is not freely available (FAS).

## 5.2 Matching Systems and Strategies

Around 30 different systems have been evaluated on the data sets listed in Table 1. The reader can refer to Thomas et al. [this volume] for an approach that matches FAS using domain ontologies and reasoning. However, only few systems (30%) implement some strategy to deal with multilingualism. As shown in Fig. 1, the first evaluation campaigns considered systems which do not include a multilingual or cross-lingual matching component. The strategy used for most of them was to preserve labels and comments in a single natural language (English or Dutch, for instance) in both input ontologies and to apply classical monolingual approaches (terminological and structural similarities). In 2008, the system *Rimon* (Li et al. 2009) introduced a first proposal on cross-lingual matching which applied a *translation-based approach* (Sect. 4). A similar approach has been used by Taxomap in 2009, while GG2WW has adopted a *corpus-based approach* which exploited the EuroWordNet thesaurus and the Dutch Wikipedia as background resources (Euzenat et al. 2009). In 2010 and 2011, the cross-lingual and multilingual tracks were discontinued, especially due to the insufficient number of participants.

---

<sup>16</sup><http://web.informatik.uni-mannheim.de/oaei-library/2012/>

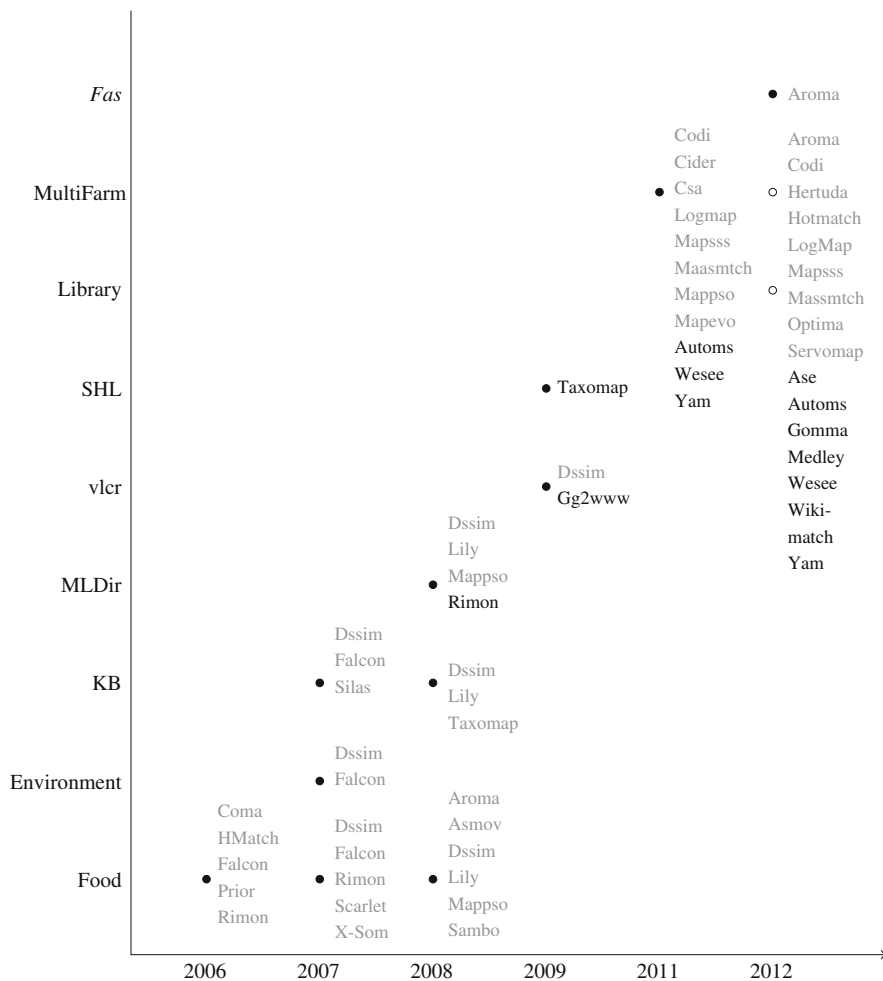
**Table 1** Comparison of data sets for cross-lingual and multilingual matching

Data set	Domain	NL coverage	Formalism	Size	Num. ontos	Problem	Eval.
Environment	Environment Agriculture	ar, bg, cs, da, de, el, en, es, et, eu, fi, fr, ja, hu, it, nl, no, pl, pt, ru, sk, sl, sv, th, zh	SKOS, OWL	Large	3	ML	Sample
FAO	Agriculture Food Fishery	en, fr, es, ar, zh, pt, cs, ja, th, sk	OWL	Medium	3	ML	Sample
Fas	Finance	en, fr, de, it	XML	Small to medium	3	ML	RA
KB	General	en, de	SKOS, OWL	Large	2	ML/CL	Sample
Library	Library	en, de, fr	SKOS, OWL	Medium	2	ML	RA Manual
MLdirectory	Cars, movie, outdoor, photo, software	en, jp	OWL	Medium	5	CL	Manual
MultiFarm	Conference	cn, cz, de, en, es, fr, nl, pt, ru	OWL	Small	7	CL	RA
SHL	Library	en, fr, de	SKOS	Large	3	CL	RA Sample
v1cr	General	en, nl	SKOS, OWL	Large	3	ML/CL	Sample

Up to 1,000 concepts, the data set is considered *small*, 1,001 to 10,000 *medium* and more than 10,000 *large*. We also consider whether there is *reference alignment* (RA) available for a data set or there was applied *manual-* or *sample-based* evaluation approach

Motivated by the lack of benchmarks for cross-lingual matching evaluation, the MultiFarm data set was offered for a first time in 2011 (OAEI 2011.5)<sup>17</sup> and again in 2012. In the OAEI 2011.5 campaign, only three participating systems (Wesee, Automsv2 and Yam++) used specific methods to handle cross- or multilinguality at all. All of them apply a *translation-based approach*. As we observed in 2012, a

<sup>17</sup><http://oaei.ontologymatching.org/2011.5/>.



**Fig. 1** Overview of matching systems participating on OAEI cross-lingual and multilingual tracks and outside OAEI campaigns (Fas). *Light grey colour* indicates the systems that do not use any kind of specific strategy to deal with multilingualism. 2011 refers to the OAEI 2011.5 intermediary campaign. In 2012, MultiFarm and Library counted with the same set of participants

progress has been taken place: seven systems (out of 24 in OAEI 2012) proposed specific methods for dealing with multilinguality: Ase, Automsv2, Gomma, Medley, Wesee, Wmatch and Yam++. Most of these systems apply a *translation-based approach*, using English as pivot language. Only Wmatch uses a *corpus-based approach* and exploits Wikipedia for extracting interlanguage links. As expected, specific methods for dealing with ontologies that are described in different languages work much better than nonspecific systems. However, the absolute results ( $\cong 0.40$  *F*-measure for the best matcher) are still not very good compared to the best

results in the original conference data set ( $\cong 0.75$  *F*-measure), thus leaving ample room for improvements.

### 5.3 Experimenting on Cross-Lingual and Multilingual Cases

To see whether current state-of-the-art matching systems deal with cross-lingual and multilingual ontologies, we performed a few experiments by applying matching systems on the library track. This data set and systems are available for us in the context of OAEI 2012. We chose this data set because it was not explicitly announced as multilingual matching in the OAEI but the ontologies contain labels in multiple languages. Since each ontology has German as well as English labels, we could simulate both cross-lingual and multilingual matching tasks.

The results are listed in Table 2. We let the systems match the ontologies only with English labels (EN) and only with German labels (DE). Further, we united the former two alignments ( $EN \cup DE$ ) and performed the usual matching task where all labels are included (ALL). When only the German labels are taken into account, all values (precision, recall, *F*-measure) are higher compared to the ontologies with the English labels. This has two reasons: on average more German labels are available per class, and the reference alignment has been created by German domain experts. Thus, if the matching systems only have a look at the English labels, their alignments will result in a low *F*-measure. If we take the union of both alignments ( $EN \cup DE$ ), the *F*-measure values are in most cases (for 10 of 13 systems) even better than the *F*-measure values when all labels are available (ALL). Especially

**Table 2** Results library track

Matcher	EN			DE			EN $\cup$ DE			ALL		
	Pre	Rec	Fmeas	Pre	Rec	Fmeas	Pre	Rec	Fmeas	Pre	Rec	Fmeas
AROMA	0.13	0.46	0.20	0.14	0.66	0.23	0.11	0.75	0.19	0.12	0.66	0.21
CODI	0.37	0.29	0.33	0.54	0.50	0.52	0.40	0.58	0.47	0.50	0.49	0.49
GOMMA	0.68	0.56	0.62	0.64	0.87	0.74	0.60	0.90	0.72	0.60	0.90	0.72
Hertuda	0.62	0.58	0.60	0.58	0.88	0.70	0.53	0.94	0.68	0.52	0.92	0.67
HotMatch	0.76	0.46	0.57	0.78	0.59	0.67	0.71	0.71	0.71	0.74	0.58	0.65
LogMap	0.76	0.50	0.61	0.77	0.75	0.76	0.71	0.82	0.76	0.78	0.65	0.71
LogMapLt	0.66	0.50	0.57	0.70	0.78	0.74	0.62	0.83	0.71	0.65	0.77	0.71
MapSSS	0.58	0.19	0.29	0.64	0.21	0.32	0.58	0.33	0.42	0.59	0.18	0.28
Optima	0.35	0.09	0.14	0.45	0.05	0.09	0.33	0.11	0.16	0.39	0.08	0.13
ServOMap	0.76	0.44	0.56	0.82	0.62	0.70	0.74	0.71	0.73	0.83	0.63	0.72
ServOMapL	0.71	0.50	0.58	0.78	0.67	0.72	0.69	0.76	0.73	0.75	0.70	0.72
WeSeE*	0.66	0.51	0.57	0.72	0.63	0.67	0.62	0.74	0.68	0.70	0.62	0.66
YAM++*	0.71	0.54	0.61	0.73	0.75	0.74	0.65	0.83	0.73	0.68	0.76	0.72

Systems marked with \* implement multilingual methods

the recall can be strongly increased, e.g. up to 17 % for LogMap. The main reason is that the matching systems usually compare the labels with each other, no matter in which language they are, compute some similarity measures and combine them. Often, labels in different languages are not quite similar, which results in a low similarity over all labels. For example, the concept  $c_1$  has four German labels “Altstadtsanierung”, “Stadterneuerung”, “Stadtsanierung” and “Stadtteilsanierung” as well as the English labels “Urban regeneration” and “Urban renewal”. The second concept  $c_2$  has the German label “Stadterneuerung” and the English label “urban renewal”. If all labels are directly compared without any translation, no correspondence is generated, although it would be a correct one. When only the English labels are taken into account, the correspondence is found since the overall similarity is much higher. Besides the recall, also the precision can be increased. Again, assume two concepts  $c_1$  with German labels “Binnensee” and “See” and the English label “lake”. Concept  $c_2$  has the German label “Sehen” and the English label “see”. Several matching systems create the incorrect correspondence between  $c_1$  and  $c_2$  based on the exact match of an English and a German label. However, the correspondence is incorrect, since the German word “See” is not the same as the English one. If only English labels are compared among each other, this error does not occur. Nevertheless, the precision of the merged alignments is in several cases lower than the precision of the original matching task. When the alignments are only merged after the matching itself, the systems cannot perform any filtering, e.g. to discard correspondences if an entity is already included in other correspondences with a higher confidence.

Based on our experiments, we can observe that state-of-the-art systems are still not able to properly tackle the case where ontologies have labels in different languages, whether shared (multilingual case) or not (cross-lingual case). In this specific setting, they even perform worse given labels in different languages although it is an additional information which could be exploited to improve the alignments.

## 6 Challenges and Future Work

To date, despite the ongoing effort thus far in developing various techniques, there is not a clear winner in solving multilingual and cross-lingual matching problems. As corroborated on recent evaluations, most of the approaches are focused on automatic translation. Novel techniques remain central to the innovation and advancement of the multilingual Semantic Web. In addition, tools that support interactions between the user and the matching process are yet to be developed, as well as infrastructures and repositories for searching, sharing and reusing existing cross-lingual and multilingual alignments.

Evaluation of new techniques designed specifically for multilingual and cross-lingual scenarios depends on the availability of multilingual data sets that are accompanied by reliable reference alignments. Although there has been an increas-



ing effort in making such data sets accessible to the research community, additional data sets accompanied by readily available reference alignments involving wider domain coverage and additional natural languages are essential for the improvement of multilingual and cross-lingual matching techniques. In addition, techniques to generate multilingual ontologies from monolingual ones have also to be taken into account.

On the one hand, as stated by Gracia et al. (2012), unexplored background knowledge sources such as the Linked Open Data and the whole Web as a big corpus can potentially be used to assist the cross-lingual and multilingual matching processes. On the other hand, this brings the need for scalable matching systems and systematic evaluation of these systems with regard to this dimension.

**Acknowledgements** Cassia Trojahn is partially supported by the CAPES-COFECUB Cameleon project number 707-11. Ondřej Zamazal has been supported by the CSF grant no. 14-14076P.

## References

- Aguirre, J. L., Eckert, K., Euzenat, J., Ferrara, A., van Hage, W. R., Hollink, L., et al. (2012). Results of the ontology alignment evaluation initiative 2012. In *Proceedings of the 7th International Workshop on Ontology Matching* (pp. 73–115).
- Beisswanger, E. (2010). Exploiting relation extraction for ontology alignment. In *Proceedings of the 9th International Semantic Web Conference* (pp. 289–296).
- Bouma, G. (2010). Cross-lingual ontology alignment using EuroWordNet and Wikipedia. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*. European Language Resources Association.
- Buitelaar, P., Cimiano, P., Haase, P., & Sintek, M. (2009). Towards linguistically grounded ontologies. In *Proceedings of the 6th European Semantic Web Conference* (pp. 111–125).
- Cheng, C., Lau, G., Pan, J., Law, K., & Jones, A. (2008). Domain-specific ontology mapping by corpus-based semantic similarity. In *Proceedings of 2008 Engineering Research and Innovation Conference*.
- Cimiano, P., Buitelaar, P., McCrae, J., & Sintek, M. (2011). LexInfo: A declarative model for the lexicon-ontology interface. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(1), 29–51.
- Cimiano, P., Montiel-Ponsoda, E., Buitelaar, P., Espinoza, M., & Gómez-Pérez, A. (2010). A note on ontology localization. *Applied Ontology*, 5(2), 127–137.
- Eger, S., & Sejane, I. (2010). Computing semantic similarity from bilingual dictionaries. In *Proceedings of the 10th International Conference on the Statistical Analysis of Textual Data* (pp. 1217–1225).
- Espinoza, M., Gómez-Pérez, A., & Mena, E. (2008). LabelTranslator: A tool to automatically localize an ontology. In *Proceedings of the 5th European Semantic Web Conference* (pp. 792–796).
- Euzenat, J., Ferrara, A., Hollink, L., Isaac, A., Joslyn, C., Malaisé, V., et al. (2009). Results of the ontology alignment evaluation initiative 2009. In *Proceedings of the 4th International Workshop on Ontology Matching* (pp. 73–126).
- Euzenat, J., Meilicke, C., Stuckenschmidt, H., Shvaiko, P., & Trojahn, C. (2011). Ontology alignment evaluation initiative: Six years of experience. *Journal on Data Semantics*, XV, 158–192.
- Euzenat, J., & Shvaiko, P. (2007). *Ontology matching*. New York: Springer.

- Fu, B., Brennan, R., & O'Sullivan, D. (2009). Cross-lingual ontology mapping: An investigation of the impact of machine translation. In *Proceedings of the 4th Annual Asian Semantic Web Conference* (pp. 1–15).
- Fu, B., Brennan, R., & O'Sullivan, D. (2010). Cross-lingual ontology mapping and its use on the multilingual semantic web. In *Proceedings of the 1st International Workshop on the Multilingual Semantic Web* (pp. 13–20).
- Fu, B., Brennan, R., & O'Sullivan, D. (2011). Using pseudo feedback to improve cross-lingual ontology mapping. In *Proceedings of the 8th Extended Semantic Web Conference* (pp. 336–351). Berlin: Springer.
- Fu, B., Brennan, R., & O'Sullivan, D. (2012). A configurable translation-based cross-lingual ontology mapping system to adjust mapping outcomes. *Web Semantics: Science, Services and Agents on the World Wide Web*, 15, 15–36.
- Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Gómez-Pérez, A., Buitelaar, P., & McCrae, J. (2012). Challenges for the multilingual web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 11, 63–71.
- Hassan, S., & Mihalcea, R. (2009). Cross-lingual semantic relatedness using encyclopedic knowledge. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (pp. 1192–1201). Stroudsburg, PA, USA: ACL.
- Hertling, S., & Paulheim, H. (2012). WikiMatch: Using wikipedia for ontology matching. In *Proceedings of the 7th International Workshop on Ontology Matching* (pp. 37–48).
- Jung, J. J., Håkansson, A., & Hartung, R. (2009). Indirect alignment between multilingual ontologies: A case study of Korean and Swedish ontologies. In *Proceedings of the 3rd KES International Symposium on Agent and Multi-Agent Systems* (pp. 233–241). Berlin: Springer.
- Kalfoglou, Y., & Schorlemmer, M. (2003). Ontology mapping: The state of the art. *Knowledge Engineering Review*, 18(1), 1–31.
- Landry, P. (2009). Multilingualism and subject heading languages: How the MACS project is providing multilingual subject access in Europe. *Catalogue & Index*, 157, 9–11.
- Li, J., Tang, J., Li, Y., & Luo, Q. (2009). RiMOM: A dynamic multi-strategy ontology alignment framework. *IEEE Transactions on Knowledge and Data Engineering*, 21(8), 1218–1232.
- Liang, A. C., & Sini, M. (2006). Mapping agrovoc and the chinese agricultural thesaurus: Definitions, tools, procedures. *The New Review of Hypermedia and Multimedia*, 12(1), 51–62.
- Lin, F., & Krizhanovsky, A. (2011). Multilingual ontology matching based on Wiktionary data accessible via SPARQL endpoint. In *Proceedings of the 13th All-Russian Conference Digital Libraries: Advanced Methods and Technologies, Digital Collections* (pp. 19–22).
- Mayr, P., & Petras, V. (2008). Building a terminology network for search: The KoMoHe project. In *Proceedings of the Conference on Dublin Core and Metadata Applications* (pp. 177–182).
- McCrae, J., Spohr, D., & Cimiano, P. (2011). Linking lexical resources and ontologies on the semantic web with lemon. In *The semantic web: Research and applications* (pp. 245–259). Heidelberg: Springer.
- Meilicke, C., Garcia-Castro, R., Freitas, F., van Hage, W. R., Montiel-Ponsoda, E., de Azevedo, R. R., et al. (2012). MultiFarm: A benchmark for multilingual ontology matching. *Journal on Web Semantics*, 15, 62–68.
- Mihic, S., & Ivetic, D. (2012). Multilingual ontology alignment based on visual representations of ontology concepts. In *Proceedings of the 5th International Conference on Advances in Computer-Human Interaction* (pp. 101–105).
- Mohammad, S., Gurevych, I., Hirst, G., & Zesch, T. (2007). Cross-lingual distributional profiles of concepts for measuring semantic distance. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 571–580).
- Montiel-Ponsoda, E., de Cea, G. A., Gómez-Pérez, A., & Peters, W. (2011). Enriching ontologies with multilingual information. *Natural Language Engineering*, 17(3), 283–309.
- Ngai, G., Carpuat, M., & Fung, P. (2002). Identifying concepts across languages: A first step towards a corpus-based approach to automatic ontology alignment. In *Proceedings of the 19th International Conference on Computational Linguistics* (pp. 1–7).

- Pazienza, M. T., & Stellato, A. (2005). Linguistically motivated ontology mapping for the semantic web. In *Proceedings of the 2nd Italian Semantic Web Workshop*.
- Pazienza, M. T., & Stellato, O. (2006). Linguistic enrichment of ontologies: A methodological framework. In *Proceedings of the 2nd Workshop on Interfacing Ontologies and Lexical Resources for Semantic Web Technologies*.
- Peters, C., Braschler, M., & Clough, P. (2012). *Multilingual information retrieval: From research to practice*. New York: Springer.
- Rahm, E., & Bernstein, P. A. (2001). A survey of approaches to automatic schema matching. *The VLDB Journal*, 10(4), 334–350.
- Shvaiko, P., & Euzenat, J. (2005). A survey of schema-based matching approaches. *Journal on Data Semantics*, 4, 146–171.
- Spohr, D., Hollink, L., & Cimiano, P. (2011). A machine learning approach to multilingual and cross-lingual ontology matching. In *Proceedings of the 10th International Semantic Web Conference* (pp. 665–680). Berlin: Springer.
- Sváb-Zamazal, O., Svátek, V., Berka, P., Rak, D., & Tomášek, P. (2005). OntoFarm: Towards an experimental collection of parallel ontologies. In *Poster Proceedings of the 4th International Semantic Web Conference*.
- Trojahn, C., Quaresma, P., & Vieira, R. (2008). A framework for multilingual ontology mapping. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, 1034–1037.
- Wang, S., Isaac, A., Schopman, B., Schlobach, S., & Meij, L. (2009). Matching multi-lingual subject vocabularies. In *Research and advanced technology for digital libraries* (pp. 125–137). Heidelberg: Springer.

# Mind the Cultural Gap: Bridging Language-Specific DBpedia Chapters for Question Answering

Elena Cabrio, Julien Cojan, and Fabien Gandon

**Abstract** In order to publish information extracted from language-specific pages of Wikipedia in a structured way, the Semantic Web community has started an effort of internationalization of DBpedia. Language-specific DBpedia chapters can contain very different information from one language to another; in particular, they provide more details on certain topics or fill information gaps. Language-specific DBpedia chapters are well connected through instance interlinking, extracted from Wikipedia. An alignment between properties is also carried out by DBpedia contributors as a mapping from the terms in Wikipedia to a common ontology, enabling the exploitation of information coming from language-specific DBpedia chapters. However, the mapping process is currently incomplete, it is time-consuming as it is performed manually, and it may lead to the introduction of redundant terms in the ontology. In this chapter, we first propose an approach to automatically extend the existing alignments, and we then present an extension of QAKiS, a system for Question Answering over Linked Data that allows to query language-specific DBpedia chapters relying on the abovementioned property alignment. In the current version of QAKiS, English, French, and German DBpedia chapters are queried using a natural language interface.

**Key Words** DBpedia • Linked data • Multilingualism • Question answering

---

E. Cabrio (✉)  
INRIA Sophia Antipolis, Sophia Antipolis, France

EURECOM, Sophia Antipolis, France  
e-mail: [elena.cabrio@inria.fr](mailto:elena.cabrio@inria.fr)

J. Cojan • F. Gandon  
INRIA Sophia Antipolis, Sophia Antipolis, France  
e-mail: [Julien.Cojan@inria.fr](mailto:Julien.Cojan@inria.fr); [Fabien.Gandon@inria.fr](mailto:Fabien.Gandon@inria.fr)

## 1 Introduction

The Semantic Web provides a framework to transform the access to information by adding machine-readable Linked Data and the semantics of their schema to the human-readable textual content, to facilitate automated processing and integration of the vast amount of available information on the web. The Semantic Web is an extension of the classical web, and the data and schemas it adds coexist with the documentary representations that were already available and linked on the web. Moreover, more and more web sites are adding direct access to the data they use to generate their pages and enhance existing services they offer by semantic data. This not only allows interoperability, reusability, and potentially unforeseen applications of opened data, but it leads to a unique situation in which large amounts of information are available, both in textual form for human consumption, as well as in structured form in line with standard shared vocabularies for consumption by machines.

A very important case of such web sites offering strongly tied texts and data is the couple Wikipedia-DBpedia. Collaboratively constructed resources, such as Wikipedia, have grown into central knowledge sources providing a vast amount of updated information accessible on the web, essentially as pages for human consumption. From such corpora, structured information has been extracted and stored into knowledge bases—for example, the DBpedia project Bizer et al. (2009)—that cover a wide range of different domains and connect entities across them. The original DBpedia project has then been mirrored at other sites for the Wikipedia content in other languages than English: we refer to the collection of such DBpedia projects as “language-specific DBpedia chapters.”<sup>1</sup> Language-specific DBpedia chapters are well connected through instance interlinking, extracted from Wikipedia (more details are provided in Sect. 2). An alignment between properties is also carried out by DBpedia contributors as a mapping from the terms used in Wikipedia to a common ontology, enabling the exploitation of information coming from the language-specific DBpedia chapters. At the same time, language-specific DBpedia chapters can contain different information from one language to another, providing more specificity on certain topics or filling information gaps. For instance, when looking for the nationality of Barack Obama on the English DBpedia chapter, we notice that there is no property *nationality* directly linking Obama to the United States. Such information can instead be found in the French DBpedia chapter, the second biggest chapter. Moreover, the knowledge of certain instances and the conceptualization of certain relations can be biased according to different cultures, and this is reflected in the structure and content of such collaboratively constructed resources. No information is provided in English Wikipedia and DBpedia, for instance, for the French musical group “Les Frères Jacques” or for the French writer Jean-Bernard Pouy.

---

<sup>1</sup><http://wiki.dbpedia.org/Internationalization/Chapters>.

Being able to exploit all the amount of multilingual information would bring several advantages to systems that harvest information from Wikipedia and DBpedia—and, more generally, from the Multilingual Semantic Web (Buitelaar et al. 2013)—automatically, both considering (1) the intersection of such resources in different languages to detect contradictions or divergences and (2) the union of such resources, to fill information gaps (cross-fertilization among languages). Also Rinser et al. (2013) highlight the importance of mapping the attributes of the infoboxes across different language versions, to increase the information quality and quantity in Wikipedia.

In the context of Natural Language (NL) Question Answering (QA) over Linked Data, a system which is able to exploit information coming from the multilingual and parallel versions of DBpedia would increase its probability to retrieve a correct answer (i.e., its recall). Given the multilingual scenario, attributes are labeled in different natural languages. The common ontology enables to query the multiple DBpedia chapters with the same vocabulary on the mapped data. Unfortunately, the cross-language mapping process of properties among language-specific DBpedia chapters is currently incomplete, it is time-consuming since it is performed manually, and it may lead to the introduction of redundant terms in the ontology, as it becomes difficult to navigate through the existing vocabulary. Moreover, several problems arise concerning both the variety and ambiguity of properties extracted from Wikipedia Infoboxes (e.g., attribute names are not always sound, often cryptic or abbreviated) and the fact that they are specific to a particular language.

In this chapter, we tackle the following research question:

*How to fill the gaps between language-specific DBpedia chapters for QA?*

Given the complexity of our research question, in this chapter we narrow its scope, answering to the following subquestions:

- (1) *How to benefit from querying language-specific DBpedia data-sets in the current mapping progress?*
- (2) *How to safely extend the property alignments?*
- (3) *How can QA systems benefit from querying language-specific DBpedia chapters?*

In this chapter, we do not make use of general alignment techniques, and we do not enter in the merits of the related discussions.<sup>2</sup> We rather exploit the existing manually created alignments.

In the first part of the chapter, we carry out a comparative analysis of property alignment in language-specific DBpedia chapters, considering English and French DBpedia chapters as a case study and highlighting the current status of the property alignment between them. Moreover, we propose an approach to automatically extend the existing alignments taking advantage of Wikipedia and DBpedia structures.

---

<sup>2</sup>For an overview, see Trojahn et al. (this volume).

In the second part of the chapter, we present an extension of QAKiS (Cabrio et al. 2012), a system for Question Answering over Linked Data that allows to query language-specific DBpedia chapters exploiting the above-mentioned property alignments. Extending QAKiS to query language-specific data-sets goes in the direction of enhancing user consumption of semantic data originally produced for a different culture and language, overcoming language barriers.<sup>3</sup>

The reminder of the chapter is structured as follows. Section 2 provides an analysis of the current status of property alignments in language-specific DBpedia chapters (focusing on the English and French versions), while Sect. 3 proposes an approach to extend the current mappings. Section 4 describes QAKiS extension to query language-specific DBpedia chapters. Section 5 discusses the related work in the literature; conclusions end the chapter.

## 2 DBpedia Property Alignment Current Status

As introduced before, DBpedia (Bizer et al. 2009) is a community effort to extract structured data from Wikipedia and to publish it as Linked Data. At the beginning, it only contained data extracted from the English Wikipedia, while in the most recent period, efforts to integrate data extracted from chapters of languages different from English have arisen (e.g., for German, Spanish, French, and Italian). However, in the current state of affairs, the content is still focused on the English chapter, due to the fact that naming conventions limit the coverage of other chapters and the fact that English is the biggest chapter.

Language-specific DBpedia chapters have been created following the Wikipedia structure (Kontokostas et al. 2012): each chapter contains therefore data extracted from Wikipedia in the corresponding language and so reflects local specificity. Data are published in Resource Description Framework (RDF) and are structured in triples  $\langle \text{subject}, \text{predicate}, \text{object} \rangle$  where the *subject* is an instance corresponding to a Wikipedia page, the *predicate* is a property from the DBpedia ontology or from other vocabularies (e.g., foaf, dublicore, georss), and the *object* is either a literal value or another instance.

Data from different DBpedia chapters are connected by several alignments: (1) *instances* are aligned according to the interlanguage links that are created by Wikipedia editors to relate articles about the same topic in different languages. As shown in Rinser et al. (2013), these correspondences are far from being perfect, but a simple filter applied before data publication in DBpedia significantly improves its quality; (2) *properties* mostly come from template attributes, that is, structured elements that can be included in Wikipedia pages to display structured information, the most common being the infoboxes. The generic template extraction that creates property URIs from their textual names has the inconvenient of generating a

---

<sup>3</sup>Currently a hot topic, see the Multilingual Question Answering over Linked Data challenge (QALD-3 and 4) <http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/index.php?x=home>.

large variety of properties, as well as ambiguous terms. For instance, both properties `propEn:birthDate`<sup>4</sup> and `propEn:dateOfBirth` appear in English DBpedia with the same meaning. On the contrary, the property `propEn:start` is used to indicate both the starting place of a route (e.g., the first station on a railway line) and the date of the beginning of an event. Moreover, as introduced before, the terms used for properties are language dependent.

To overcome these limitations, a common ontology and mappings from template definitions to the ontology vocabulary are being collaboratively edited by the DBpedia community.<sup>5</sup> For instance, the attributes *date of birth* and *birth date* are mapped to the ontology property `dbo:birthDate`<sup>4</sup> in the description of a person, and the attribute *start* is mapped to `dbo:routeStart` when describing a road, to `dbo:startDate` when describing an event. This term normalization effort has the goal to improve the alignment of properties among language-specific DBpedia chapters. It is, however, ongoing work and needs constant maintenance as Wikipedia templates evolve over time. Assistance tools for mapping editions, as well as automated techniques to extend the resulting alignments, are becoming therefore important issues to address.

As a case study to analyze the current state of affairs of property alignment in language-specific DBpedia chapters, we consider the datasets of English and French DBpedia. While the English chapter is the biggest and the most complete, with about 400 million triples<sup>1</sup> and 345 templates mapped, the French chapter is the second chapter in size (~130 million triples and 42 templates mapped). In our analysis, for each object property `prop`, we compare the triples  $\langle \text{subject}, \text{prop}, \text{object} \rangle$  from English and French DBpedia on aligned pairs of instances `subject` and `object`. That is, triples  $\langle \text{subject}_{fr}, \text{prop}, \text{object}_{fr} \rangle$  from French DBpedia are transposed into  $\langle \text{subject}_{en}, \text{prop}, \text{object}_{en} \rangle$ , where `subjecten` and `objecten` are, respectively, instances of English DBpedia related to `subjectfr` and `objectfr` through the relation `owl:sameAs`. These triples are compared with triples  $\langle \text{subject}, \text{prop}, \text{object} \rangle$  from English DBpedia such that `subject` and `object` are also related to French instances with relation `owl:sameAs`.

Figure 1 describes the possible outcomes of such a comparison. In case (a) we have the same value for the property in both the English and the French chapters. For instance, for the subject *Barack Obama* the property `birthPlace` is present in both the English and the French versions, with the same value.

In this case, the French chapter does not bring new information, except a confirmation of values found in the English chapter. In (b) we also have the same property in both English and French chapters but this time with different values. In (c) we have values for the property in the English chapter only: in the example of *Barack Obama*, the property `residence` is present for the English chapter

<sup>4</sup>For simplification, we use here the shorthand `propEn:` for <http://en.dbpedia.org/property/> and `dbo:` for <http://dbpedia.org/ontology/>.

<sup>5</sup>On the wiki <http://mappings.dbpedia.org>.



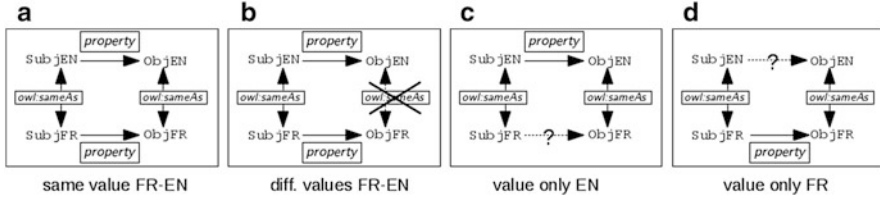


Fig. 1 Outcomes of the comparison between EN and FR chapters

(with the value *White House*), while it is missing in the French version. In (d) we have a value for the property in the French chapter only; again in the example of *Barack Obama*, the property *nationality* is missing for the English DBpedia chapter, while it is present in the French version (with the value *États-Unis*, i.e., *United States*).

There can be two reasons for differing values in case (b) (Fig. 1): (1) a disagreement between the two datasets, produced either by an error in one of them or reflecting a different viewpoint (e.g., for properties of type `owl:functionalProperty`) or (2) the values reported in the two chapters are complementary, often providing a different granularity level (e.g., city vs. country for the birthplace of *Henry Lawson*). The first case can be interestingly exploited to automatically detect inconsistencies among the data which can help the Wikipedia community to improve information quality across language versions. The second one brings additional information on the subject, but it could also help to infer relationships between the values (for instance, that the city where *Henry Lawson* was born is in his country of birth).

The same comparison has been carried out for datatype properties over triples `<subject, prop, val>` with aligned instances `subject`. For every property `prop`, we count (a) how many `subject` have the same values with `prop` in French and English, (b) how many have at least one different value, and how many have only values either (c) in the English or (d) in the French DBpedia.

We observed that the ratio between the number of values that are the same in English and French chapters and the number of values that are different is lower for datatype properties than for object properties. This is true in particular for string literals, as most of them are expressed in their respective chapter language (we did not compare neither instance labels nor abstracts). Nevertheless, we kept these properties in our comparison, as some of them bring information that can be exploited in a different language, for instance, for people's names.

Reflecting the different progression of the mapping task between French and English DBpedia, 217 ontology properties are currently used in French DBpedia, compared to more than 1,000 in English DBpedia.

Table 1 shows some statistics resulting from the comparison between English and French DBpedia. In particular, it shows some of the (object) properties for which French DBpedia presents the highest number of values not present in the English version, that is, the properties to which the French chapter can contribute most.

**Table 1** Statistics resulting from the comparison of the FR and EN DBpedia chapters

	(a) Same value FR-EN (%)	(b) Diff. values FR-EN (%)	(c) Value only EN (%)	(d) Value only FR (%)
dbo:nationality	1,536 (3.8)	437 (1)	11,825 (29.6)	26,074 (65.6)
dbo:birthPlace	14,139 (17.4)	1,965 (2.5)	49,754 (61.3)	15,279 (18.8)
dbo:region	22,178 (44.5)	676 (1.4)	14,397 (29)	12,502 (25.1)
<b>Total object properties</b>	239,321 (14.6)	40,232 (2.6)	1,046,532 (64.3)	305,452 (18.7)
<b>Total datatype properties</b>	104,262 (7.6)	134,995 (9.8)	976,025 (71.2)	155,134 (11.4)
<b>Total</b>	343,583 (11.4)	175,227 (5.8)	2,022,557 (67.3)	460,586 (15.5)

For every property *prop*, column (a) shows how many *subject* have the same values with *prop* in French and English, column (b) shows how many have at least one different value, and columns (c) and (d) show how many have only values either in the English or in the French DBpedia, respectively (values in percentages are reported between brackets)

Moreover, it provides the total number of pairs (subject, property) that (a) have a value in common in English and French chapters, (b) have different values in the two chapters, (c) have only values in English chapter, and (d) have only values in French chapter.

Two intermediate sums are also given for the object properties and for the datatype properties. These sums show overall that the aligned data from the French and English chapters are quite complementary. About 47 % of the data from the French DBpedia expressed in the common ontology cannot be found in English DBpedia (column *d* vs.  $a + b + d$ ), and about 80 % of the data from the English DBpedia expressed in the common ontology cannot be found in French DBpedia (column *c* vs.  $a + b + c$ ). The values provided in Table 1 for the column (d) “*only FR value*” confirm our initial intuition that being able to exploit language-specific DBpedia chapters provides an additional amount of information both specific to a certain culture (for instance, concerning French habits, food, or minor musical groups) and to fill information gaps (for instance, missing links in the English chapters).

### 3 Extending the Existing Alignment

A large portion of the data extracted by the DBpedia community comes from the templates that are used in Wikipedia articles for synthetic descriptions. Templates define a set of attributes to describe a certain kind of entity (e.g., authors, football players, cars, planets). The task of mapping templates consists in matching attributes of a given template to properties of the DBpedia ontology. The DBpedia ontology is relatively large (more than 1,500 properties for DBpedia—version 3.9), and manually finding the appropriate property to be mapped can take some time. However, many attributes are used with the same meaning in several templates, for instance, *name*, *birth date*, or *nationality* in templates for person’s description.

Avoiding the need to repeat these mappings would save DBpedia contributors a lot of time and would speed up the mapping process.

We propose therefore an approach to expand the property mappings to all nonambiguous attributes, that is, attributes that have always been manually mapped to the same ontology property. This results in the extension of the alignments between the properties textually generated from the attributes and the ontology properties. And so, it extends the alignment between language-specific datasets. By nonambiguous attributes, we mean the terms that have not proven to be ambiguous in the existing mappings. The integration of the extended mappings into the mapping data would require human validation in order to check for incorrect alignments. In the following, we evaluate the possible gain obtained from the approach we propose. We use a simple heuristic to select mappings that are likely to be correctly propagated: we select only the attributes that have been mapped consistently to the same ontology property multiple times.

Concerning the mapping frequency of non ambiguous attributes in French DBpedia to the DBpedia ontology properties, 47 attributes are mapped at least twice, 18 attributes are mapped at least three times (i.e., *lieu de décès* → `dbo:deathPlace`), and only one is mapped at least ten times (i.e. *nom* → `foaf:name`). Since we assume that the mapping frequency is a good indicator of the correctness of the mapping, in the rest of the section, we will consider only the mappings that were mapped at least twice (i.e., frequency  $\geq 2$ ). Moreover, we carry out a manual validation of the 47 mappings appearing more than twice, to check if they are correct according to the attribute names. The results of such evaluation confirm that in 83 % of the cases (i.e., 35 mappings), the mappings are correct. The validity of the remaining ones can be biased by the context in which they appear, since the attribute terms are either vague or polysemous (i.e., could have different meanings). For instance, mapping the attribute *division* to `dbo:locatedInArea` seems correct for geographic places, but *division* could be used to indicate also a football league or an organization department, and in those cases the mapping is incorrect.

Table 2 provides for each mapping a comparison between the number of instances that have a value for the generic property (build from the attribute occurrence) and the number of instances that have a value for the mapped ontology property. For instance, the property `propFr:lieuDeDécès` is present for more than 25,000 instances (column *values for p*, Table 2) and `dbo:deathPlace` for more than 17,000 (column *values for po*). Note that *lieu de décès* is not the only attribute to be mapped to `dbo:deathPlace` (i.e., also *lieu décès*, *décès*, and other variants). The column *values for both* indicates how often the mapping *lieu de décès* to `dbo:deathPlace` is actually applied, and it gives the number of instances that have values for both the generic and the ontology properties (i.e., 13,314). The potential gain of this mapping extension is given by the number of instances that have a value for the generic property but no values for the ontology property, that is,  $25,477 - 13,314 = 12,163$  additional values for `dbo:deathPlace`. Over the 47 mappings that can be extended, the potential gain is  $1,326,200 - 543,824 = 782,376$ , corresponding to an increase of about 59 %.

**Table 2** Comparison of values between generic and ontology properties for the extended mappings in French DBpedia

Generic prop. (p) propFr:	Ontology prop. (po) dbo:	Values for p	Val. for p in po range (%)	Values for po	Values for both	Same values (%)
lieuDeDécès	deathPlace	25,477	14,615 (57.3)	17,190	13,314	7,579 (56.9)
région	region	87,917	79,853 (90)	51,713	46,077	45,993 (99)
nationalité	nationality	44,345	10,071 (22.7)	46,985	34,884	8,887 (25.4)
lieuDeNaiss.	birthPlace	66,262	37,326 (56.3)	49,430	41,716	24,569 (58.8)
<b>Total object prop</b>		645,719	391,044 (60.5)	482,444	28,4201	20,9692 (73.7)
<b>Total datatype prop</b>		680,481	111,876 (16.4)	517,368	259,623	59,047 (22.7)
<b>Total</b>		1,326,200	502,920 (37.9)	999,812	543,824	268,739 (49.4)

Column *values for p* reports the number of instances that have values for the generic properties; column *val. for p in po range* reports the instances for which the generic property values are coherent with the ontology property signature; column *values for po* reports the instances that have values for the mapped ontology property. Column *values for both* reports the number of instances that have values for both the generic and the ontology properties; column *same value* reports those for which the generic property and the ontology property have the same value

Column *same values* gives the number of instances for which the generic property and the ontology property have the same value. However, the comparison with the number of co-occurrence of the two properties is not fair, as the extractor that generates the values for the ontology property is guided by the property signature (in the example of `dbo:deathPlace`, the expected value is an instance), whereas the generic property is more subject to noise and may generate another output from the same attribute value (for instance, a number if the attribute value begins with a street number). So for this comparison, we narrow our scope to the instances for which the generic property values are coherent with the ontology property signature (column *values for p in po range*). Out of the 25,477 instances that have a value for `propFr:lieuDeDécès`, only 14,615 have an object value. However, every time there is an object value for `propFr:lieuDeDécès` and a value for `dbo:deathPlace`, these are the same. In a symmetric way, we calculated the mapping frequency of nonambiguous attributes in English DBpedia to ontology properties. As expected, many more attributes are mapped more frequently than in the French chapter (i.e., 689 attributes are mapped at least twice, 296 attributes mapped at least five times, and 160 mapped at least ten times, e.g., *twin* to `dbo:twinCity` or *successor* to `dbo:successor`).

To evaluate the quality of the data obtained applying the above-presented approach to extend the mapping among language-specific DBpedia chapters, we compare the values obtained from the mapping extension for the French chapter, with the values obtained for the English chapter as previously done in Sect. 2 for the existing alignments. Table 3 summarizes the results obtained from such comparison. More specifically, it provides the number of values that were added through this process (column *new values w.r.t. DBpedia En and Fr*) with respect to the values already available through ontology properties in English and French DBpedia. For instance, the mapping extension (*lieu de naissance* to `dbo:birthPlace`)

**Table 3** Comparison between values obtained with the mappings extension in French and English DBpedia

Generic property <i>propFr</i> :	Ontology property <i>dbo</i> :	New values w.r.t. DBpedia En and Fr	Same values (%)	Diff. values
<i>lieuDeDécès</i>	<i>deathPlace</i>	4,393	4,016 (89.3)	479
<i>région</i>	<i>region</i>	16,491	18,496 (82.5)	3,906
<i>nationalité</i>	<i>nationality</i>	358	870 (81.3)	200
<i>lieuDeNaissance</i>	<i>birthPlace</i>	6,934	7,016 (89)	862
<b>Total object prop</b>		85,951	733,06 (82.7)	15,250
<b>Total datatype prop</b>		16,155	45,177 (90)	5,001
<b>Total</b>		102,106	118,483 (85.4)	20,251

Column *new values w.r.t. DBpedia En and Fr* provides the number of values that were added through the mapping extension process w.r.t. the values already available through ontology properties in English and French DBpedia

considered earlier generates 6,934 new values. Among the values that were already present in the English chapter, 7,016 are the same and 862 differ (89 % identical). We can notice that this is about the same ratio as for the comparison between values for the same ontology property in Sect. 2, that is, 14,139 identical values (column a, Table 1) and 1,965 different (columns a+b, Table 1), that is, 87 % identical. We can consider it as a positive result, as it suggests that most of the differences in the values are generated by differences between the two chapters of DBpedia, rather than from mapping mistakes.

Concerning the 47 mappings described in Sect. 3, we have 118,483 identical values and 20,251 different values (respectively, columns *same values* and *different values* in Table 3). If we consider object properties and datatype properties separately, we obtain now a better correlation between values of English and French chapters for datatype properties (90 % instances with same values) than for object properties (82 %). This may be explained by the fact that many datatypes are not specified for generic properties (e.g., for strings), so we selected the values that fit the range of the property as specified in the ontology and we removed values that generated noise in the comparison described in Sect. 2.

## 4 QA Experimental Setting

To benefit from the amount of information coming from the aligned language-specific datasets described before, we extended QAKiS, our system for open domain Question Answering over Linked Data (Cabrio et al. 2012), to query language-specific DBpedia chapters (Sect. 4.1). To enhance users interactions with the Web of Data, query interfaces providing a flexible mapping between natural language expressions and concepts and relations in structured knowledge bases are becoming particularly relevant. More specifically, QAKiS allows end users to submit a query to an RDF triple store in English and obtain the answer in the same language, hiding

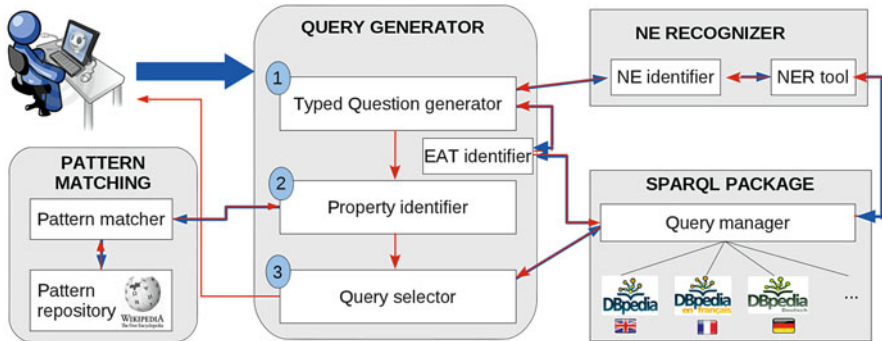


Fig. 2 QAKiS workflow

the complexity of the nonintuitive formal query languages involved in the resolution process. At the same time, the expressiveness of these standards is exploited to scale to the huge amounts of available semantic data.

We evaluate QAKiS extension to query English, French, and German DBpedia chapters with two sets of experiments, described in Sects. 4.2 and 4.3.

#### 4.1 QA System Description: QAKiS

Question Answering wiKiFramework-based System (QAKiS)<sup>6</sup> (Cabrio et al. 2012) addresses the task of QA over structured knowledge bases (e.g., DBpedia), where the relevant information is expressed also in unstructured forms (e.g., Wikipedia pages). It implements a relation-based matching for question interpretation, to convert the user question into a query language (e.g., SPARQL). More specifically, it makes use of relational patterns—automatically extracted from Wikipedia and collected in the WikiFramework repository (Mahendra et al. 2011)—that capture different ways to express a certain relation in a given language.<sup>7</sup>

QAKiS is composed of four main modules (Fig. 2):

- The *query generator* takes the user question as input, generates the typed questions, and then generates the SPARQL queries from the retrieved pattern.
- The *pattern matcher* takes as input a typed question and retrieves the patterns (among those in the repository) matching it with the highest similarity.
- The *SPARQL package* handles the queries to DBpedia.
- A *named entity (NE) recognizer*.

<sup>6</sup>A demo is available at <http://qakis.org/qakis2/>.

<sup>7</sup>Gerber et al. (this volume) describe another framework (i.e., BOA) to address the challenge of extracting structured data as RDF from unstructured data.

The actual version of QAKiS targets questions containing an NE related to the answer through one property of the ontology, as *Which river does the Brooklyn Bridge cross?* or *In which country does the Nile starts?* Such questions match a single pattern (i.e., one relation).

Before running the *pattern matcher* component, the target of the question is identified using the Stanford Core NLP NE Recognizer,<sup>8</sup> together with a set of strategies based on the comparison with the labels of the instances in the DBpedia ontology. Then a *typed question* is generated by replacing the question keywords (e.g., who, where) and the NE by their types and supertypes. A Word Overlap algorithm is then applied to match such typed questions with the patterns for each relation. A similarity score is provided for each match: the highest represents the most likely relation. A set of patterns is retrieved by the pattern matcher component for each typed question and sorted by decreasing matching score. For each of them, a set of SPARQL queries is generated and then sent to the SPARQL endpoint for answer retrieval.

#### 4.1.1 QAKiS Extension to Query Language Specific DBpedia Chapters

To allow QAKiS to query the ontology properties of language-specific DBpedia chapters, we modified QAKiS architecture at the *SPARQL package* level. The typed question generation and the pattern matching steps work as before, but now, instead of sending the query to English DBpedia only, the *query manager* reformulates the query and sends it to multiple DBpedia chapters. As only the English chapter contains labels in English, this change has no impact on the NE recognition. The main difference lies in the query selection step. As before, patterns are considered in order of decreasing matching score, the generated query is then evaluated and if no results are found the next pattern is considered, and so on. However, as queries are now evaluated on several DBpedia chapters, it is more likely to get results, terminating query selection with a higher matching score. Currently, the results of an SPARQL query are aggregated by the set union. Other strategies could be considered, such as using a voting mechanism to select the most frequent answer or enforcing a priority according to data provenance (e.g., English chapter could be considered as more reliable for questions related to English culture). In the current version, QAKiS allows to query English, French, and German DBpedia chapters.

## 4.2 Evaluation on QALD-2 Dataset

As a first step of our experiments, we evaluate if the integration of the French and German DBpedia datasets has an impact on QAKiS performances on the

---

<sup>8</sup><http://nlp.stanford.edu/software/CRF-NER.shtml>.

standard benchmark of the QALD-2 challenge<sup>3</sup> (DBpedia track). It is provided by QALD organizers to compare different approaches and systems that mediate between a user, expressing his or her information need in natural language, and semantic data. Since in the actual version of the system it targets only questions containing an NE related to the answer through one property of the ontology (e.g., *In which military conflicts did Lawrence of Arabia participate?*), we extracted from the complete benchmark the questions corresponding to such criteria. Out of 100 questions available for testing, the questions containing an NE related to the answer through one property of the ontology amount to 32, which we used in our experiment. The discarded questions require either some forms of reasoning (e.g., counting or ordering) on data, aggregation (from datasets different from DBpedia), involve n-ary relations, or they are Boolean questions. We run both QAKiS<sub>EN</sub> (i.e., the system taking part into the challenge) and QAKiS<sub>EN+FR</sub> and QAKiS<sub>EN+DE</sub> (the versions enriched with the French and German DBpedia chapters, respectively) on the reduced set of questions.

Since the answer to QALD-2 questions can be retrieved from the English DBpedia, we do not expect multilingual QAKiS to improve its performances. On the contrary, we want to verify that QAKiS performances do not decrease (due to the choice of the wrong relation triggered by a different pattern that finds an answer in language-specific DBpedia chapters). QAKiS<sub>EN</sub> correctly answers to 15/32 questions and partially correctly to 4/32 questions (e.g., in *Give me all companies in Munich*, the list provided by QAKiS using `foundationPlace` as relation and *Munich* as subject is only partially overlapping with the one proposed by the organizers). The extended QAKiS often selects patterns that are different with respect to the one selected by the original system, but except in one case the identified target relation is the same, meaning that performances are not worsened when querying several language-specific DBpedia chapters.

### 4.3 *Separate Evaluations on French and German DBpedia Chapters*

As introduced before, the questions created for QALD-2 challenge are thought to find an answer in the English DBpedia, so they cannot be used to evaluate the contribution resulting from the extension of property alignments to language-specific DBpedia chapters. Since we are not aware of any standard list of questions whose answers can be found in French and German DBpedia chapters only, we created our reference set to evaluate the extension in QAKiS<sub>EN+FR</sub> and QAKiS<sub>EN+DE</sub>'s coverage performing the following steps:

1. We take the sample of 32 questions from QALD-2.
2. We extract the list of triples present in French (and German) DBpedia only (as described in Sect. 2).



3. In each question, we substitute the named entity with another entity for which the asked relation can be found in the French (or German) chapter only.

For instance, for QALD-2 question *How tall is Michael Jordan?*, we substitute the Named Entity *Michael Jordan* with the entity *Margaret Simpson*, for which we know that the relation `height` is not present in English DBpedia, but it is linked in the French chapter. As a result, we obtain the question *How tall is Margaret Simpson?* that we submit to QAKiS<sub>EN+FR</sub>. Following the same procedure for German, in *Who developed Skype?* we substituted the NE *Skype* with the entity *IronPython*, obtaining the question *Who developed IronPython?*<sup>9</sup> For some properties (i.e., `Governor`, `Battle`, `FoundationPlace`, `Mission`, and `RestingPlace`), no additional links are provided by language-specific DBpedia chapters, so we discarded related questions.

QAKiS precision on the new set of questions over French and German DBpedia is in line with QAKiS<sub>EN</sub> on English DBpedia (~50%) (i.e., out of 27 questions, QAKiS<sub>EN+FR</sub> correctly answers to 14 questions and partially correctly to 1 question). To double-check, we run the same set of questions on QAKiS<sub>EN</sub> (which relies on the English chapter only), and in no cases it was able to detect the correct answer, as expected. This second evaluation did not have the goal to show improved performances of the extended QAKiS with respect to its precision, but to show that the integration of language-specific DBpedia chapters in the system is easily achievable and that the expected improvements on its coverage are really promising and worth exploring (see Table 1).

## 5 Related Work

In this chapter, we have exploited existing alignments over DBpedia data to compare and aggregate data from different Wikipedia chapters. The instance alignments are manually edited by the Wikipedia community (as interlanguage links) and the property alignments by the DBpedia community. The field of ontology alignment tackles questions about automated or partially automated alignment techniques. Rahm and Bernstein (2001) and Shvaiko and Euzenat (2013) present surveys on the topic.

Several works address the more specific topic of data integration from Wikipedia chapters directly from the article content. Rinser et al. (2013) provide an overview of instance-based template attributes matching approaches over language-specific Wikipedia chapters. They also present their own very thorough approach. First, several criteria are taken into account to improve the instance matching resulting from the interlanguage links (i.e., based on this instance alignment, a template

---

<sup>9</sup>The obtained set of transformed questions is available at <http://qakis.org/qakis2/>.

alignment is computed according to their use in matched instances). Then, attributes of aligned templates are matched according to the instances and values they relate.

To predict the matching probability of pairs of infobox attribute instances across different language versions, Adar et al. (2009) employ self-supervised machine learning with a logistic regression classifier using a broad range of features (e.g., n-gram/word overlap of attribute keys and values, wiki link overlap). Bouma et al. (2009) perform a matching of infobox attribute based on instance data. Bouma (2010) describes a system for linking the thesaurus of the Netherlands Institute for Sound and Vision to EnglishWordNet and DBpedia, using EuroWordNet and Dutch Wikipedia as intermediaries for the two alignments. Tacchini et al. (2009) provide several strategies for merging data extracted from different Wikipedia chapters. They present a software framework for fusing RDF datasets based on different conflict resolution strategies, and they apply it to fuse infobox data that is extracted from multilingual editions of Wikipedia.

Apro시오 et al. (2013) define a methodology to increase DBpedia coverage in different languages. Information is bootstrapped through cross-language links, starting from the available mappings in some pivot languages and then extending the existing DBpedia datasets comparing the classifications in different languages. When such classification is missing, supervised classifiers are trained on the original DBpedia (relying on the Distant Supervision paradigm).

A survey on the field of Question Answering is provided by Lopez et al. (2011), with a focus on ontology-based QA. Moreover, they examine the potential of open user-friendly interfaces for the SW to support end users in reusing and querying SW content. State-of-the-art QA systems over Linked Data generally address the issue of question interpretation mapping a natural language question to a triple-based representation. For instance, Freya (Damljanovic et al. 2012) uses syntactic parsing in combination with ontology-based lookup for question interpretation, partly relying on the user's help in selecting the entity that is most appropriate as match for some natural language expressions. One of the problems of that approach is that often end users are unable to help, in case they are not informed about the modeling and vocabulary of the data. PowerAqua (Lopez et al. 2009) accepts user queries expressed in NL and retrieves answers from multiple semantic sources on the SW. It follows a pipeline architecture, according to which the question is (1) transformed by the linguistic component into a triple-based intermediate format and (2) passed to a set of components to identify potentially suitable semantic entities in various ontologies, and then (3) the various interpretations produced in different ontologies are merged and ranked for answer retrieval. PowerAqua's main limitation is in its linguistic coverage.

Pythia (Unger and Cimiano 2011) relies on a deep linguistic analysis to compositionally construct meaning representations using a vocabulary aligned to the vocabulary of a given ontology. Pythia's major drawback is that it requires a lexicon, which has to be manually created. More recently, an approach based on Pythia (Unger and Cimiano 2011) but more similar to the one adopted in QAKiS is presented (Unger et al. 2012). It relies on a linguistic parse of the question to produce an SPARQL template that directly mirrors the internal structure of the question

(i.e., SPARQL templates with slots to be filled with URIs). This template is then instantiated using statistical entity identification and predicate detection.

## 6 Conclusions and Future Work

In the first part of this chapter, we have proposed an in-depth comparative analysis of language-specific DBpedia chapters, focusing in particular on the French and the English DBpedia chapters, proving that most of their content is complementary: each chapter brings a significant amount of data that cannot be found in the other chapter (about half of the data from the French DBpedia and 80 % of the data from the English DBpedia). To perform this comparison, we have first considered the existing alignments and compared the two chapters to highlight their differences. Then, we have proposed an approach to extend the existing property alignments to all the occurrences of nonambiguous attributes (i.e., attributes that humans have always mapped to the same ontology properties).

Since the DBpedia ontology is continuously evolving, maintaining its consistency is a complex task that needs continual updates. Some studies have been carried out to evaluate the quality of the DBpedia ontology: being able to automatically compare the values of several chapters, as we showed in our work, could provide interesting indicators of errors or vandalism in one chapter and detect discrepancies among vocabulary used among chapters or even among topics of the same chapter.

In the second part of this chapter, we have considered Question Answering over Linked Data scenario. To show the interesting potential for NLP applications resulting from the property alignments in language-specific DBpedia chapters, we have extended the QAKiS so that it can query the ontology properties of the French and German DBpedia chapters. We show that this integration extends the system coverage (i.e., the recall), without having a negative impact on its precision.

We plan to extend the presented work in a number of directions. First, we would like to improve the mapping extension approach by taking into account instance types to disambiguate attributes. We also plan to use alignment tools (e.g., Silk<sup>10</sup>) to suggest additional property alignments based on the similarity of their use in their respective chapters (e.g., considering the number of equivalent pairs that two properties have in common). Moreover, since the pieces of information obtained by querying distributed SPARQL endpoints may provide different results for the same query, leading to an inconsistent set of information about the same topic, we are investigating the problem of reconciling information obtained by distributed SPARQL endpoints. In particular, we plan to address this problem by combining the AI non-monotonic reasoning framework called argumentation theory to reason over inconsistent sets of information and provide nevertheless a unique and motivated

---

<sup>10</sup><http://lod2.eu/Project/Silk.html>.

answer to the user. We are currently working at the implementation and evaluation of such a framework in QAKiS (Cabrio et al. 2013).

## References

- Adar, E., Skinner, M., & Weld, D. S. (2009). Information arbitrage across multi-lingual Wikipedia. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining (WSDM)* (pp. 94–103). New York: ACM.
- Apro시오, A. P., Giuliano, C., & Lavelli, A. (2013). Towards an automatic creation of localized versions of DBpedia. In *Proceedings of the International Semantic Web Conference (ISWC)* (pp. 494–509).
- Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., et al. (2009). DBpedia: A crystallization point for the Web of Data. *Web Semantics*, 7(3), 154–165.
- Bouma, G. (2010). Cross-lingual ontology alignment using EuroWordNet and Wikipedia. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*.
- Bouma, G., Duarte, S., & Islam, Z. (2009). Cross-lingual alignment and completion of Wikipedia templates. In *Proceedings of the Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies (CLIAWS3)* (pp. 21–29). Stroudsburg, PA, USA: Association for Computational Linguistics.
- Buitelaar, P., Choi, K.-S., Cimiano, P., & Hovy, E. H. (2013). The multilingual semantic Web (Dagstuhl Seminar 12362). *Dagstuhl Reports*, 2(9), 15–94.
- Cabrio, E., Cojan, J., Palmero, A., Magnini, B., Lavelli, A., & Gandon, F. (2012). QAKiS: An open domain QA system based on relational patterns. In *Proceedings of the International Semantic Web Conference (ISWC Demos)*.
- Cabrio, E., Cojan, J., Villata, S., & Gandon, F. (2013). Hunting for Inconsistencies in Multilingual DBpedia with QAKiS. In *Proceedings of the International Semantic Web Conference (ISWC Posters & Demos)* (pp. 69–72).
- Damljanovic, D., Agatonovic, M., & Cunningham, H. (2012). FREyA: An interactive way of querying linked data using natural language. In *Proceedings of the 8th International Conference on the Semantic Web (ESWC)* (pp. 125–138). Berlin: Springer.
- Kontokostas, D., Bratsas, C., Auer, S., Hellmann, S., Antoniou, I., & Metakides, G. (2012). Internationalization of linked data: The case of the Greek DBpedia edition. *Web Semantics: Science, Services and Agents on the World Wide Web*, 15, 51–61.
- Lopez, V., Uren, V. S., Sabou, M., & Motta, E. (2009). Cross ontology query answering on the semantic Web: An initial evaluation. In *Proceedings of the 5th International Conference on Knowledge Capture (K-CAP)* (pp. 17–24).
- Lopez, V., Uren, V. S., Sabou, M., & Motta, E. (2011). Is question answering fit for the semantic web? A survey. *Semantic Web*, 2(2), 125–155.
- Mahendra, R., Wanzare, L., Bernardi, R., Lavelli, A., & Magnini, B. (2011). Acquiring relational patterns from Wikipedia: A case study. In *Proceedings of the 5th Language and Technology Conference (LTC)*.
- Rahm, E., & Bernstein, P. A. (2001). A survey of approaches to automatic schema matching. *The VLDB Journal*, 10(4), 334–350.
- Rinser, D., Lange, D., & Naumann, F. (2013). Cross-lingual entity matching and infobox alignment in Wikipedia. *Information Systems*, 38(6), 887–907.
- Shvaiko, P., & Euzenat, J. (2013). Ontology matching: State of the art and future challenges. *IEEE Transactions on Knowledge and Data Engineering*, 25(1), 158–176.
- Tacchini, E., Schultz, A., & Bizer, C. (2009). Experiments with Wikipedia cross-language data fusion. In *CEUR Workshop Proceedings ISSN 1613-0073* (Vol. 449).

- Unger, C., Bühmann, L., Lehmann, J., Ngonga Ngomo, A. -C., Gerber, D., & Cimiano, P. (2012). Template-based question answering over RDF Data. In *Proceedings of the 21st International Conference on World Wide Web (WWW)* (pp. 639–648). New York, NY, USA: ACM.
- Unger, C., & Cimiano, P. (2011). Pythia: Compositional meaning construction for ontology-based Question Answering on the Semantic Web. In *Proceedings of the 16th International Conference on Applications of Natural Language to Information Systems (NLDB)* (pp. 153–160).

# Multilingual Extraction Ontologies

David W. Embley, Stephen W. Liddle, Deryle W. Lonsdale, Byung-Joo Shin,  
and Yuri Tijerino

**Abstract** The growth of multilingual web content and increasing internationalization portends the need for cross-language query processing. We offer ML-OntoES (a MultiLingual **O**ntology-based **E**xtraction **S**ystem) as a solution for narrow-domain/data-rich applications. Based on language-independent extraction ontologies (Embley et al., *Conceptual modeling foundations for a web of knowledge*. In: Embley D, Thalheim B (eds) *Handbook of conceptual modeling: theory, practice, and research challenges*. Springer, Heidelberg, Germany, pp 477–516, 2011a), ML-OntoES enables semantic search over domain-specific, semistructured information. Key ideas of ML-OntoES include: (1) monolingual semantic indexing and query interpretation with extraction ontologies and (2) conceptual-level cross-language translation. A prototype implementation, along with experimental work showing good extraction accuracy in multiple languages, demonstrates the viability of the ML-OntoES approach of using multilingual extraction ontologies for cross-language query processing.

**Key Words** Conceptual-level cross-language information transfer • Cross-language query processing • Extraction ontologies • Monolingual query interpretation • Monolingual semantic indexing

---

D.W. Embley (✉) • S.W. Liddle • D.W. Lonsdale  
Brigham Young University, Provo, UT, USA  
e-mail: [embley@cs.byu.edu](mailto:embley@cs.byu.edu); [liddle@byu.edu](mailto:liddle@byu.edu); [lonz@byu.edu](mailto:lonz@byu.edu)

B.-J. Shin  
Kyungnam University, Kyungnam, Korea  
e-mail: [shinbyungjoo@gmail.com](mailto:shinbyungjoo@gmail.com)

Y. Tijerino  
Kwansei Gakuin University, Kobe-Sanda, Japan  
e-mail: [ontologist@gmail.com](mailto:ontologist@gmail.com)

## 1 Introduction

An ideal cross-language query system would allow users to pose queries and receive answers in their own language when executing queries against foreign-language source documents. A user  $U$ , for example, who speaks only English, may wish to enquire about nearby restaurants while visiting Japan. Using an iPhone,  $U$  may wish to pose a query to find a “BBQ restaurant with typical prices < \$40.” Figure 1 shows an interface with the query in a type-in text field, the English version of the answers retrieved, and a “see further information button” to tap on to obtain more details such as hours of operation, payment method, and rating. Figure 2 gives actual answers retrieved from the web for this sample query (all in Japanese, of course), and this is the challenge—to query the Japanese in Fig. 2 with the English in Fig. 1.



**Fig. 1** English query over Japanese data with results in English

店名	住所	ジャンル	予算
新肉屋	梅田1-10-19	焼肉	2000
肉屋	梅田1-11-29	焼肉	3000
美味肉	梅田2-30-22	焼肉	1500
焼肉屋	梅田3-19-28	焼肉	3000
焼き焼き	梅田2-18-26	焼肉	1000

Fig. 2 Results extracted from Japanese web pages

Queries like the English-Japanese BBQ restaurant query call for CLIR (Cross-Language Information Retrieval) (Olive et al. 2011; Peters et al. 2012). Interest in CLIR and related technologies is growing, and international initiatives are helping mature the field.<sup>1</sup> A typical approach to CLIR consists of query translation followed by monolingual retrieval and retranslation of results. Our approach to CLIR, which we describe in detail in Sect. 2, differs substantially: rather than translate a query at the language level, we first interpret it with respect to a conceptualization with both query and conceptualization in the same language; we then translate the query to an identical conceptualization in the target language, and having previously semantically annotated target documents with respect to the target-language conceptualization, we then retrieve results and reverse the conceptual translation to return final results in the language of the query.

The approach we take is not entirely unprecedented; several other types of systems use an “interlingua” to mediate processing of content between two or more languages. Since the days of symbolic pivot-based machine translation (Mahesh 1996), ontologies of various sorts have served in crosslinguistic applications including information extraction (Declerck et al. 2010; Aggarwal et al. 2013). Recently, ontology localization (Tijerino 2010) has become viable in boosting lexical content for translation. Some support translation via mappings between language-specific ontologies (Fu et al. 2012). Others, with the advent of statistical methods in natural language processing, use hybrid approaches in translating extraction-ontology content (Montiel-Ponsoda et al. 2011).

Because our approach is symbolic and ontology based and implements first-order (but not higher-order) logic for inference, the concerns raised by Hirst (this volume) could be relevant. We note, however, that the technologies for our system at present originate from the conceptual-modeling and data-extraction communities rather than from natural language processing and computational linguistics, though we foresee being able to orient our work more toward the nexus of all of these areas. In particular, our ontologies do not model the lexicon; they model conceptual relations, with relevant grounding in lexical entries, and the assertions they represent are more “data”-like than “information”-like and thus do not suffer as severely from the issues Hirst raises (this volume). In addition, since creating a domain ontology

<sup>1</sup>See, for example, <http://www.clef-initiative.eu>.



is within the purview of end users, they can either develop a writer-centered view of the data (i.e., more directly modeling the document type) or a reader-centered view (i.e., more oriented to which concepts are of most use to them). To avoid the grand pitfalls in Hirst's warning (this volume), we concentrate on data-rich, narrow-domain applications known a priori and consider our knowledge sources useful, if imperfect, artifacts. Furthermore, we adopt a multifaceted engineering approach for cross-language mappings, and while recognizing the equivalency problem, we allow for various types of correspondence beyond one-to-one mappings (Embley et al. 2011c).

What distinguishes our approach is the narrow, domain-specific, user-definable nature of our ontologies and their construction, as well as the role of these ontologies at the center of a larger infrastructure (Embley et al. 2011c). Our ontologies tend to be less elaborate than others and hence less rich in the types of context required for successful treatment by statistical translation methods. Our work is situated in the space of linguistically grounded, end-user-developed ontologies that incorporate various lexical resources and mappings at various levels of conceptualization.

These semantic conceptualization requirements limit our approach to applications that are easily conceptualizable—those that are data-rich and narrow in scope. Although limited, the applications are significant and practically important covering areas such as service finding like the restaurant example illustrated in Figs. 1 and 2, retail purchasing while shopping abroad, information seeking while traveling and sightseeing, and multicultural topical research such as family history where ancestors have immigrated to a country with a different language.

We call our cross-language query engine ML-OntoES (**M**ulti**L**ingual **O**ntology **E**xtraction **S**ystem) and describe its architecture in Sect. 2. Like search engines, ML-OntoES assumes the existence of an indexed document collection. Indexes for ML-OntoES, however, are not just for keywords but are also for recognized semantic concepts. Extraction ontologies (Embley et al. 2011a), which we describe in Sect. 2.1, allow ML-OntoES to semantically index a document collection with respect to an ontological conceptualization. Extraction ontologies also allow ML-OntoES to interpret queries with respect to an ontological conceptualization, as we describe in Sect. 2.2. ML-OntoES matches conceptualized queries with the conceptualized semantic index to retrieve results. When the query language differs from the document-collection language, ML-OntoES invokes a conceptual-level translation as we explain in Sect. 2.3. In order for ML-OntoES to work well, semantic recognition accuracy must be high and extraction-ontology construction costs must be low; we address these issues in Sect. 3. In Sect. 4, we conclude by summarizing the principles and practicalities required to make ML-OntoES work successfully.

## 2 ML-OntoES Architecture

Figure 3 sketches the architecture of ML-OntoES by giving a retail-sales example in which ML-OntoES processes a French query against a collection of Korean car advertisements. Before query processing begins, ML-OntoES applies its Korean

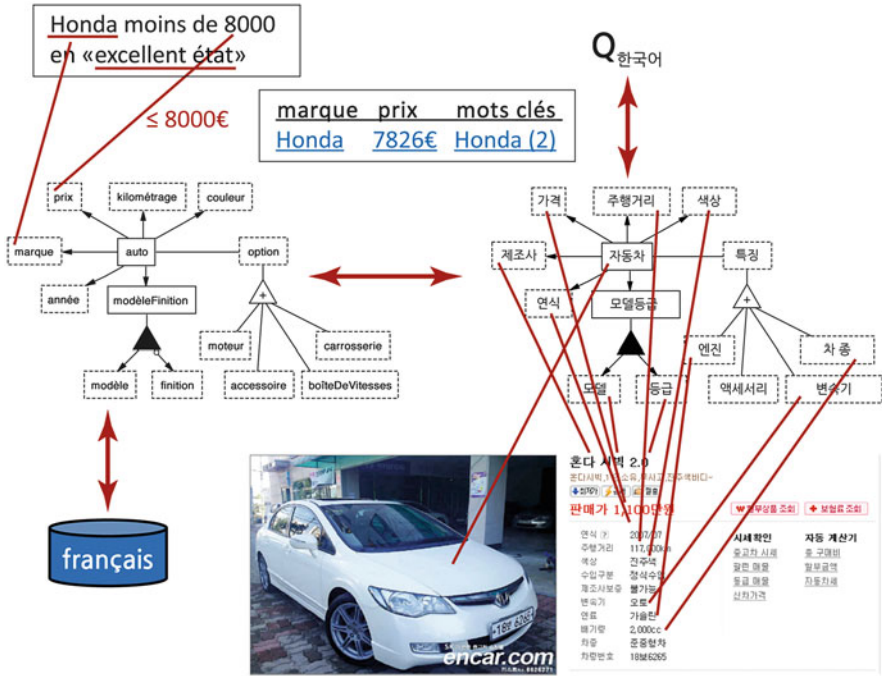


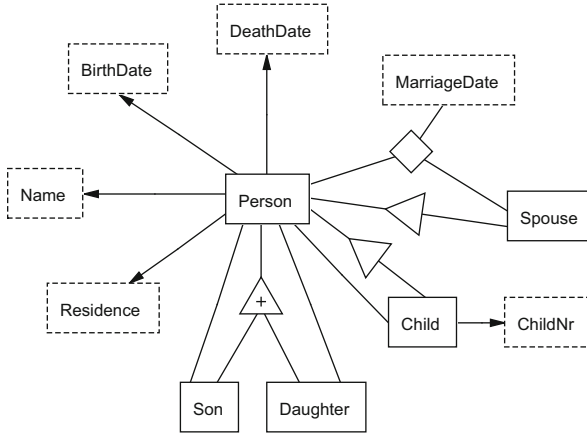
Fig. 3 Cross-language query processing

extraction ontology to Korean source pages to create a semantic index. Once semantic indexes have been built, query processing can begin: as Fig. 3 illustrates, ML-OntoES (1) applies a French car-ad extraction ontology to the query to recognize and conceptualize the query’s semantic constraints and to remove semantic-constraint words from query, leaving and thus identifying the keywords; (2) maps the French conceptualization and keywords to the Korean conceptualization and keywords (note that the conceptualizations are structurally one to one, allowing for identical select-project-join processing); (3) matches the Korean conceptualization and keywords with the previously constructed semantic and keyword indexes; (4) maps the resulting Korean conceptualizations and keywords back into French; and (5) displays the results. As Fig. 3 shows, query processing of a Korean query  $Q_{\text{제조사}}$  over the French repository, *français*, is symmetrical.

### 2.1 ML-OntoES Extraction Ontologies

An *extraction ontology* (see Figs. 4 and 5) is a 5-tuple  $(O, R, C, I, L)$ :

$O$  : Object sets—one-place predicates whose instance values are either all *lexical*, denoted by named dashed-border rectangles in Fig. 4, or all *nonlexical*, denoted



**Fig. 4** Ontological conceptualization for assertion extraction

by solid-border rectangles (e.g., *BirthDate* is lexical with values such as “June 7, 1949” and *Person* is nonlexical with object-identifier values)

*R* : Relationship sets— $n$ -place predicates,  $n \geq 2$ , represented by lines connecting object-set rectangles (e.g., *Person–Name* in Fig. 4) and also by black-triangle aggregation symbols connecting holonyms (e.g., *modèleFinition* in Fig. 3) to meronyms (e.g., *modèle* and *finition*)

*C* : Constraints—closed formulas, as implied by the notation (e.g.,  $\forall x(Person(x) \Rightarrow \exists!y(Person-BirthDate(x, y)))$ )—one of the many functional constraints denoted by the arrowhead on the range side of the *Person–BirthDate* relationship set;  $\forall x(Child(x) \Rightarrow Person(x))$ —a hypernym/hyponym constraint denoted by the triangle, which may optionally also specify mutual exclusion among its hyponym sets by a “+” symbol (e.g., mutual exclusion of *Son* and *Daughter* in Fig. 4) or specify that the hypernym set is a union of its hyponym sets (“U”) or both (“ $\cup$ ”) to form a partition among its hyponyms)

*I* : Inference rules—logic rules specified over predicates (e.g.,  $Person-Gender(x, 'Female') :- Daughter(x)$ )

*L* : Linguistic groundings—text recognizers for populating object and relationship sets and collections of interrelated object and relationship sets (e.g., recognizers for *Name* and *BirthDate* in Fig. 5)

The conceptual foundation for an extraction ontology is a restricted fragment of first-order logic, but its most distinguishing feature is its linguistic grounding,<sup>2</sup> which turns an ontological specification into an extraction ontology. Each object set has a *data frame* (Embley 1980), which is an abstract data type augmented with linguistic recognizers that specify textual patterns for recognizing instance

<sup>2</sup>Similar to the linguistic grounding discussed in Buitelaar et al. (2009), but different in its details.

```

Name
  external representation: \b{FirstName}\s{LastName}\b
  external representation: \b{FirstName}\s[A-Z]\w+\b
  ...
BirthDate
  external representation: \b1[6-9]\d\d\b
  left context: b\s
  right context: [.]
  context keywords: \bborn\b(\sin\b)?|...
  ...
  input method: DateStringToJulianDate
  output method: JulianDateToDateString
  operator methods:
    LessThan(p1: BirthDate, p2:BirthDate) returns (Boolean)
  external representation: (before|earlier than|<)\s{p2} ...
  ...

```

Fig. 5 Sample recognizers for linguistically grounding the ontology in Fig. 4

243311. Abigail Huntington Lathrop (widow), Boonton, N. J., b. 1810, dau. of Mary Ely and Gerard Lathrop; m. 1835, Donald McKenzie, West Indies, who was b. 1812, d. 1839.  
 (The widow is unable to give the names of her husband's parents.)  
 Their children:

1. Mary Ely, b. 1836, d. 1859.
2. Gerard Lathrop, b. 1838.

243312. William Gerard Lathrop, Boonton, N. J., b. 1812, d. 1882, son of Mary Ely and Gerard Lathrop; m. 1837, Charlotte Brackett Jennings, New York City, who was b. 1818, dau. of Nathan Tilestone Jennings and Maria Miller. Their children:

1. Maria Jennings, b. 1838, d. 1840.
2. William Gerard, b. 1840.
3. Donald McKenzie, b. 1840, d. 1843. } Twins.
4. Anna Margareta, b. 1843.
5. Anna Catherine, b. 1845.

Fig. 6 An excerpt from p. 419 of *The Ely Ancestry*

values, context keywords, applicable operators, and operator parameters. The data frame for *BirthDate* in Fig. 5 illustrates recognizers for both instance values and operator applicability. Although any kind of textual pattern recognizer is possible, our current implementation supports only regular expressions or combinations of regular expressions and dictionaries. Relationship sets may also have data-frame recognizers. Recognizers for larger ontological components are also possible—*Ontology Snippets*, as we call them.

We explain how the linguistic recognizers work by showing how they apply to an OCRed excerpt from the *The Ely Ancestry* (Beach et al. 1902) in Fig. 6.

- *Lexical object-set recognizers* identify lexical instances in terms of external representations, context, exclusions, and dictionaries. One of the possibly many **external representations** for *BirthDate* in Fig. 5 is “\b1[6–9]\d\d\b”, representing years between 1600 and 1999, with an immediate **left context** of “\b\s”, an immediate **right context** of “[.]”, and **context keywords** that include “\bborn\b(\sin)?”, which may appear close to but not necessarily immediately adjacent to the birth year. Note that these regular-expression patterns match all the birth years in Fig. 6. The **external representations** for *Name* in Fig. 5 illustrate the use of dictionaries and mixed dictionaries and regular expressions. A name in curly braces within a regular expression references a named regular expression (e.g., “{FirstName}” references a dictionary of given names: “AaronlAbdullAbbeyl..”). An **input method** converts a recognized string into an appropriate internal representation—for example, a Julian-date representation in Fig. 5, and an **output method** converts an internal representation to a standard format for display to a user. Applicable **operator methods** are particularly useful for constraints in queries like “List Mary Ely’s children born before 1840” where parameter  $p_1$  comes from an extracted value and  $p_2$  follows “before”.
- *Nonlexical object-set recognizers* identify nonlexical objects through object existence rules, which identify text such as proper nouns, that designate the existence of objects. The object existence rule “{Name}” for the nonlexical object set *Person*, for example, references the regular expressions in the *Name* object set, and when a name is recognized, ML-OntoES generates a *Person* object and associates it with the recognized name.
- *Relationship-set recognizers* identify phrases that relate objects. For example, the regular expression “^{\d{1,2}}\.\s{Person},\sb\s{BirthDate}[.]” for the *Person–BirthDate* relationship set relates Maria Jennings to 1838 and William Gerard to 1840—two of the *Person–BirthDate* relationships that appear in Fig. 6.
- *Ontology-snippet recognizers* identify text patterns that provide instances for groups of object and relationship sets. Recognizers for ontology snippets consist of regular expressions with capture groups and predicate mappings.

To effectively recognize semantic object and relationship instances in text, we must often tune extraction ontologies to the view of the text provided by its author (e.g., tune Figs. 4 and 5 to the author’s view in Fig. 6). An author’s view, however, may differ in its organization and content from the view we wish to have as we query the extracted information. We can obtain the view we want (e.g., Fig. 7) by using the inference-rule component of ML-OntoES.

In our prototype implementation, we use the Jena reasoner (<http://jena.apache.org>) over RDF triples to specify inference rules. Since ML-OntoES is fundamentally specified as a set of  $n$ -ary predicates ( $n \geq 1$ ), the Jena reasoner immediately applies. Moreover, its results are also  $n$ -ary predicates, which lets us conveniently augment an ML-OntoES ontology. We can, for example, have the rules

```
target:Person(x) :- source:Person(x)
target:Person–Gender(x, ‘Male’) :- source:Son(x)
target:Father(x) :- target:Person–Child(x,y),target:Person–Gender(x, ‘Male’)
```

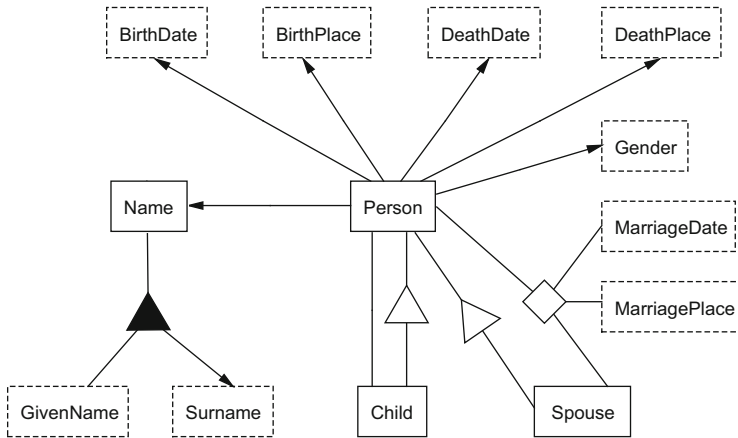


Fig. 7 Target ontology of desired biographical assertions

which, respectively, specify that persons in a source ontology (e.g., Fig. 4) become persons in the target ontology (e.g., Fig. 7), that sons are male, and that persons who have a child and are male are fathers. Furthermore, the Jena reasoner defines a set of built-in predicates that is extensible, and we can create extensions to specify predicates that, for example, can split a name such as “William Gerard Lathrop” into two given names and a surname and that can infer the surname of the children for the culture in which *The Ely Ancestry* was written as the surname of the father. Inferred object and relationship sets may have data-frame recognizers, thus making inferred assertions directly queryable.

In addition to inferring assertions, ML-OntoES also has the ability to reason over the stated and implied assertions to do entity resolution. In our prototype implementation, we use the Duke entity resolver (<http://code.google.com/p/duke>) and generate OWL *same-as* relationships when, for example, Duke discovers that of the three “Mary Ely”s in Fig. 6, only the first and third are the same.

## 2.2 ML-OntoES Monolingual Query Processing

Before query processing begins, ML-OntoES preprocesses a document collection and creates a keyword index and a semantic index. In our prototype implementation, ML-OntoES creates its keyword index with Lucene (<http://lucene.apache.org>) and its semantic index with extraction ontologies. ML-OntoES applies extraction ontologies to text documents to find instance values in the documents with respect to the object and relationship sets in the ontology as explained in Sect. 2.1 and illustrated for Korean in Fig. 3. ML-OntoES returns its semantic index as RDF triples.

Assuming a known context—an identified extraction ontology—ML-OntoES first distinguishes between semantic and keyword text in the query and processes semantics through the semantic index and keywords through the keyword index. ML-OntoES then combines the results and subsequently ranks and displays retrieved documents, for example, as suggested by Fig. 1, allowing users to click on results to view original documents from which information was extracted and, in the case of inferred results, to also see the reasoning chains.

For the French query in Fig. 3, the data-frame recognizers in the French car-ad extraction ontology recognize “Honda” and “moins de 8000” and convert them to the constraints *marque* = “Honda” and *prix* < 8000€. For monolingual query processing, ML-OntoES generates a SPARQL query from these constraints that not only finds cars that satisfy the constraints in its semantic index but also retrieves information about references to its cached copies of the web pages from which ML-OntoES extracted the information—thus making the semantic index an actual index into its known web pages.

Assuming that users wish to have as many of the semantic constraints satisfied as possible and knowing that users may query for constraints not specified in source documents, ML-OntoES generates conjunctive queries and allows SPARQL constraint satisfaction to be optional. Then, for acyclic conceptualizations (e.g., the application ontologies in Fig. 3), ML-OntoES generates queries in a straightforward way: join over edges in the ontologies that connect identified nodes, and filter conjunctively on identified conditions. For the query in Fig. 3, for example, ML-OntoES produces the SPARQL equivalent of  $\pi_{\text{marque,prix}}\sigma_{\text{marque}='Honda' \wedge \text{prix}<8000}(\text{auto-marque} \bowtie \text{auto-prix})$ .<sup>3</sup> For cycles, ML-OntoES identifies all possible paths in the conceptual-model graph that cover identified object and relationship sets and then either acknowledges the ambiguity and returns answers for all paths or discovers that the query explicitly identifies one or more of the paths and returns answers only for these paths.

ML-OntoES processes free-form queries conjunctively. However, like standard search engines, it also provides for advanced-search capabilities for queries that involve disjunctions and negations. When a user requests the advanced-search option for an application, ML-OntoES dynamically generates a form from the application’s extraction ontology. The form provides for negations with a checkbox, disjunctions with click-extended OR buttons, and comparators for all declared comparison operations in the application’s data frames.

For keyword query processing to work well, it is necessary to remove stopwords plus words and phrases intended to convey semantic constraints or result types. Thus, ML-OntoES removes stopwords such as “de” and “en” and a phrase like “moins de 8000”, which it recognizes as generating a semantic constraint. Semantic-phrase removal prevents terms such as “moins” from matching irrelevant tokens in documents. ML-OntoES also removes semantic phrases expressing equality constraints such as “Marque égale Honda”, but for recognized equality constraints,

---

<sup>3</sup>By  $\pi$  and  $\sigma$ , we mean projection and selection, respectively.

it leaves the value word or phrase as a keyword. Thus, in our example, “Honda” becomes a keyword. ML-OntoES also passes quoted phrases, such as «excellent état», to Lucene to process as single-phrase keywords.

### 2.3 ML-OntoES Cross-Language Query Processing

Given a query  $Q$  in language  $L_1$  and an interpretation of  $Q$  with respect to a conceptualization also in language  $L_1$ , ML-OntoES maps  $Q$  from the conceptualization in language  $L_1$  to a corresponding conceptualization in language  $L_2$ . Cross-language conceptualizations are structurally identical, and therefore since the semantic concepts and constraints have a one-to-one correspondence, the implied select-project-join operations for query  $Q$  will be the same in both conceptualizations. Thus, for example, the SPARQL equivalent of the French query  $\pi_{\text{marque.prix}}\sigma_{\text{marque}='Honda' \wedge \text{prix} < 8000}$  (*auto-marque*  $\bowtie$  *auto-prix*) becomes a SPARQL equivalent of the Korean query  $\pi_{\text{제조사,가격}}\sigma_{\text{제조사}='\text{혼다}' \wedge \text{가격} < 11700800}$  (자동차-제조사  $\bowtie$  자동차-가격).

For narrow-domain, data-rich applications, we expect native-language extraction ontologies for different languages/locales to be similar, but not necessarily identical. Thus, when adding a new extraction ontology to ML-OntoES for a new language or new localization of an existing language, we check structural consistency and make adjustments as necessary to retain the structural one-to-one correspondence across all ontologies. In Korean car ads, for example, mention of accidents is common. Assuming the accident concept is not yet part of the existing conceptualizations, we can either drop the concept from the Korean ontology (deeming it not essential) or add it to all other ontologies for the application.

For keywords and instance values in semantic constraints, ML-OntoES uses existing services for currency conversions, keyword translation, unit conversions, and transliterations and uses existing language resources and pay-as-you-go construction for lexicon and commentary translations<sup>4</sup>:

- *Lexicons*. Lexicon mappings substitute one word by another or one word by a small number of others. For common concepts such as colors, corresponding translations are available in cross-language dictionaries. Interestingly, these mappings are not always one to one (e.g., “blue” in Korean is 파랑색 and 파란색 and 청색).
- *Units and Measures*. ISO standard conversion formulas for units and measures are commonly available, and coding them is straightforward. In our implemen-

<sup>4</sup>Our mapping typology here resonates with that of León-Araúz and Faber (this volume), though our lexical type inventory is not as finely articulated.



tation, we use, for example, kilometers for mileage, integers for car years, Julian calendar specifications for dates, and a 24-hour clock for time.

- *Currency*. Because services exist that directly convert amounts in one currency to amounts in any other currency, mappings for currency conversions are direct from one language/localization to another.
- *Transliteration*. Like direct conversion among currencies, transliteration mappings are direct from one language to another.
- *Keywords*. Since keywords can be any word or quoted phrase, we use a general translation service.
- *Commentary*. Ontologies may contain free-form commentary to explain unfamiliar concepts, such as localized tipping protocols.

For answer values returned, we use the mappings to transform values and keywords back into the original language. In Fig. 3, for example, ML-OntoES maps the Korean car make *혼다* first into its language-agnostic equivalent and then into the French “Honda”, and the currency converter converts the Korean Won price 1,100만 원 into 7,826€ and the twice-appearing keyword *혼다* via a general translation service into “Honda (2)”.

Development and maintenance of ML-OntoES cross-language mappings agree in spirit with the principles of Bosca et al. (this volume). Our methods and tools, however, obviously vary somewhat.

### 3 Practicalities

How well ML-OntoES works in practice primarily depends on the accuracy of its linguistic grounding, which, in turn, depends on the quality of its knowledge engineering. For ML-OntoES to be successful, we must sufficiently increase semantic recognition accuracy and sufficiently decrease engineering construction costs.

#### 3.1 Recognition Accuracy

Cross-language query-processing accuracy depends on (1) extraction accuracy in all languages when indexing the semantics in a document collection and (2) cross-language query transformation so that nothing is lost or spuriously added.

To check extraction accuracy, we built French and Korean extraction ontologies for car-ad and obituary applications. The combinations represent typological variety across languages and document diversity in degree of semistructuredness. From 500 French car ads, 1,500 French obituaries, 430 Korean car ads, and 502 obituaries, gathered from several different online sites, we randomly selected about 100 of each of the four combinations to constitute validation and blind test sets (respectively, 20 and 80 of the 100) and used the rest for training (in the sense that we looked at many of them as we built our ontologies).

**Table 1** Car ad within-language extraction results

		Make (%)	Model (%)	Year (%)	Price (%)	Color (%)	Mileage (%)
French	Recall	87	76	96	89	82	98
	Precision	65	67	90	95	47	92
Korean	Recall	99	99	100	100	100	95
	Precision	99	99	100	100	100	95

**Table 2** Obituary within-language extraction results

		Title	Name (%)	Death Date (%)	Funeral		
					Date (%)	Time (%)	Place (%)
French	Recall	76 %	42	80	69	43	38
	Precision	99 %	63	88	70	30	83
Korean	Recall	N/A	97	97	50	50	100
	Precision		97	97	100	100	67

**Table 3** Cross-language query transformation results

Car-ad queries	Recall			Precision		
	$\sigma$ (%)	$\pi$ (%)	$\kappa$ (%)	$\sigma$ (%)	$\pi$ (%)	$\kappa$ (%)
French-to-English	77	86	100	81	90	74
Korean-to-English	98	100	100	93	99	52

Tables 1 and 2 show the results. The car-ad domain is ontologically narrow, and accordingly, our extraction ontologies perform quite well on this domain (as we have come to expect (Embley et al. 2011a)). Precision and recall for Korean car ads are high because these ads mostly have a regular structure, allowing our Korean expert to quickly tune the extraction ontology. The French car ads are more free-form, and so the results are lower. The obituary domain is much broader, and extraction is more challenging—particularly for names and places. Even so, our Korean expert was able to quickly tune the extraction ontology, and performance for most concepts was remarkably high. French extraction was hampered by greater variability and complex sentence structures. For example, there are only 187 names in our Korean surname dictionary, compared with 228,429 in our French surname dictionary, which partially explains the relatively high performance for Korean name extraction.

To check cross-language query transformation accuracy, we asked students in two senior-level database classes to generate car-ad queries which they felt an earlier demo version of a free-form query processor should interpret correctly. The students generated 137 syntactically unique queries, of which 113 were suitable for testing ML-OntoES. To obtain Korean and French queries, we faithfully translated 50 of these 113 into each language.

Table 3 shows the results of interpreting the queries in their respective languages and transforming the internal representation of each query, as understood, into the

internal representation of the query in English. In the table,  $\sigma$  and  $\pi$ , respectively, represent query selection (i.e., conditionals such as “Price < \$12,000”) and query projection (i.e., choice of results to include, e.g., the make and model of a car), and  $\kappa$  represents keywords. Since  $\sigma$  and  $\pi$  translations are always correct, the less-than-perfect  $\sigma$  and  $\pi$  results come from inaccurate within-language query interpretation. The lower recall and precision for French conditionals ( $\sigma$ ) points to a need for better recognizers. More complete synonym sets for French ontological concepts ( $\pi$ ) would increase recall but may decrease precision. Expanded stopword lists in French would remove spurious keywords ( $\kappa$ ) like “list” and “want”. Stopwords in Korean make little sense because most of the standard English-like stopwords are prefixes and suffixes and become part of glyphs. An attempt to remove them after translation often fails because translations themselves are often poor; for example,  $\text{인}$ , which in our query should translate as “which is”—both English stopwords—instead was translated as “inn” (or “hotel”).

### 3.2 Construction Cost

The ML-OntoES architecture requires a substantial amount of information that must be encoded, either by hand or through some automated means. The difficulty of eliciting or otherwise acquiring such data from domain experts—Feigenbaum’s “knowledge engineering bottleneck” (Feigenbaum 1984)—is a decades-old issue.

Our approach substantially mitigates, without completely solving, this problem: our system uses narrow, domain-dependent ontologies that a typical user should be able to specify. We have developed interactive tools for designing and populating ontologies with the requisite types of knowledge, and we are investigating the use of machine learning and linguistic analysis to reduce the cost of developing recognizers for linguistically grounding ontologies. Furthermore, we advocate and practice re-using to the degree possible already extant knowledge sources, and we resonate with similar work being done by other researchers to leverage a wide variety of resources in the boosting of ontology content for crosslinguistic extraction while minimizing the cost (Fu et al. 2012), also convincingly advocated by Bond et al. (this volume).

We assume that end users knowledgeable in a particular domain can create focused, narrow-in-scope ontologies that involve extraction of relevant content from data-rich knowledge sources. In the context of crosslinguistic extraction, ontology creators need to know the languages for which they are designing ontologies. Creation of the ontologies involves specifying concepts, relationships, constraints, and lexical items useful for extraction. Three methods are available for ontology creation and population: (1) programmers can hand-populate them by entering data directly into the data structure; (2) experienced users can interact with the data structure via our custom-designed ontology editor, a tool for specifying ontology content; or (3) domain experts with limited experience can interact with a form-driven interface

that guides the user through design decisions necessary to provide content. The time and effort involved for developing an ontology typically involve one person's efforts over several days, perhaps at the most a week or two, less time if the user has expertise in language, lexicons, and text processing techniques. As with any knowledge engineering task, there is a point of diminishing returns in specifying expert knowledge: more time and effort can be spent developing content to increase performance but at the risk of experiencing the knowledge-engineering bottleneck. A short but representative list of resource types we have used or are considering using for ontology creation and population follows:

- *Lexical databases*: Several publicly available lexical resources—monolingual and multilingual—provide comprehensive information on lexical semantic relations: synonymy, hypernymy, hyponymy, meronymy, word senses, and crosslinguistic mappings. Example resources include the WordNet (<http://wordnet.princeton.edu>), the GlobalWordNet (<http://www.globalwordnet.org>), and the BabelNet (<http://lcl.uniroma1.it/babelnet>).
- *Lexicons*: Specialized lists of narrow-domain words of interest are readily found on the Web: gazetteers for place names, census indexes for person names, and product name databases are some examples. For our evaluation work in Sect. 3.1, we mined pull-down menus from <http://paruvendu.fr> which contains all French automobile make/model combinations and mined tabs from <http://www.encar.com> which lists Korean makes and models.
- *Term banks*: The computerization and subsequent web deployment of vast terminology banks, such as TermiumPlus (<http://www.termiumpus.gc.ca>) and EuroTerm (<http://www.euroterm.org/test1/glossary>), has put literally millions of concepts and their single-word and multiword terms within easy reach of the general public. In prior work, we have shown how to integrate terminological resource content into our ontologies (Lonsdale et al. 2002).
- *Transliteration services*: When crosslinguistic mappings involve different character sets, services can perform character conversion. In our current implementation, we use a Hangul/Roman transliterator (<http://sori.org/hangul/conv2kr.cgi>) for Korean to/from English. Unfortunately, no general transliteration resource appears to be currently available.
- *Translation services*: LabelTranslator (<http://www.neon-toolkit.org>), for example, provides translation (called by others “localization services”) for ontology labels between three European languages. For general-purpose translation, services based on statistical machine translation systems can be used; we currently use Bing (<http://api.microsofttranslator.com/V2/Http.svc/Translate>) when more direct methods are not readily available.

The crosslinguistic aspect of our system involves a star-based architecture similar to notion in Dorr et al. (2006) that maps between languages at the conceptual-model level (Embley et al. 2011c). At the center of the star is a language-agnostic pivot that mediates between language-specific extraction ontologies. Since conceptual

associations are routinely direct, this removes the necessity to translate between languages and allows for recovering the mappings from the isomorphic ontological content. Furthermore, the effort required to add another language to the system only involves developing the relevant knowledge sources for the new language. The complexity of adding a new language to the system is thus reduced from  $O(n^2)$  to  $O(n)$ .

As ML-OntoES becomes more reliant on external resources, it also becomes subject to what Hoekstra calls the “knowledge reengineering bottleneck” in the context of the Semantic Web, with its four new challenges (Hoekstra 2010): (1) Our system is *data dependent* since its effectiveness, robustness, and scalability depend on the appropriateness and quantity of data we incorporate from elsewhere. (2) We have *limited control* over the dirtiness of the data we process and over the coverage of the resources we adopt. (3) ML-OntoES becomes subject to *increased complexity* as disparate resources are integrated into the system. (4) As our system transitions from small-scale systems to large-scale web applications, it assumes *increased importance*. With the star-based architecture of the system and through careful selection of relevant knowledge resources, we hope to be able to strike a pragmatic balance among these issues, at least for data-rich, narrow-domain applications.

## 4 Conclusion

ML-OntoES processes cross-language, hybrid query and keyword-search requests for narrow-domain, data-rich applications in accord with three principles: (1) monolingual semantic indexing based on extraction ontologies, (2) monolingual extraction-ontology-based semantic analysis of user queries, and (3) structurally identical application ontologies to facilitate conceptual-level cross-language mappings:

1. For query processing to work in reasonable time, semantic indexes must exist. ML-OntoES creates semantic indexes by crawling web pages and documents on the web with application-dependent, monolingual extraction ontologies. Then, for each assertion found (as explained in Sect. 2.1), we can record the assertion’s objects in their identified ontological object sets and its relationships among the objects in its identified ontological relationship sets and associate the object and the relationship pointers into a cached copy of the page or document.
2. When a user submits a query, it is best if the system already knows the context in which the query is asked—that is, already knows which ontology or set of ontologies, prepopulated with assertions, should be used to return an answer. Otherwise, the system must search for an application ontology (or a set of application ontologies) by applying candidate extraction ontologies to the query and checking the coverage. Indexes over words and common conceptualizations such as dates and currencies can speed up the process of locating appropriate

ontologies for the query. Then, as explained in Sect. 2.2, ML-OntoES can monolingually construct a query with respect to the structure of the ontology.

3. As noted in Sect. 2.3, since all language-and-locale versions of extraction ontologies for a particular application are structurally identical, generated query expressions have the same form in all versions, and only the instance values, if any, need translation. ML-OntoES uses cross-language dictionaries for word substitutions, standard conversion formulas for units and measures, online currency converters for currency exchange, and transliteration services for name conversions. Keyword and commentary translation are more difficult to translate accurately. But rough approximations, as provided by online translators, are often sufficient. For critical vertical applications where specialized keywords and jargon words matter in hybrid queries, special application-dependent keyword and keyword-phrase cross-language dictionaries can be developed as a supplement for online translators. Likewise, when commentary is critical, such as for business transactions and detailed instructions, careful translations would need to be written, if they do not already exist.

Our prototype implementation demonstrates feasibility, but as a practical matter, for ML-OntoES to be successful, extraction-ontology recognition accuracy must be high (Sect. 3.1), and extraction-ontology construction costs must be low (Sect. 3.2). Summarizing our discussion of these issues in Sect. 3, we point out that the knowledge engineering required for car ads and obituaries returned reasonably good precision and recall results for French and particularly good for Korean, and that the time and effort required to develop the extraction ontologies, given the lexical resources available to us, are within reason. This “knowledge-engineering bottleneck” is, however, a drawback of ML-OntoES.

Because of this drawback, our current and expected future efforts for ML-OntoES are focused on mitigating extraction-ontology construction costs. Focusing on the vertical domain of historical documents and particularly family-history documents (Embley et al. 2011b), we are exploring ways to automate the construction of extraction ontologies. For lists, which are commonly found in family-history documents, we have been able to generate both regular-expression and HMM recognizers that accurately extract genealogical assertions of interest and insert them into ontological structures (Packer and Embley 2013). We are currently working on automating the extraction of more general text patterns found in semistructured documents and on combining a dependency parser with a semantic reasoner to generate assertions that can be inserted into a target ontology. The domain of family history is particularly in need of cross-language query processing, especially for untrained users because many people have ancestors who have come from countries with a language foreign to their own.

**Acknowledgments** We are grateful to Tae Woo Kim and Rebecca Brinck for annotating our Korean and French document sets. We are also grateful to the reviewers for their insightful feedback and challenging questions.

## References

- Aggarwal, N., Polajnar, T., & Buitelaar, P. (2013). Cross-lingual natural language querying over the web of data. In E. Métais, F. Meziane, M. Saraee, V. Sugumaran, & S. Vadera (Eds.), *Natural Language Processing and Information Systems: 18th International Conference on Applications of Natural Language to Information Systems (NLDB 2013)*. *Lecture Notes in Computer Science* (Vol. 7934, pp. 152–163). New York: Springer.
- Beach, M., Ely, W., & Vanderpoel, G. (1902). *The ely ancestry*. New York: The Columer Press. On-line book: <http://www.archive.org/details/-elyancestrylinea00beac>.
- Buitelaar, P., Cimiano, P., Haase, P., & Sintek, M. (2009). Towards linguistically grounded ontologies. In *Proceedings of the 6th European Semantic Web Conference (ESWC'09)* (pp. 111–125). Heraklion: Greece.
- Declerck, T., Krieger, H.-U., Thomas, S., Buitelaar, P., O’Riain, S., Wunner, T., et al. (2010). *Ontology-based multilingual access to financial reports for sharing business knowledge across Europe* (pp. 67–76). Budapest: Memolux Kft.
- Dorr, B., Hovy, E., & Levin, L. (2006). Machine translation: Interlingual methods. In *Natural language processing and machine translation. Encyclopedia of Language and Linguistics* (2nd ed., pp. 383–394). Oxford, UK: Elsevier Ltd.
- Embley, D. (1980). Programming with data frames for everyday data items. In *Proceedings of the 1980 National Computer Conference*, Anaheim, CA (pp. 301–305).
- Embley, D., Liddle, S., & Lonsdale, D. (2011a). Conceptual modeling foundations for a web of knowledge. In D. Embley & B. Thalheim (Eds.), *Handbook of conceptual modeling: Theory, practice, and research challenges* (Chap. 15, pp. 477–516). Heidelberg, Germany: Springer.
- Embley, D., Liddle, S., Lonsdale, D., Machado, S., Packer, T., Park, J., et al. (2011b). Enabling search for facts and implied facts in historical documents. In *Proceedings of the International Workshop on Historical Document Imaging and Processing (HIP 2011)*, Beijing, China (pp. 59–66).
- Embley, D., Liddle, S., Lonsdale, D., & Tijerino, Y. (2011c). Multilingual ontologies for cross-language information extraction and semantic search. In *Proceedings of the 30th International Conference on Conceptual Modeling (ER 2011)*, Brussels, Belgium (pp. 147–160).
- Feigenbaum, E. (1984). Knowledge engineering: The applied side of artificial intelligence. *Annals of the New York Academy of Sciences*, 426, 91–107.
- Fu, B., Brennan, R., & O’Sullivan, D. (2012). A configurable translation-based cross-lingual ontology mapping system to adjust mapping outcomes. *Web Semantics: Science, Services and Agents on the World-Wide Web*, 15, 15–36.
- Hoekstra, R. (2010). The knowledge reengineering bottleneck. *Semantic Web—Interoperability, Usability, Applicability*, 1, 1–5.
- Lonsdale, D., Ding, Y., Embley, D., & Melby, A. (2002). Peppering knowledge sources with SALT: Boosting conceptual content for ontology generation. In *Proceedings of the AAAI Workshop: Semantic Web Meets Language Resources: The 18th National Conference on Artificial Intelligence*, Edmonton, AB, Canada (pp. 30–36).
- Mahesh, K. (1996). Ontology development for machine translation: Ideology and methodology. Technical Report MCCS-96-292. Albuquerque, NM: Computing Research Laboratory, University of New Mexico.
- Montiel-Ponsoda, E., de Cea, G. A., Gómez-Pérez, A., & Peters, W. (2011). Enriching ontologies with multilingual information. *Natural Language Engineering*, 17(3), 283–309.
- Olive, J., Christianson, C., & McCary, J. (Eds.). (2011). *Handbook of natural language processing and machine translation: DARPA global autonomous language exploitation*. New York: Springer.
- Packer, T., & Embley, D. (2013). Cost effective ontology population with data from lists in OCRed historical documents. In *Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing (HIP 2013)*, Washington, DC, USA (pp. 1–8).

- Peters, C., Braschler, M., & Clough, P. (2012). *Multilingual information retrieval: From research to practice*. New York: Springer.
- Tijerino, Y. (2010). Cross-cultural and cross-lingual ontology engineering. In *Proceedings of the 2010 Workshop on Cross-Cultural and Cross-Lingual Aspects of the Semantic Web*, Shanghai, China (pp. 44–53).



# Collaborative Management of Multilingual Ontologies

Alessio Bosca, Mauro Dragoni, Chiara Di Francescomarino,  
and Chiara Ghidini

**Abstract** The usage of multilingual resources has registered a significant increase in the last decade. Multilinguality is used in several fields of computer science, and the necessity of managing multilingual information has become an important as well as critical task. In this chapter, we face the problem of the management of multilingual ontologies by describing which problems arise in its context, with a particular emphasis on the collaborative modeling aspect. We present a tool that provides features able to support the collaborative management of multilingual ontologies, and we describe a real-world use case in which the exploitation of multilingual ontologies improves the effectiveness of information technology systems.

**Key Words** Collaborative modeling • Multilingual ontologies • Semantic applications

## 1 Introduction

With the recent rapid diffusion over the Web of worldwide distributed document bases, the question of multilinguality is becoming increasingly relevant. So far, research and development activities have been concentrated on monolingual environments, and in the large majority of cases, the default language has been English. Although English admittedly tends to play a predominant role in international communications, the diversity of the world's languages and cultures gives rise to an enormous wealth of knowledge and ideas. A clear example of this scenario is represented by users throughout the world that, independently of their native tongue, want to access the massive volumes of information available over the networks and, in particular, over the World Wide Web.

---

A. Bosca  
Celi s.r.l., Via S. Quintino 31, 10131 Torino, Italy  
e-mail: [alessio.bosca@celi.it](mailto:alessio.bosca@celi.it)

M. Dragoni (✉) • C. Di Francescomarino • C. Ghidini  
FBK-IRST, Trento, Italy  
e-mail: [dragoni@fbk.eu](mailto:dragoni@fbk.eu); [dfmchiara@fbk.eu](mailto:dfmchiara@fbk.eu); [ghidini@fbk.eu](mailto:ghidini@fbk.eu)

Together with the growth of these requests, the ontology engineering community has started exploring the possibility to use multilinguality for increasing the expressiveness of the semantic artifacts, which are more and more used for enhancing the information associated to the available resources. A recent example, which witnesses the importance of multilinguality in the field of ontology engineering, is provided by the Monnet Project,<sup>1</sup> which targets the problem of multilingual information access at the semantic level (Declerck et al. 2010).

Building multilingual ontologies, however, is a complex activity (Espinoza et al. 2008a) that requires to tackle a number of problems spanning from the translation of labels and descriptions associated to a given ontology entity to the adaptation of the ontology to a concrete language and cultural community. Moreover, building and maintaining such artifacts demand for a complex collaboration between different (geographically distributed) users with different competencies. Indeed, besides the well-known collaboration issues between domain experts and ontology engineers, which have been already widely investigated in the literature, in such a scenario, a new role, the language expert, who is in charge of supervising the translation of the ontology in different languages, is involved.

In this chapter, we focus on the collaborative ontology engineering in a multilingual environment, and we provide a technical solution for supporting the issues it raises, ranging from semantic approaches to face multilinguality to Web-based modeling tools to tackle collaboration. Moreover, we present a practical use case: the construction of a multilingual ontology in the context of a European-funded project and its exploitation for a cross-language information retrieval (CLIR) task.

This chapter is structured as follows: Sect. 2 presents a brief state of the art related to the multilingual ontology engineering and exploitation and a quick overview of the available tools. In Sect. 3, we introduce typical collaboration issues which arise when working in a multilingual ontology engineering and evolution context. In Sect. 4, we present a collaborative tool for creating and evolving ontologies and the customizations that have been implemented for addressing the specific requirements of a multilingual environment. Section 5 describes the Organic.Lingua EU-funded project and the role of the ontology in its context, while in Sect. 6 we show how multilingual ontologies can be exploited for the CLIR task on a concrete use case in the context of the Organic.Lingua project. Finally, Sect. 7 concludes.

## 2 Related Work

In the recent years, the usage of multilingual knowledge in the Semantic Web environment has considerably increased, and several works concerning the modeling and the exploitation of multilingual artifacts have been published.

---

<sup>1</sup><http://www.monnet-project.eu>.

On the modeling side, the use of automatic translation processes for building the multilingual layer of an ontology has been explored by Espinoza et al. (2008b). Here, the authors propose LabelTranslator, a system that automatically localizes ontologies by adapting them to the concrete target language and cultural community. LabelTranslator takes as input an ontology whose labels are described in a source natural language and returns, for each ontology label, the most probable translation into a target natural language.

On the exploitation side, multilingual ontologies have been used in the development of multilingual search systems and portals. Stamou et al. (2004) presented a structured multilingual conceptual repository that has been employed as the backbone of a conceptual indexing and retrieval system. Their conceptual warehouse originates from a multilingual semantic network, called BalkaNet, and its Interlingual Index, which was enriched with domain ontology information inherited from the Suggested Upper Merged Ontology (SUMO) ontology. Other examples of the multilinguality exploitation for the enhancement of information browsing and search portals are discussed by Vouros et al. (2005) and by Bo et al. (2003). In the former, the authors present a multilingual information system that enables users to search for information in an ontology-driven and content-based way, supported by lexical resources and reasoning services. In the latter, the authors present an extension of the Developing Ontology-Grounded Methods and Applications (DOGMA) ontology engineering framework by equipping it with features for coping with context and multilinguality issues.

Also the ontology-matching problem took advantage of multilinguality. Spohr et al. (2011) discuss several approaches in which multilinguality is used to learn a matching function between two ontologies starting from a small set of manually aligned concepts and evaluate them on different pairs of financial accounting standards. Differently, in the work presented by dos Santos et al. (2008), the mappings between multilingual ontologies are computed by using, first, a lexical database and applying, in a second step, a set of specialized agents adopting different mapping approaches. A survey about multilingual and cross-lingual ontology matching is presented by Trojahn et al. (this volume). In particular, the authors offer a classification of existing multilingual and cross-lingual matching approaches, as well as an overview of the resources (data sets, systems, and strategies) used for their evaluation.

Concerning the modeling of specific domains, among which the medical one is the most common, the use of multilinguality has been explored in several works. Nyulas et al. (2012) presented the work carried out in cooperation with the World Health Organization (WHO) for modeling the International Classification of Traditional Medicine (ICTM), a standardized system for encoding and collecting health statistics data related to traditional medicine practice throughout the world. They describe how the multilingual content has been modeled, the Web platform used for editing, and some of the challenges encountered in the multilingual modeling and in the use of the platform. Other works related to the medical domain are described by Elberichi et al. (2012), where they addressed the issue

of the classification of multilingual Web documents based on an ontology in the biomedical domain, and by Collier et al. (2006), where they presented a method for developing a new conceptual structure and a multilingual terminological resource that focuses on priority pathogens and on the diseases that they cause.

Multilingual ontologies have been applied also in other domains beyond the medical one. Kerremans et al. (2003) presented the termontology method, a method for the representation of multilingual and culture-specific knowledge, in the domain of value-added tax (VAT). In the work of Liu and Ma (2010), an ontology-based methodology for the development of research and development project management systems with multilingual support is presented. In the discussed approach, a four-layer multilingual ontology consisting of domain model, application model, user model, and linguistic model is proposed.

Unfortunately, all these approaches mainly focus on modeling and exploitation of multilingual ontologies while neglecting collaboration aspects and issues that the management of a multilingual resource necessarily raises.

Similarly, very few tools are available for collaborative management of multilingual ontologies. The only instrument supporting the management of multilinguality in ontologies is *NeOn* (Espinoza et al. 2008a); however, it does not provide facilities for supporting collaboration. On the contrary, tools like *Knoodl*<sup>2</sup> and *Protégé* (Gennari et al. 2003) only recently have been extended to support the collaboration between users for the creation and the evolution of ontologies, but they do not deal with multilinguality issues, which have significantly grown in importance during the last years (Peters et al. 2008).

### 3 Collaboration Aspects in the Management of Multilingual Ontologies

It is nowadays well established that creating ontologies has become a teamwork activity, as it requires a range of knowledge and skills hardly findable all together in a single person. For this reason, collaborative aspects in ontology modeling have been investigated, and several works to support and enhance collaboration in this context have been presented (see, e.g., Palma et al. 2011; Sure et al. 2002; Tudorache et al. 2010; Dimitrova et al. 2008; Di Francescomarino et al. 2012). The requirements and features that have emerged from these studies highlight the need to support collaboration in an articulated way: from supporting the collaboration between who understands the domain to be represented (the domain expert) and who has proper expertise in ontology modeling (the knowledge engineer) to supporting communication, discussion, and decision making between (geographically) distributed teams of ontology contributors.

---

<sup>2</sup><http://www.knoodl.com>.

It is not rare to witness a situation where the modeling team is geographically distributed and/or users may not be able to participate in physical meetings. Supporting collaboration requires enabling the awareness of the user on the evolution of the modeling artifacts, favoring the coordination of the modeling effort within the team, as well as fostering the communication of the modeling choices and decisions made among the modeling actors.

Multilinguality adds the linguistic problem to the classical collaborative ones. Indeed, the construction of multilingual layer of an ontology cannot be reduced to the simple translation of concepts, labels, and definitions, but it demands that each term is adapted to the culture of the target language. Therefore, besides the classical domain expert and knowledge engineer roles, a new role is required to manage the ontology multilingual layer: the *language expert*. The task of the language expert is managing the translations carried out on the ontology entities, providing term translation, and coordinating the translation activities, by approving translations suggested by other actors involved in the ontology management process.

Coordinating all these (possibly geographically distributed) experts with their backgrounds, skills, and tasks demands for appropriate instruments able to guide and support them through the collaborative building and maintenance of the multilingual artifact. For example:

- Domain experts need to be supported in contributing to the ontology construction, in starting or participating to discussions, in suggesting actions to be taken on the ontology, and in commenting on existing issues.
- Knowledge engineers should be notified when new concepts are added or existing ones are updated in order to be able to make decisions on them; moreover, they should be provided with the appropriate means to either approve or discard provided suggestions and proceed with creations and updates on the main-language version of the ontology.
- Language experts should be put in the condition to provide, check, and revise the translations of entity labels and definitions, either manually or supported by automated translation services.

The specific tasks performed by each expert and the basic flow among them seem hence quite straightforward: knowledge engineers are usually in charge to formalize, refine, and accept the domain experts' input, while language experts are in charge to translate or revise translations of added/updated concepts and descriptions. In this view, no strict and rigid methodologies are required but rather a flexible way to support and guide experts in the collaborative management of multilingual ontologies.

In the next section, we introduce a wiki system for the collaborative multilingual ontology authoring which addresses all the abovementioned desiderata, thus enabling an effective collaborative modeling of multilingual ontologies.

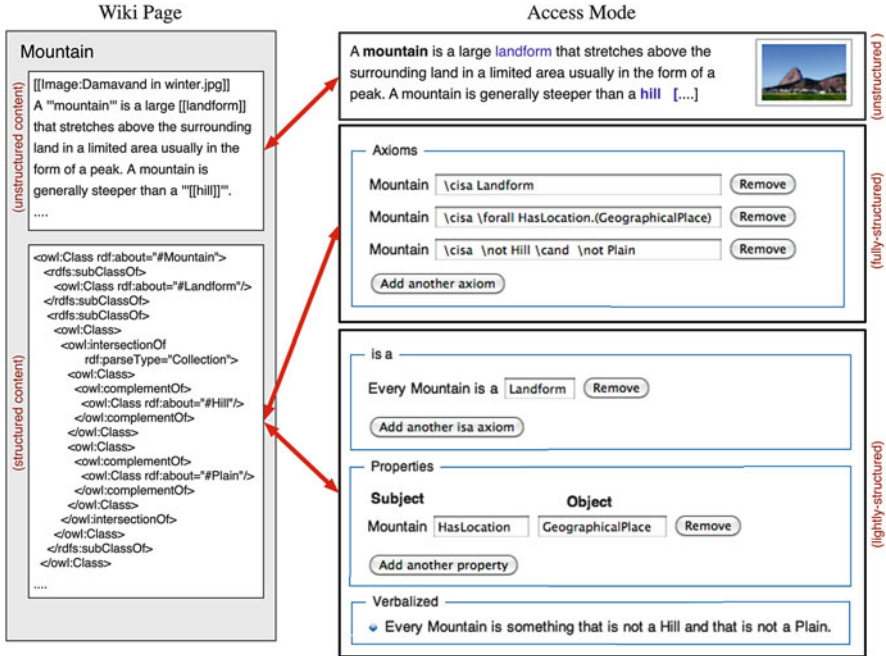


Fig. 1 A page and the access modes in MoKi

## 4 The MoKi Tool

MoKi<sup>3</sup> is a collaborative MediaWiki-based (Wikimedia Foundation, n.d.) tool for modeling ontological and procedural knowledge in an integrated manner.<sup>4</sup> MoKi is grounded on three main pillars, which we briefly illustrate with the help of Fig. 1:

- Each basic entity of the ontology (i.e., concepts, object and datatype properties, and individuals) is associated to a wiki page. For instance, the concept Mountain in Fig. 1 is associated to a wiki page which contains its description.
- Each wiki page describes an entity by means of both unstructured (e.g., free text, images) and structured (e.g., OWL axioms) contents.
- A multimodal access to the page content is provided to support easy usage by users with different skills and competencies.

The multimodal access is the key feature that permits the support of collaboration between different types of users. Figure 1 shows the three access modes,

<sup>3</sup><http://moki.fbk.eu>.

<sup>4</sup>Though MoKi allows to model both ontological and procedural knowledge, here we will limit our description only to the features for building ontologies.

implemented in MoKi, for accessing the unstructured and structured content of the wiki page:

- The unstructured access mode allows the user to edit/view the unstructured part of a MoKi page. Editing/viewing occurs in the standard MediaWiki way.
- The fully structured access mode allows the user to edit/view the structured part of a MoKi page using the full OWL 2 expressiveness<sup>5</sup> and is meant to be used by knowledge engineers to author the formal statements describing the entity associated to the wiki page.
- The lightly structured access mode enables users to edit/view the content of the structured part of the MoKi page in a simplified way. This access mode consists of a form meant to be used by domain experts and contains statements that correspond to all the axioms in the fully structured access mode. In the upper part, the user can view and edit simple statements which can be easily converted to/from OWL statements. An example is the uppermost statement “Every Mountain is a Landform” in the lightly structured access mode of Fig. 1. The bottom part of the form provides a verbal description [automatically obtained via the OWL 2 Verbalizer (Kaljurand and Fuchs 2007)] of those OWL statements which cannot be intuitively translated/edited as simple statements in the upper part of the page. The purpose of this verbal description is to give the domain expert a flavor of the complex statements that the knowledge engineer has formalized. If doubtful about some of the statements, the domain expert can mark them and ask for a clarification using, e.g., the discussion mechanism.

Moreover, MoKi presents a set of MediaWiki-based collaborative editing functionalities, such as:

- *Discussions*: To discuss about challenging issues related to the ontology modeling. It is possible to discuss on single ontology entities or on (a part of) the whole model. Comments in the discussion pages are organized in threads, with details on the user and date/time associated to each comment.
- *Watchlist*: To monitor interesting ontology entities. Any change performed on monitored ontology entities is notified (with messages and email alerts) to the user.
- *Notifications*: To inform users about ontology changes that are relevant for them. E-mail or message notifications are automatically sent, in case changes to pages in the users watchlist occur. Users can also send specific notifications, soliciting a confirmation or revision on some aspects of the ontology from particular users.
- *History and revision*: To track changes and comments added on a specific ontology entity.

A comprehensive description of MoKi is presented by Ghidini et al. (2012).

MoKi, already equipped with a dedicated view for the domain experts and with the collaborative functionalities typical of the wiki systems, has been customized

---

<sup>5</sup>We adopt the syntax of *latex2owl*: <https://dkm.fbk.eu/index.php/Latex2owl>.

in order to meet the specific needs of the collaborative multilingual ontology management. In detail, it has been enhanced with (a) a set of multilingual features for enabling both manual and automatic translation of labels and descriptions associated to the ontology entities and (b) a set of collaborative features specifically targeting linguistic issues. Translating domain-specific ontologies, in fact, demands that experts discuss and reach an agreement not only on modeling choices but also on (automated) term translations. These facilities enable the language expert to manage the translations carried out on the ontology entities.

## 4.1 Supporting the Collaborative Management of Multilingual Ontologies with MoKi

In this subsection, we briefly describe the main customizations implemented in MoKi for providing support to the collaborative management of multilingual ontologies.

### 4.1.1 Domain and Language Expert View

The semistructured access mode, dedicated to the domain and language experts, has been equipped with functionalities that permit language experts to accomplish the revisions of the linguistic layer. This set of functionalities allows them to revise the translations of names and descriptions of each entity (concepts, individuals, and properties). For facilitating the browsing and the editing of the translations, a quick view box has been inserted into the mask (see Fig. 2); this way, language experts are able to navigate through the available translations, invoke possible third-party translation services connected to MoKi for retrieving a translation suggestion, or, alternatively, edit the translation by themselves (Fig. 3).

Multilingual component

Select language: English ▾

Translation in the language:	<b>English</b>
Concept name	agricultural method
Concept description	Practices used to enhance crop and livestock health and prevent weed, pest or disease problems without the use of chemical substances.

Fig. 2 Multilingual box for facilitating the entity translation



The image shows a software interface for editing entity translations. It is titled "Entity translation" and is divided into two main sections. The first section is for the entity name "agricultural method". It has a "Translation:" field with the text "método agrícola" and a "Suggest translation" button below it. The second section is for the entity description. The description is "Practices used to enhance crop and livestock health and prevent weed, pest or disease problems without the use of chemical substances." The "Translation:" field contains the Spanish text "Practicas empleadas para mejorar la salud de los animales y los cultivos, así como para prevenir problemas con las adventicias, las plagas u otras enfermedades sin utilizar sustancias químicas." Below this field is another "Suggest translation" button. At the bottom of the interface are "Save" and "Cancel" buttons.

Fig. 3 Quick translation box for editing entity translations

#### 4.1.2 Approval and Discussion Facilities

Given the complexity of translating domain-specific ontologies, translations often need to be checked and agreed upon by a community of experts. This is especially true when ontologies are used to represent terminological standards which need to be carefully discussed and evaluated. To support this collaborative activity, we foresee the usage of the wiki-style features of MoKi, expanded with the possibility of assigning specific tasks of ontology entity translation to specific experts who need to monitor, check, and approve the suggested translations. This customization promotes the management of the changes carried out on the ontology (both at domain and linguistic layer) by providing the facilities necessary to manage the ontology entity life cycle.

These facilities may be split in two different sets of features. The first group may be considered as a monitor of the activities performed on each entity page. When changes are committed, approval requests are created. They contain the identification of the expert in charge of approving the change, the date in which the change has been performed, and a natural language description of the change. Moreover, a mechanism for managing the approvals and for maintaining the history of all approval requests for each entity is provided. The second set of features contains the facilities for managing the discussions associated with each entity page. A user interface for creating the discussions has been implemented together with a notification procedure that alerts users when new topics/replies, related to the discussions they are following, are posted.

List all Concepts

Number of concepts in the Domain Model: 62







Select language: English		Select language: Italiano	
Concept	Description	Concept translation	Description translations
Activity	A type of action performed by an agent in general sense.	attività	
agricultural method	Practices used to enhance crop and livestock health and prevent weed, pest or disease problems without the use of chemical substances.	agrario metodo	le pratiche vegetali e animali usati per promuovere la salute e la prevenzione delle malattie, parassiti e infestanti problemi senza l'uso di sostanze chimiche.  
europaean agricultural method	Agricultural techniques used in Europe.	metodo agricolo europeo	le tecniche agricole utilizzate in europa.  
animal origin processed product	Any product of animal origin canned, cooked, frozen, concentrated, pickled or otherwise prepared to assure its preservation in transport, distribution and storage, but does not include the final cooking or preparation of a food product for use as a meal or part of a meal such as may be done by restaurants, catering companies or similar establishments where	animale sorgente processed prodotto	

Fig. 4 View for comparing entity translations

### 4.1.3 Quick Translation Feature

For facilitating the work of language experts, we have implemented the possibility of comparing two lists of translations side by side. In this way, the language expert in charge of revising the translations can avoid to navigate through the entity pages and speed up the revision process. Figure 4 shows such a view, by presenting the list of concepts in English and the corresponding Italian translation. At the right of each element of the table, a link allows to invoke a translation box (as the one in Fig. 3) that gives the opportunity to quickly modify the translation without opening the corresponding entity page. Finally, in the last column, the presence of a flag indicates that some changes have been performed on that concept and that a revision/approval of the changes is required.

### 4.1.4 Ontology Translator Component

This component manages the translation operations provided by MoKi through external automatic translation services. When a translation, for an entity name or description, is requested, the ontology translator invokes the external translation services. The component sends the request to the APIs exposed by the third-party translation services, and after the retrieval of the result, the representation of the entity is updated with the returned information.

### 4.1.5 Ontology and Interface Multilingual Facilities

In order to complete the set of features available for managing the multilingual aspects of the tool, MoKi has been equipped with two further components that permit users to manage the language of the ontology and of the tool interface, respectively.

Indeed, besides allowing users to select the language to be used for showing the ontology among the available ones, MoKi also gives the possibility to add a new language to the ontology. In this last case, the ontology translator component described above is invoked for retrieving, for each entity described in the ontology, the translation of its label and description in the new language. Also the ontology export functionality, which enables the ontology export in the OWL format, has been revisited in the light of multilinguality by adding the possibility to choose, among the available languages, the ones in which the ontology has to be exported. This customization has not been implemented for addressing the management of the multilingual ontology per se but for improving the usability of the tool in a multilingual context.

Similarly, concerning the tool interface, besides the possibility to switch among the available languages, MoKi also provides a module to add a new language to the tool interface and to manage the translation of its labels. This module has been implemented on top of the multilingual features of MediaWiki.

### 4.1.6 Linked Open Data Service

In order to permit the exposure of the ontology artifact to third-party components, MoKi has been equipped with a service that exposes entity information by using the Linked Open Data format. Such a service offers the possibility to perform remotely operations on the ontology; examples of available remote operations are the retrieval of the entire ontology, the retrieval of part of it, or the possibility to edit the ontology, e.g., by adding a new translated label. The service provides a RESTful interface for receiving the requests, while the results are exposed by using the Simple Knowledge Organization System (SKOS) language.<sup>6</sup> This customization has been implemented for providing an exposure feature that permits the linking between MoKi and external third-party tools that want to exploit the multilingual artifact.

The customized version of MoKi has been extensively used in the context of the Organic.Lingua EU project, which provides a valid test-bed for the application of the tool in a real-world scenario. In the next section, we present the project, and we describe how a modeled multilingual ontology has been used in that context.

---

<sup>6</sup><http://www.w3.org/2004/02/skos/>.

## 5 The Organic.Lingua Project

Organic.Lingua (<http://www.organic-lingua.eu>) is an EU-funded project that aims at providing automated multilingual services and tools facilitating the discovery, retrieval, exploitation, and extension of digital educational content related to organic agriculture and agroecology. In particular, the project aims at providing, on top of a Web portal, cross-lingual facility services enabling users to (a) find resources in languages different from the ones in which the query has been formulated and/or the resource described (e.g., providing services for cross-lingual retrieval), (b) manage metadata information for resources in different languages (e.g., offering automated metadata translation services), and (c) contribute to evolve the content (e.g., providing services supporting the users in content generation).

The accomplishment of these objectives is reached in the Organic.Lingua project by means of two components: on the one hand, a Web portal offering software components and linguistic resources able to provide multilingual services and, on the other hand, a conceptual model (formalized in the *organic agriculture* ontology) used for managing information associated with the resources provided to the final users and shared with other components deployed on the Organic.Lingua platform. In a nutshell, the usage of the *organic agriculture* ontology is twofold:

- Resource annotation: Each time a content provider inserts a resource in the repository, the resource is annotated with one or more concepts extracted from the ontology. The list of available concepts is retrieved by using an ontology service deployed in the ontology management component (shown in Sect. 4). Then, this list is exploited for annotating the learning resources published on the Web portal.
- Resource retrieval: When Web users perform queries on the system, the ontology is used by the back-end information retrieval system to perform advanced searches based on semantic techniques. Moreover, the ontology is also used by the CLIR component for performing cross-language queries.

The next section focuses on the resource retrieval use case, showing how the exploitation of a multilingual ontology is able to increase the effectiveness of CLIR systems.

## 6 Exploiting Multilingual Ontologies: A Use Case

Among the possible ways of exploiting multilingual ontologies, cross-language retrieval is one of the most useful. In this section, we describe the concrete exploitation of the *organic agriculture* multilingual ontology on a resource retrieval use case in the context of the Organic.Lingua project.

As introduced in Sect. 5, the *organic agriculture* multilingual ontology, used for annotating documents, is also exploited by the CLIR system to retrieve documents in different languages. The CLIR system uses the multilingual ontology annotations,

obtained invoking the MOKi ontology service, for creating a search index for each document, i.e., for enriching the document context with multilingual information. In detail, not only the multilingual component of the ontology is used to build the search index but also the ontology structure. Each concept used for annotating the document is expanded by considering its ontological parents and indexing them according to a decreasing weight that depends on their semantic distance from the concept (Dragoni et al. 2012).

This mechanism allows the CLIR system to exploit the translated labels at query time. For instance, given a document containing both text and annotations in Spanish, the translated labels of each annotation are retrieved from the ontology and stored with the document content into the index. In this way, annotated documents may be retrieved by performing queries in any available language. This approach allows the system to continuously evolve the information contained in the search index by integrating, at each ontology update, the most recent labels or the newly added languages. In order to assess the effectiveness of the CLIR system and to estimate the contribution of the ontology in the retrieval process, an automatic evaluation has been performed on the Organic.Lingua resource repository.

## ***6.1 Experimental Settings***

The evaluation of the Organic.Lingua CLIR system has been inspired by the activities of the Cross-Language Evaluation Forum (CLEF<sup>7</sup>), one of the major conferences concerning the evaluation of multilingual information access systems. Based on this methodology, the resources used for such an evaluation include:

1. A set of queries that express information needs in a given language identified with a unique ID. The approach adopted for selecting the queries consisted in choosing the most popular searches performed by real users on the Organic.Lingua portal filtered by domain experts. In this way, we are able to cover as many topics as possible while avoiding similar queries.
2. A collection of documents that satisfies the information needs expressed in the queries. In the Organic.Lingua test environment, this corpus is composed of a multilingual collection of about 12,000 documents.
3. A gold standard that, for each query, provides the list of the relevant documents used to evaluate the results provided by the CLIR system. In the provided evaluation, the gold standard was manually created by the domain experts. It contains only results that are related to queries expressed or translated in English and that have at least one field (either a textual or an annotation one) in English.

---

<sup>7</sup><http://www.clef-initiative.eu>.

For evaluating the effectiveness of the CLIR system, different standard metrics have been adopted. Besides the well-known Precision and Recall measure, other metrics emerged in the IR community. By keeping as reference the CLEF evaluation campaigns, the metrics used in recent years include R-Precision, Precision@X (representing the Precision obtained after X documents, i.e., P@10 is the precision after 10 docs), and the mean average precision (MAP). Since the evaluation of the Organic.Lingua CLIR system is based on the methodology introduced by CLEF (Braschler and Peters 2003; Agosti and Ferro 2007), the same metrics will be used for evaluating the described system.

## 6.2 Evaluation and Discussion of the Results

The set of queries considered in the experiment is composed of queries in 11 different languages: French, Italian, Spanish, German, Polish, Portuguese, Hungarian, Turkish, Estonian, Latvian, and Greek. The queries have been translated in English by using the translation module of the CLIR system, and they have been used to perform the retrieval from the Organic.Lingua document collections. The CLIR system has been evaluated by adopting two different configurations, and the results have been compared with the gold standard, according to the metrics described above:

1. *Base configuration*: Each query is translated in English by using the CLIR system, and it is performed on the textual fields (i.e., title, abstract, and content) of the indexed documents.
2. *Configuration with semantic expansion*: This setting exploits the multilingual ontology labels used for enriching the representation of each document. Each query is translated in English, and it is performed on both the textual and the annotation fields of the indexed documents. In this way, not only the documents in the same language of the query, but any of the documents annotated with corresponding concepts in another language, can be retrieved.

Tables 1 and 2 report the results of the performed evaluation, grouped by configuration type.

Observing the results, we can notice that the CLIR system effectiveness is in line with the state of the art emerged in CLEF campaign (Ferro and Peters 2010). In particular, the usage of the semantic expansion setting shows a relevant increase of the Precision for the higher parts of the produced ranks (Precision@5, Precision@10, Precision@20) and also for the corresponding average Recall. Table 3 presents the percentage gain of the average Recall and Precision@10. Despite a significant increase in the Precision and Recall, the results show as well a substantial invariance for what concerns the MAP and the average Precision@Recall values. Such a phenomenon originates from the higher number of documents retrieved with

**Table 1** Base configuration

Lang	MAP	Precision@5	Precision@10	Precision@20	Avg. Recall	Avg. Precision @Recall
en	0.7261	0.7917	0.6896	0.5865	0.9635	0.6897
el	0.3731	0.3833	0.3479	0.3104	0.8253	0.3756
lv	0.3348	0.325	0.3187	0.2948	0.703	0.3483
pl	0.2559	0.3	0.2708	0.2552	0.678	0.2671
it	0.4175	0.4208	0.3729	0.3458	0.813	0.4223
fr	0.3557	0.4042	0.3568	0.3193	0.7915	0.3545
tr	0.3478	0.3917	0.3646	0.3482	0.8134	0.3486
hu	0.2406	0.2667	0.2708	0.251	0.6898	0.2385
et	0.3263	0.3667	0.3438	0.3281	0.6234	0.3596
de	0.2362	0.2458	0.1979	0.1906	0.6436	0.2549
es	0.358	0.4042	0.3521	0.3042	0.8356	0.3498
pt	0.5048	0.5708	0.4896	0.425	0.904	0.4807

**Table 2** Semantic expansion configuration

Lang	MAP	Precision@5	Precision@10	Precision@20	Avg. Recall	Avg. Precision @Recall
en	0.7351	0.7667	0.6875	0.5906	0.9803	0.6826
el	0.37	0.4292	0.3896	0.3448	0.8412	0.343
lv	0.3429	0.3917	0.35	0.3198	0.7059	0.3451
pl	0.2698	0.3417	0.3062	0.2708	0.7084	0.2692
it	0.3972	0.4458	0.3792	0.3323	0.8266	0.3675
fr	0.3587	0.4167	0.4027	0.3402	0.7961	0.3588
tr	0.3331	0.425	0.375	0.3398	0.8297	0.3412
hu	0.2167	0.2917	0.2792	0.2344	0.7152	0.2184
et	0.3177	0.4	0.3667	0.3438	0.6363	0.3394
de	0.2217	0.2792	0.25	0.2406	0.6409	0.2427
es	0.3708	0.4458	0.4167	0.3573	0.8518	0.3591
pt	0.4633	0.55	0.4729	0.4219	0.9099	0.4504

the semantic expansion configuration w.r.t. the base configuration (as proved by the increase of the Recall values). These additional documents include some relevant items presented in the higher part of the ranked result set (being the reason for the increase in Precision@10 and Precision@20) as well as a certain number of non-relevant documents presented in the lower part of the ranked result set (being the reason for a nonincreased MAP and Precision@Recall values). However, by taking into account the real usage of search engines, where the majority of search result click activity (89.8%) happens on the first page of search results (Spink et al. 2006) (i.e., meaning that users only consider the first 10–20 documents), we can say that the effectiveness obtained on the higher part of the ranked result constitutes the

**Table 3** Percentual gain

Lang	P@10 Base	P@10 SE	Gain (%)	Avg. Recall Base	Avg. Recall SE	Gain (%)
en	0.6896	0.6875	-0.30	0.9635	0.9803	1.71
el	0.3479	0.3896	11.99	0.8253	0.8412	1.89
lv	0.3187	0.35	9.82	0.703	0.7059	0.41
pl	0.2708	0.3062	13.07	0.678	0.7084	4.29
it	0.3729	0.3792	1.69	0.813	0.8266	1.65
fr	0.3568	0.4027	12.86	0.7915	0.7961	0.58
tr	0.3646	0.375	2.85	0.8134	0.8297	1.96
hu	0.2708	0.2792	3.10	0.6898	0.7152	3.55
et	0.3438	0.3667	6.66	0.6234	0.6363	2.03
de	0.1979	0.25	26.33	0.6436	0.6409	-0.42
es	0.3521	0.4167	18.35	0.8356	0.8518	1.90
pt	0.4896	0.4729	-3.41	0.904	0.9099	0.65

most relevant aspect for judging the quality of the presented results. Therefore, we can conclude that semantic expansion provided a valuable increase in the overall performance of the CLIR system.

## 7 Conclusions

In this chapter, we have presented an approach to collaborative management of multilingual ontologies. We have discussed the issues concerning the multilinguality in modeling tasks, and we have emphasized the collaborative issues. Moreover, we have described a collaborative modeling wiki-based tool that provides a set of features able to support the management of multilingual ontologies in a collaborative environment.

Finally, we have shown a possible exploitation of multilingual ontologies concerning their usage in a CLIR systems. There, the multilingual layer of the ontology has been used for enriching document representations at indexing time. The obtained results demonstrate that the usage of multilingual ontologies may lead to the improvement of CLIR system effectiveness.

## References

- M. Agosti, Di Nunzio, G. M., & Ferro, N. (2007). Scientific data of an evaluation campaign: Do we properly deal with them? In *Lecture notes in computer science* (Vol. 4730). Springer.
- Bo, J. D., Spyns, P., & Meersman, R. (2003). Creating a “dogmatic” multilingual ontology infrastructure to support a semantic portal. In *OTM Workshops* (pp. 253–266).



- Braschler, M., & Peters, C. (2003). Clef 2002 methodology and metrics. In *Lecture notes in computer science*, (Vol. 2785). Springer.
- Collier, N., Kawazoe, A., Jin, L., Shigematsu, M., Dien, D., Barrero, R., Takeuchi, K., & Kawtrakul, A. (2006). A multilingual ontology for infectious disease surveillance: Rationale, design and challenges. *Language Resources and Evaluation*, 40(3–4), 405–413.
- Declerck, T., Krieger, H.-U., Thomas, S. M., Buitelaar, P., O’Riain, S., Wunner, T., Maguet, G., McCrae, J., Spohr, D., & Montiel-Ponsoda, E. (2010). Ontology-based multilingual access to financial reports for sharing business knowledge across europe. In József Roóz & János Ivanyos (Hrsg.), *Internal financial control assessment applying multilingual ontology framework. MONTIFIC - ECQA Joint Conference, the Current Financial Crisis and Competences to Address Problems on the European Market*, Budapest, Hungary, 30 September–1 October 2010. nyomdájában, Budapest: Készült a HVG Press Kft (9/2010) [ISBN 978-963-08-0012-9]
- Di Francescomarino, C., Ghidini, C., & Rospocher, M. (2012). Evaluating wiki-enhanced ontology authoring. In *Proceedings of the 18th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2012)* (Vol. 7603, pp. 292–301).
- Dimitrova, V., Denaux, R., Hart, G., Dolbear, C., Holt, I., & Cohn, A. G. (2008). Involving domain experts in authoring owl ontologies. In *Proceedings of the 7th International Semantic Web Conference (ISWC 2008). Lecture notes in computer science* (Vol. 5318/2010, pp. 1–16). Berlin: Springer.
- dos Santos, C. T., Quaresma, P., & Vieira, R. (2008). A framework for multilingual ontology mapping. In *LREC*. European Language Resources Association.
- Dragoni, M., da Costa Pereira, C., & Tettamanzi, A. (2012). A conceptual representation of documents and queries for information retrieval systems by using light ontologies. *Expert Systems with Applications*, 39(12), 10376–10388.
- Elberichi, Z., Taibi, M., & Belaggoun, A. (2012). Multilingual medical documents classification based on mesh domain ontology. *CoRR*, abs/1206.4883.
- Espinoza, M., Gómez-Pérez, A., & Mena, E. (2008a). Enriching an ontology with multilingual information. In *Proceedings of the 5th European Semantic Web Conference (ESWC’08)* (pp. 333–347). Berlin: Springer.
- Espinoza, M., Gómez-Pérez, A., & Mena, E. (2008b). Enriching an ontology with multilingual information. In S. Bechhofer, M. Hauswirth, J. Hoffmann, & M. Koubarakis (Eds.), *ESWC. Lecture notes in computer science* (Vol. 5021, pp. 333–347). Berlin: Springer.
- Ferro, N., & Peters, C. (2010). Clef 2009 ad hoc track overview: TEL & persian tasks. In *Lecture notes in computer science* (Vol. 6241). Springer Berlin Heidelberg.
- Gennari, J., Musen, M., Ferguson, R., Grosso, W., Crubézy, M., Eriksson, H., Noy, N., & Tu, S. (2003). The evolution of protégé: An environment for knowledge-based systems development. *The International Journal of Human-Computer Studies*, 58(1), 89–123.
- Ghidini, C., Rospocher, M., & Serafini, L. (2012). Conceptual modeling in wikis: A reference architecture and a tool. In *eKNOW2012, Valencia, Spain* (pp. 128–135).
- Kaljurand, K., & Fuchs, N. E. (2007). Verbalizing owl in attempto controlled english. In *Proceedings of Third International Workshop on OWL: Experiences and Directions*, Innsbruck, Austria, 6th–7th June 2007 (Vol. 258).
- Kerremans, K., Temmerman, R., & Tummers, J. (2003). Representing multilingual and culture-specific knowledge in a vat regulatory ontology: Support from the termontology method. In *OTM Workshops* (pp. 662–674).
- Liu, O., & Ma, J. (2010). A multilingual ontology framework for r&d project management systems. *Expert Systems with Applications*, 37(6), 4626–4631.
- Nyulas, C., Tudorache, T., Tu, S. W., & Musen, M. A. (2012). Experiences with multilingual modeling in the development of the international classification of traditional medicine ontology. In P. Buitelaar, P. Cimiano, D. Lewis, J. Pustejovsky, & F. Sasaki (Eds.), *MSW. CEUR Workshop Proceedings* (Vol. 936), CEUR-WS.org.
- Palma, R., Corcho, O., Gómez-Pérez, A., & Haase, P. (2011). A holistic approach to collaborative ontology development based on change management. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(3), 299–314.

- Peters, C., Braschler, M., Nunzio, G. D., Ferro, N., Gonzalo, J., & Sanderson, M. (2008). From research to application in multilingual information access: The contribution of evaluation. In *LREC*. European Language Resources Association.
- Spink, A., Jansen, B., Blakely, C., & Koshman, S. (2006). A study of results overlap and uniqueness among major web search engines. *Information Processing & Management*, 42(5), 1379–1391.
- Spohr, D., Hollink, L., & Cimiano, P. (2011). A machine learning approach to multilingual and cross-lingual ontology matching. In L. Aroyo, C. Welty, H. Alani, J. Taylor, A. Bernstein, L. Kagal, N. F. Noy, & E. Blomqvist (Eds.), *International Semantic Web Conference (1). Lecture notes in computer science* (Vol. 7031, pp. 665–680). New York: Springer.
- Stamou, S., Nenadic, G., & Christodoulakis, D. (2004). Exploring balkanet shared ontology for multilingual conceptual indexing. In *LREC*. European Language Resources Association.
- Sure, Y., Erdmann, M., Angele, J., Staab, S., Studer, R., & Wenke, D. (2002). Ontoedit: Collaborative ontology development for the semantic web. In *Proceedings of the First International Semantic Web Conference on The Semantic Web, ISWC '02* (pp. 221–235). London: Springer.
- Tudorache, T., Falconer, S. M., Noy, N. F., Nyulas, C., Üstün, T. B., Storey, M.-A. D., & Musen, M. A. (2010). Ontology development for the masses: Creating icd-11 in webprotégé. In P. Cimiano & H. S. Pinto (Eds.), *Proceedings of the 17th International Conference on Knowledge Engineering and Management by the Masses , EKAW 2010*, Lisbon, Portugal, 11–15 October 2010 (Vol. 6317, pp. 74–89).
- Vouros, G. A., Eumeridou, E., Tselios, P., & Kotis, K. (2005). Multilingual, ontology-driven, content-based search and navigation of information items. *Applied Artificial Intelligence*, 19(7), 691–719.
- Wikimedia Foundation. (n.d.). Mediawiki. <http://www.mediawiki.org>.

# From RDF to Natural Language and Back

Daniel Gerber and Axel-Cyrille Ngonga Ngomo

**Abstract** Most knowledge sources on the Data Web were extracted from structured or semistructured data sources. Thus, they encompass solely a small fraction of the information available on the document-oriented Web. In this chapter, we present Bootstrapping Linked Data (BOA), a framework that aims to facilitate the extraction of Resource Description Framework (RDF) from text. The idea behind BOA is to extract natural language patterns that represent predicates found on the Data Web from unstructured data by using background knowledge from the Data Web. These patterns are then used to extract instance knowledge from unstructured data sources. This knowledge can finally be fed back into the Data Web. The approach followed by BOA is quasi-independent of the language in which the corpus is written. We demonstrate our approach by applying it to four different corpora and two different languages. We evaluate BOA on these data sets using DBpedia as background knowledge. Our results show that we can extract several thousand new facts in one iteration with high accuracy. Moreover, we provide the first multilingual repository of natural language representations (NLR) of predicates found on the Data Web. Finally, we present two applications of the natural language patterns generated by BOA, i.e., the fact validation framework DeFacto and the question answering engine Template - based SPARQL Learner (TBSL).

**Key Words** Fact validation • Natural language processing • Question answering • Relation extraction • Semantic Web

## 1 Introduction

The population of the Data Web has been mainly carried out by transforming semi-structured and structured data available on the Web into RDF. Yet, while these approaches have successfully generated the more than 30 billion triples currently available on the Data Web (Auer et al. 2011), they rely on background data that encompasses solely 15–20% (Gaag et al. 2009) of the information

---

D. Gerber (✉) • A.-C.N. Ngomo  
AKSW, Institut für Informatik, Universität Leipzig, Postfach 100920, 04009 Leipzig, Germany  
e-mail: [dgerber@informatik.uni-leipzig.de](mailto:dgerber@informatik.uni-leipzig.de); [ngonga@informatik.uni-leipzig.de](mailto:ngonga@informatik.uni-leipzig.de)

on the Web, as the rest of the information in the document-oriented Web is only available in unstructured form. Consequently, the data in the Data Web suffers from a lack of coverage and actuality that has been eradicated from the Web, by Web 2.0 and crowdsourcing approaches. For example, while the Wikipedia text fragment "...reputedly designed by Robert Mills, architect of the Washington Monument..." states that the triple `dbr:Washington_Monument dbo:architect dbr:Robert_Mills` holds, this triple is not included in DBpedia 3.7. In addition, being able to convert natural language to structured data makes manifold novel applications possible. For example, it allows mapping the string `born in` from questions such as `Which actors were born in Germany?` to the relation `dbo:birthPlace` and thus enables question answering based on SPARQL Protocol And RDF Query Language (SPARQL) as presented by Unger et al. (2012). Moreover, it becomes possible to check for the occurrence of RDF triples such as "`dbr:Washington_Monument dbo:architect dbr:Robert_Mills`" in text with the aim of validating them as carried out by the DeFacto framework (Lehmann et al. 2012).

In this chapter, we present the BOA framework,<sup>1</sup> which aims to address the challenge of extracting structured data as RDF from unstructured data. Unlike many approaches (e.g., Carlson et al. 2010) which start with their own ontologies and background knowledge as seeds, BOA makes use of the Data Web to retrieve high-confidence multilingual natural language patterns that express the predicates available in the Data Web. In contrast to its previous model (Gerber and Ngonga Ngomo 2011), BOA uses a supervised machine-learning approach trained on a set of manually annotated patterns to recognize high-confidence patterns. Based on these patterns, BOA can extract new instance knowledge (i.e., both new entities and relations between these new entities) from the Human Web with high accuracy. Our approach is completely agnostic of the knowledge base upon which it is deployed. It can thus be used on the whole Data Web. In addition, our extension of BOA implements generic pattern extraction algorithms that can be used to retrieve knowledge from sources written in different languages. Consequently, it can also be used on the whole Human Web.

The main contributions of this chapter are as follows: (1) We present the novel approach implemented by the BOA framework and apply it to corpora written in English and in German. (2) We provide a multilingual library of natural language representations (NLRs) of predicates found on the Data Web (especially in DBpedia). (3) We present a set of features that can be used to distinguish high-quality from poor natural language patterns for Data Web predicates. (4) We evaluate our machine-learning approach and the BOA framework on four text data sets against DBpedia and show that we can achieve a high-accuracy extraction in both languages. (5) We present how this library can be applied for fact validation and question answering. The rest of this chapter is structured as follows: In Sect. 2,

---

<sup>1</sup>A demo of the framework can be found at <http://boa.aksw.org>. The code of the project is at <http://boa.googlecode.com>.

we give an overview of previous work that is related to our approach. Thereafter, in Sect. 3, we present our bootstrapping framework and several insights that led to the approach currently implemented therein. In Sect. 4, we evaluate our approach on two different data sets and show its robustness and accuracy. DeFacto and TBSL, two applications enabled by BOA, are presented in Sect. 5. Finally, we sum up our results and conclude. This chapter is an extended version of Gerber and Ngonga Ngomo (2012), and the application sections are based on Unger et al. (2012) and Lehmann et al. (2012).

## 2 Related Work

BOA is related to a large number of disciplines due to the different areas of knowledge from which it borrows methods. Like information extraction approaches, BOA aims to detect entities in text. Three main categories of natural language processing (NLP) tools play a central role during the extraction of knowledge from text: keyphrase extraction (KE; Kim et al. 2010), named-entity recognition (NER; Finkel and Manning 2010), and relation extraction (RE; Mintz et al. 2009). While these three categories of approaches are suitable for the extraction of facts from NL, the use of the Data Web as source for background knowledge for fact extraction is still in its infancy. Mintz et al. (2009) coined the term “distant supervision” to describe this paradigm but developed an approach that led to extractors with a low precision (approx. 67.6%). Services such as Alchemy,<sup>2</sup> FOX (Ngonga Ngomo et al. 2011), and Spotlight (Mendes et al. 2011) reach better precision scores and allow to extract entities and relations from text. Yet, they do not rely on the Data Web as training data and are thus restricted with respect to the number of relations they can detect. The problem of extracting knowledge from the Web at large scale, which is most closely related to this chapter, has been the object of recent research, especially in the projects ReadTheWeb and PROSPERA. The aim of the ReadTheWeb project<sup>3</sup> (Carlson et al. 2010) is to create the never-ending language learner (NELL) that can read webpages. To achieve this goal, NELL is fed with the ClueWeb09<sup>4</sup> data set and an initial ontology. In each iteration, NELL uses the available instance knowledge to retrieve new instances of existing categories and relations between known instances by using pattern harvesting. The approach followed by PROSPERA (Nakashole et al. 2011) is similar to that of NELL but relies on the iterative harvesting of n-gram-itemset patterns that allow generalizing NL patterns found in text. Another closely related approach is presented by Demey et al. (this volume). They use fact based modeling to verbalize n-ary relations.

Our approach goes beyond the state of the art in two key aspects. First, it is the first approach to extract multilingual natural language patterns from the Data

---

<sup>2</sup><http://www.alchemyapi.com>.

<sup>3</sup><http://rtw.ml.cmu.edu>.

<sup>4</sup><http://lemurproject.org/clueweb09>.

Web. In addition, it makes use of the Data Web as background knowledge, while the approaches ReadTheWeb and PROSPERA rely on their own ontology for this purpose. Moreover, BOA can generate RDF and can thus be used to populate a knowledge base that can be readily made available for querying via SPARQL, integrating, and linking. Finally, our experiments show that our approach can extract a large number of statements (like PROSPERA and Mintz et al. 2009) with a high precision (like ReadTheWeb).

### 3 From RDF to Natural Language and Back

The idea behind the BOA framework (“BOotstrapping Linked Data”) is to facilitate the iterative extraction of RDF data from the Human Web. An overview of the workflow implemented by BOA is given in Fig. 1. The input for the BOA framework consists of a knowledge base and a text corpus. For each predicate  $p$  found in the input knowledge base, BOA carries out a sentence-level statistical analysis of the co-occurrence of pairs of labels of resources that are linked via  $p$ . BOA uses a supervised machine-learning approach to compute the score of patterns extracted from a given corpus. In a final step, our framework uses the best-scoring patterns for each relation to generate RDF data. This data and the already available background knowledge can now be used for a further iteration of the approach. In this chapter, we will describe each of the core steps of BOA in detail and focus especially on the pattern and feature extraction, as well as on the score function approaches underlying BOA.

#### 3.1 Pattern Extraction

Let  $\mathcal{K}$  be the knowledge base that is used as background knowledge. The first and optional step of the pattern extraction is the computation of surface forms  $\mathcal{S}_r$  for the

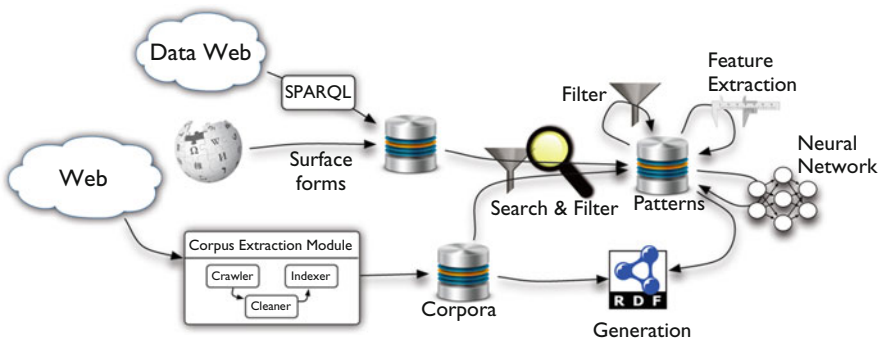


Fig. 1 Overview of the BOA approach

```

1 dbr:Empire_State_Building dbo:architect dbr:Shreve,_Lamb_and_Harmon
2 dbr:Empire_State_Building rdfs:label 'Empire State Building'@en
3 dbr:Shreve,_Lamb_and_Harmon rdfs:label 'Shreve, Lamb and Harmon'@en
    
```

**Listing 1** RDF snippet used for pattern search

**Table 1** Example sentences for pattern search

Sentence with $\lambda(s)$ before $\lambda(o)$	Sentence with $\lambda(o)$ before $\lambda(s)$
“... <b>Shreve, Lamb, and Harmon</b> <i>also designed the</i> <b>Empire State Building</b> ”	“The <b>Empire State Building</b> <i>was designed by</i> <b>William F. Lamb</b> ...”

subject and objects of a relation  $p$  for which patterns are to be extracted. To extract surface forms for resources  $r$  in  $\mathcal{K}$ , we use Wikipedia’s redirect and disambiguation pages as described by Mendes et al. (2011). Overall, the average number of surface forms per resource was 1.66 for German and 2.36 for English. The pattern search is carried out independently for each predicate. Let  $p \in \mathfrak{P}$  be an RDF predicate whose NLRs are to be detected, where  $\mathfrak{P}$  is the set of all RDF predicates under consideration. We use the symbol “ $\in$ ” between triples and knowledge bases to signify that a triple can be found in a knowledge base. The starting point for the pattern search for  $p$  is the set of pairs  $\mathcal{I}(p) = \{(s, o) : (s p o) \in \mathcal{K}\}$  that instantiate  $p$ . In the following, we use  $\mu(x)$  to signify  $x$ ’s URI. For each  $(s, o) \in \mathcal{I}(p)$ , we retrieve all sentences from the input corpus which contains at least one of all possible combinations of a subject label  $l_s \in \mathcal{S}_s$  and an object label  $l_o \in \mathcal{S}_o$ , respectively. For each found sentence, we delete all tokens that are not found between  $l_s$  and  $l_o$  in  $\sigma$ . To facilitate readability, the labels are then replaced with the placeholders D for  $l_s$  and R for  $l_o$ . We call the resulting string a *NLR* of  $p$  and denote it with  $\theta$ . Each distinctly extracted  $\theta$  is used to create a new instance of a BOA pattern.

**Definition 1 (BOA Pattern).** A BOA pattern is a pair  $\mathcal{P} = (\mu(p), \theta)$ , where  $\mu(p)$  is  $p$ ’s URI and  $\theta$  is a NLR of  $p$ .

**Definition 2 (BOA Pattern Mapping).** A BOA pattern mapping is a function  $\mathcal{M}$  such that  $\mathcal{M}(p) = \mathfrak{S}$ , where  $\mathfrak{S}$  is the set of NLRs for  $p$ .

For example, consider the RDF snippet from Listing 1 derived from DBpedia. Querying the index of an underlying corpus for sentences which contain both entity labels returns the sentences depicted in Table 1 among others. We can replace “Empire State Building” with D, because it is a label of the subject of the `:architect` triple, as well as replace “Shreve, Lamb, and Harmon” and “William F. Lamb” (a surface form  $l_r \in \mathcal{S}_r$ ) with R because it is one label of the object of the same triple. These substitutions lead to the BOA patterns `(:architect, “D was designed by R”)` and `(:architect, “R also designed the D”)`. For the sake of brevity and in the case of unambiguity, we also call  $\theta$  “pattern.” Patterns are only considered for storage and further computation if they withstand a first filtering process. For example, they must contain more than one non-stop word, have a token count between certain thresholds, and may not begin with a conjunction. In

addition to  $\mathcal{M}(\mathfrak{p})$  for each  $\mathfrak{p}$ , we compute the number  $f(\mathcal{P}, s, o)$  of occurrences of  $\mathcal{P}$  for each element  $(s, o)$  of  $\mathcal{I}(\mathfrak{p})$  and the ID of the sentences in which  $\mathcal{P}$  was found. Based on this data, we can compute (1) the total number of occurrences of a BOA pattern  $\mathcal{P}$ , dubbed  $f(\mathcal{P})$ ; (2) the number of sentences that led to  $\theta$  and that contained  $\lambda(s)$  and  $\lambda(o)$  with  $(s, o) \in \mathcal{I}(\mathfrak{p})$ , which we denote  $l(s, o, \theta, \mathfrak{p})$ ; and (3)  $\mathcal{I}(\mathfrak{p}, \theta)$  is the subset of  $\mathcal{I}(\mathfrak{p})$  which contains only pairs  $(s, o)$  that led to  $\theta$ . Thereafter, we apply a second filtering process, where we eliminate the long tail of patterns which have only been learned by a single pair  $(s, o)$ . We denote the set of predicates, such that the pattern  $\theta \in \mathcal{M}(\mathfrak{p})$  as  $\mathfrak{M}(\theta)$ . Note that pattern mappings for different predicates can contain the same pattern.

### 3.2 Feature Extraction

Feature extraction is applied on all patterns which overcome both filtering processes.

Note that although BOA is designed to work independently of the language of the underlying corpus, it can be tailored toward a given language. For example, the ReVerb and IICM feature exploit knowledge that is specific to English. The first three features BOA relies upon are the support, specificity, and typicality as described by Gerber and Ngonga Ngomo (2011). In addition, we rely on the three supplementary features dubbed IICM, ReVerb, and tf-idf. The **Intrinsic Information Content Metric (IICM)** captures the semantic relatedness between a pattern’s NLR and the property it expresses. This similarity measure was introduced in Seco et al. (2004) and is based on the Jiang-Conrath similarity measure (Jiang and Conrath 1997). We apply this measure to each BOA pattern mapping independently. First, we retrieve all synsets for each token of the pattern mappings associated *rdfs:label* from WordNet. If no such synsets are found, we use the tokens of the *rdfs:label* of  $\mathcal{M}(\mathfrak{p})$ . We then apply the IICM measure pairwise to these tokens and the tokens derived from one  $\mathcal{M}(\mathfrak{p})$  assigned pattern’s NLR. The IICM score for one pattern is then the maximal value of the similarity values of all pairs. **ReVerb** has been introduced by Fader et al. (2011) and distinguishes good from bad relation phrases by measuring how well they abide to a predefined part-of-speech-based regular expression. Since the input of ReVerb is a POS-tagged sentence, but a pattern is only a substring of a sentence, we use all sentences we found the pattern in (see Sect. 3.1) as ReVerb’s input. For all of ReVerb’s extracted relations of a particular sentence, we check if it matches the pattern in question and use ReVerb’s trained logistic regression classifier to assign a confidence score to this extraction. Note that BOA focuses on the relation between two given resources and discards all other extractions, since those are not mappable to the background knowledge. Finally, we calculate a pattern’s ReVerb feature as the average of all scored extractions. The **tf-idf** features are an adaption of the tf-idf score used in information retrieval and text mining. The intuition behind this feature is to distinguish relevant from irrelevant patterns for a given pattern mapping  $\mathcal{M}(\mathfrak{p})$ . In the BOA case, a document is considered to be all tokens of all patterns (without stop words and the placeholders



“D” and “R”) of one pattern mapping. In other words, the total number of documents is equal to the number of pattern mappings with patterns. We then calculate the features  $idf(p)$  and  $tf(p)$  for each token of the patterns NLR as follows:

$$idf(p) = \sum_{t \in \mathcal{T}(p)} \log \left( \frac{|\mathcal{M}(p)|}{df(t) + 1} \right) + 1 \quad tf(p) = \sum_{t \in \mathcal{T}(p)} \sqrt{f(t)}$$

where  $df(t)$  is the document frequency of  $t$ ,  $f(t)$  the term frequency of  $t$ , and  $\mathcal{T}(p)$  the set of tokens for a pattern  $p$ .

### 3.3 Scoring Approach

Given the number of features that characterize the input data, devising a simple scoring function transforms into a very demanding task. In this work, we address the problem of computing a score for each BOA pattern by using feedforward neural networks. The input layer of our network consists of as many neurons as features for patterns exist, while the output neuron consists of exactly one neuron whose activation was used as score. We used the sigmoid function as transfer function. For each data set, we trained the neural network by using manually annotated patterns (200 in our experiments). The patterns were extracted from the set of all patterns generated by BOA by first randomly sampling the same number of patterns for each predicate (seven in our experiments) and selecting a subset of these patterns for annotation.

### 3.4 RDF Generation

The generation of RDF out of the knowledge acquired by BOA is the final step of the extraction process and is carried out as follows: For each pattern  $\theta$  and each predicate  $p$ , we first use the Lucene index to retrieve sentences that contain  $\theta$  stripped from the placeholders “D” and “R.” These sentences are subsequently processed by an NER tool that is able to detect entities that are of the `rdfs:domain` and `rdfs:range` of  $p$ . Thereafter, the first named entities within a limited distance on the left and right of  $\theta$  which abide by the domain and range information of  $p$  are selected as labels for subject and object of  $p$ . Each of the extracted labels is then fed into the URI retrieval and disambiguation service implemented by the FOX framework. If this service returns a URI, then we use it for the label detected by BOA. Else, we create a new BOA URI. By applying our approach, we were able to extract the triples shown in Listing 2 from the text fragment “...reputedly designed by Robert Mills, architect of the Washington Monument.”

```

1 dbr:Washington_Monument dbo:architect dbr:Robert_Mills .
2 dbr:Washington_Monument rdf:type dbo:Building .
3 dbr:Washington_Monument rdfs:label "Washington Monument"@en .
4 dbr:Robert_Mills rdf:type dbo:Architect .
5 dbr:Robert_Mills rdfs:label "Robert Mills"@en .

```

**Listing 2** RDF snippet generated by BOA

Note that `dbr:Washington_Monument dbo:architect dbr:Robert_Mills` is not included in DBpedia but explicitly stated in Wikipedia.

## 4 Evaluation

The aim of our evaluation was threefold. First, we aimed at testing how well BOA performs on different languages. To achieve this goal, we applied BOA to German and English corpora. Our second goal was to determine the accuracy of BOA's extraction. For this purpose, we sampled 100 triples from the data extracted by BOA from each corpus and had two annotators measure the precision of these samples manually. Finally, we wanted to compute the amount of (new) knowledge that can be extracted by BOA. For this purpose, we compute the number of new triples that we were able to extract. We excluded temporal properties from the evaluation as BOA does not yet distinguish between different time expressions and conjugations. We evaluated our approach on the four corpora described in Table 2.

### 4.1 Score Function

We began the evaluation by annotating 200 patterns per corpus by hand. Each training data set was annotated independently by the authors, who agreed on the annotations in 89% of the cases. The annotations upon which the authors disagreed were resolved by both authors. High-quality patterns were assigned a score of 1; else they were assigned a 0. We then trained four different neural networks (one for each data set) to distinguish between the high-precision and poor patterns. In our experiments, we varied the size of the hidden layer between one and three times the size of the input layer. In addition, we varied the error rate to which they were trained. The maximal number of training epochs was set to 10,000. The accuracy of the networks was measured by using a tenfold cross validation. Patterns whose score was above 0.5 were considered to be good patterns, while all others were considered to be poor. The best neural network was set to be the smallest network that reaches the maximal accuracy. The networks trained to achieve an error rate of maximally 5%, and having a greater or equal number of hidden layer neurons than features, performed best in our experiments.

**Table 2** Statistical overview of German and English text corpus

Corpus	Sentences	Tokens	Unique tokens	Tokens per sentence
en-wiki	58.0M	1,240.6M	6.8M	21.4
en-news	214.3M	4,745.1M	17.6M	22.1
de-wiki	24.6M	428.4M	6.7M	17.4
de-news	112.8M	2,062.1M	18.0M	18.3

**Table 3** Results of one iteration of the BOA framework

	en-wiki	de-wiki	en-news	de-news
Number of pattern mappings	125	44	66	19
Number of patterns	9,551	586	7,366	109
Number of new triples	78,944	22,883	10,138	883
Number of known triples	1,829	798	655	42
Number of found triples	80,773	3,081	10,793	925
Precision top 100 triples (%)	92	70	91	74

## 4.2 Multilinguality

Enabling BOA to process languages other than English requires solely the alteration of the NER tools and POS taggers. As the results on German show, languages with a wide range of morphosyntactical variations demand the analysis of considerably larger corpora to enable the detection of meaningful patterns. For example, while we trained the neural network by using the same number of patterns, we were not able to detect any triples with a score above 0.5 when using the German Wikipedia and DBpedia. Yet, when using a larger German news corpus data set, we were able to detect new patterns with an acceptable precision (see subsequent section).

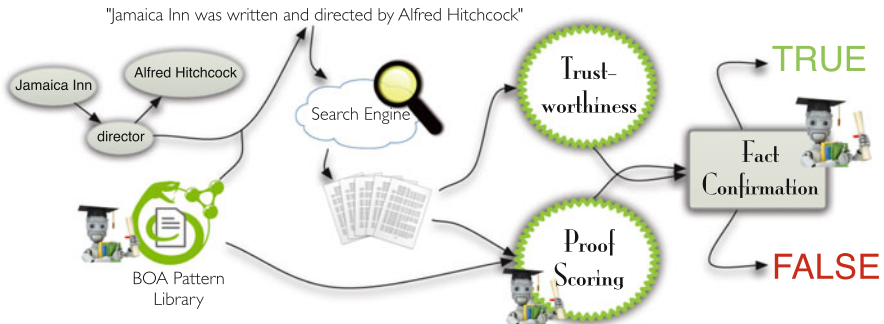
## 4.3 Accuracy

The results of our experiments on accuracy are shown in Table 3. We measured the precision of the extraction carried out by BOA as well as the number of new triples that we were able to extract in one iteration. For the top 100 scored triples, we achieved a precision over 90% overall on the English data sets. This value is comparable to that achieved by the previous versions of BOA (Gerber and Ngonga Ngomo 2011). Yet, the addition of surface forms for the extraction yields the advantage of achieving a considerably higher recall both with respect to the number of patterns extracted as well as with respect to the total number of triples extracted. For example, when using the English Wikipedia, we can extract more than twice the amount of triples. The same holds for the number of patterns and pattern mappings as shown in Table 3.

```

1 Chinese spokenIn Malaysia .
2 Chinese spokenIn China .
3 Weitnau administrativeDistrict boa:Oberallgäu .
4 Memmingerberg administrativeDistrict boa:Unterallgäu .
5 ESV_Blau-Rot_Bonn ground Bonn .
6 TG_Würzburg ground Würzburg .
7 Intel_Corporation subsidiary McAfee .
8 Iomega subsidiary ExcelStor_Technology .
    
```

**Listing 3** RDF extracted by BOA. If not stated otherwise, all instances and properties use the DBpedia namespace



**Fig. 2** Overview of the DeFacto framework

An excerpt of the new knowledge extracted by BOA is shown in Listing 3. Note that the triple `Iomega subsidiary ExcelStor_Technology` is wrong. Although Iomega planned to buy ExcelStor, the deal was never concluded. Our approach finds the right patterns in the sentences describing the deal and thus extract this triple.

## 5 Applications

In this chapter, we present two applications in which the BOA pattern library has been applied successfully. The first application is DeFacto, a framework to evaluate the validity of RDF triples, and the second application, TBSL, is a question answering tool for RDF knowledge bases.

### 5.1 DeFacto

The DeFacto system consists of the components depicted in Fig. 2. The system takes an RDF triple as input and returns a confidence value for this triple as well as possible evidence for the fact. The evidence consists of a set of webpages, textual

excerpts from those pages, and meta-information on the pages. The text excerpts and the associated metainformation allow the user to quickly get an overview over possible credible sources for the input statement: Instead of having to use search engines, browsing several webpages, and looking for relevant pieces of information, the user can more efficiently review the presented information. Moreover, the system uses techniques which are adapted specifically for fact validation instead of only having to rely on generic information retrieval techniques of search engines.

The first task of the DeFacto system is to retrieve webpages which are relevant for the given task. The retrieval is carried out by issuing several queries to a regular search engine. These queries are computed by verbalizing the RDF triple using natural language patterns extracted by the BOA framework. As a next step, the highest ranked webpages for each query are retrieved. Those webpages are candidates for being sources for the input fact. Both the search engine queries as well as the retrieval of webpages are executed in parallel to keep the response time for users within a reasonable limit. Once a webpage has been retrieved, we extract plain text by removing HTML markup. We can then apply our fact confirmation approach on this text. In essence, the algorithm decides whether the webpage contains a natural language formulation of the input fact. This step distinguishes DeFacto from information retrieval methods. If no webpage confirms a fact according to DeFacto, then the system falls back on lightweight NLP techniques and computes whether the webpage does at least provide useful evidence. In addition to fact confirmation, the system computes different indicators for the trustworthiness of a webpage as presented by Nakamura et al. (2007). These indicators are of central importance, because a single trustworthy webpage confirming a fact may be a more useful source than several webpages with low trustworthiness. In addition to finding and displaying useful sources, DeFacto also outputs a general confidence value for the input fact. This confidence value ranges between  $[0, 1]$  and serves as an indicator for the user: Higher values indicate that the found sources appear to confirm the fact and can be trusted. Low values mean that not much evidence for the fact could be found on the Web and that the websites that do confirm the fact (if such exist) only display low trustworthiness. A prototype implementing the above steps is available at <http://defacto.aksw.org>. The generated provenance output, we use the PROV Ontology,<sup>5</sup> can also be saved directly as RDF. The source code of both, the DeFacto algorithms and user interface, is openly available.<sup>6</sup>

## 5.2 Evaluation

Our main objective in the evaluation was to find out whether DeFacto can effectively distinguish between true and false input facts. In the following, we describe how we

---

<sup>5</sup><http://www.w3.org/2011/prov/>.

<sup>6</sup><https://github.com/AKSW/DeFacto>.

trained DeFacto using DBpedia, which experiments we used, and then discuss the results of those experiments.

We focus our tests on the top 60 most frequently used properties in DBpedia. The system can easily be extended to cover more properties by extending the training set of BOA to those properties. Note that DeFacto itself is also not limited to DBpedia, i.e., while all of its components are trained on DBpedia, the algorithms can be applied to arbitrary URIs.

For training a supervised machine-learning approach, positive and negative examples are required. We use facts contained in DBpedia as positive examples, which are chosen randomly for each property. We obtain 600 statements this way and verified them manually. It turned out that some of the obtained triples were incorrectly extracted, e.g., obviously violated domain and range restrictions, or could not be confirmed by an intensive search on the Web within 10 min. Overall, 473 out of 600 checked triples were facts which we subsequently used as positive examples.

The generation of negative examples is more involved than the generation of positive examples. In order to effectively train DeFacto, we considered it essential that many of the negative examples are similar to true statements. In particular, most statements should be meaningful subject-predicate-object phrases. For this reason, we derive the negative examples from positive examples by modifying them but following domain and range information. Assume the input triple  $(s, p, o)$  in a knowledge base  $\kappa$  is given and let  $dom$  and  $ran$  be functions returning the domain and range of a property. We used the following methods to generate the negative example sets dubbed *domain*, *range*, *domain-range*, *property*, *random*, and *20%mix* (in that order):

1. A triple  $(s', p, o)$  is generated where  $s'$  is an instance of  $dom(p)$ , the triple  $(s', p, o)$  is not contained in  $\kappa$ , and  $s'$  is randomly selected from all resources which satisfy the previous requirements.
2. A triple  $(s, p, o')$  is generated analogously by taking  $ran(p)$  into account.
3. A triple  $(s', p, o')$  is generated analogously by taking both  $dom(p)$  and  $ran(p)$  into account.
4. A triple  $(s, p', o)$  is generated in which  $p'$  is randomly selected from our previously defined list of 60 properties, and  $(s, p', o)$  is not contained in  $\kappa$ .
5. A triple  $(s', p', o')$  is generated where  $s'$  and  $o'$  are randomly selected resources,  $p'$  is a randomly selected property from our defined list of 60 properties and  $(s', p', o')$  is not contained in  $\kappa$ .
6. Twenty percent of each of the above-created negative training sets were randomly selected to create a heterogeneous test set.

We performed tenfold cross validations for our experiments. In each experiment, we used our created positive examples but varied the negative example sets described above to see how changes influence the overall behavior of DeFacto.

The results of our experiments are shown in Tables 2, 3, and 4. J48 decision trees show the most promising results. Given the challenging tasks,  $F$ -measures up to 78.8% for the combined negative example set appear to be very positive

**Table 4** Classification results for linear regression, SVM, and J48 decision trees

Classifier	P	R	F <sub>1</sub>	AUC	RMSE	P	R	F <sub>1</sub>	AUC	RMSE
	domain					range				
LR	0.799	0.753	0.743	0.83	0.4151	0.881	0.86	0.859	0.844	0.3454
SVM	0.811	0.788	0.784	0.788	0.4609	0.884	0.867	<b>0.865</b>	0.866	0.3409
J48	0.835	0.827	<b>0.826</b>	0.819	0.3719	0.869	0.862	0.861	0.908	0.3194
	Domain–range					Property				
LR	0.871	0.85	0.848	0.86	0.3495	0.822	0.818	0.818	0.838	0.3792
SVM	0.88	0.863	0.861	0.855	0.3434	0.819	0.816	0.816	0.825	0.3813
J48	0.884	0.871	<b>0.87</b>	0.901	0.3197	0.834	0.832	<b>0.832</b>	0.828	0.3753
	Combined negative examples					Random 20% mix				
LR	0.855	0.854	0.854	0.908	0.3417	0.665	0.645	0.634	0.785	0.4516
SVM	0.855	0.854	0.854	0.906	0.3462	0.734	0.729	0.728	0.768	0.4524
J48	0.876	0.876	<b>0.876</b>	0.904	0.3226	0.8	0.79	<b>0.788</b>	0.782	0.405

indicators that DeFacto can be used to effectively distinguish between true and false statements, which was our primary evaluation objective. In general, DeFacto also appears to be stable against the various negative example sets. In particular, the algorithms with overall positive results also seem less affected by the different variations. When observing single runs of DeFacto manually, it turned out that our method of generating positive examples is particularly challenging for DeFacto: For many of the facts in DBpedia, only few sources exist in the Web. In general, DeFacto performs better when the subject and object of the input triple are popular on the Web, i.e., there are several webpages describing them. In this aspect, we believe our training set is indeed challenging upon manual observation.

### 5.3 SPARQL Template-Based Question Answering

A second domain of application for the BOA pattern library is question answering. The basic intuition behind this application is that we can use the BOA patterns to detect expressions in questions which correspond to known relations from the Data Web. We implemented this approach in the template-based question answering system (TBSL), whose overview is given in Fig. 3. The input question, formulated by the user in natural language, is first processed by a POS tagger. On the basis of the POS tags, lexical entries are created using a set of heuristics. These lexical entries, together with predefined domain-independent lexical entries, are used for parsing, which leads to a semantic representation of the natural language query, which is then converted into a SPARQL query template. The query templates contain *slots*, which are missing elements of the query that have to be filled with URIs. In order to fill them, our approach first generates natural language expressions for possible slot fillers from the user question using WordNet expansion. In a next step, entity identification approaches are used to obtain URIs for those natural

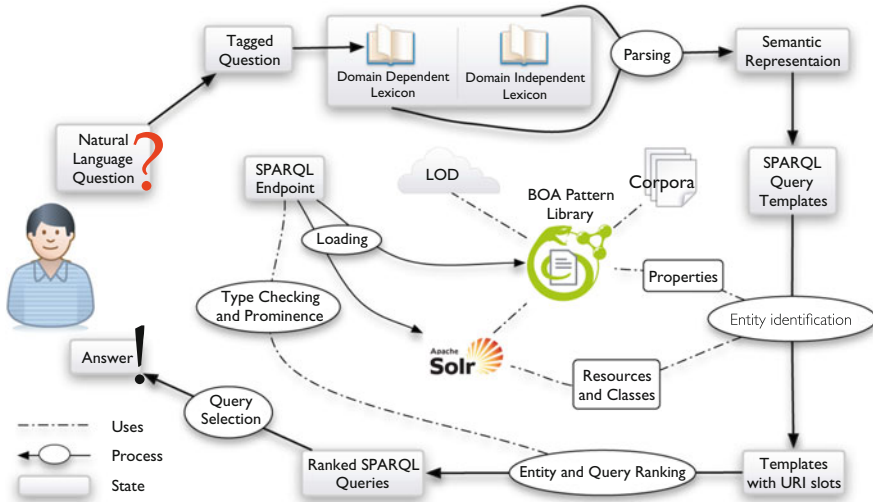


Fig. 3 Overview of the TBSL question answering pipeline

language expressions. These approaches rely both on string similarity as well as on natural language patterns which are compiled from existing structured data in the Linked Data Cloud and text documents. This yields a range of different query candidates as potential translations of the input question. It is therefore important to rank those query candidates. To do this, we combine string similarity values, prominence values, and schema conformance checks into a score value. The highest ranked queries are then tested against the underlying triple store, and the best answer is returned to the user.

The evaluation of the approach was based on the QALD-1<sup>7</sup> benchmark on DBpedia (Lehmann et al. 2013). It comprises two sets of 50 questions over DBpedia, annotated with SPARQL queries and answers. We only considered the questions from the test set and evaluated them w.r.t. precision and recall. The results reported are based on natural language questions tagged with ideal (manual annotation) part-of-speech information.

## 5.4 Evaluation Results

Of the 50 training questions provided by the QALD-1 benchmark, 11 questions rely on namespaces which we did not incorporate for predicate detection: FOAF<sup>8</sup>

<sup>7</sup><http://www.sc.cit-ec.uni-bielefeld.de/qald>.

<sup>8</sup><http://www.foaf-project.org/>.



```
1 Who was Tom Hanks married to?  
2 Which actors were born in Germany?  
3 Which presidents were born in 1945?  
4 Who wrote the book The pillars of the Earth?
```

**Listing 4** QALD queries answered with the help of the BOA pattern library

and YAGO.<sup>9</sup> Especially the latter poses a challenge, as YAGO categories tend to be very specific and complex. We did not consider these questions; thus, only 39 questions are processed by our approach. Of these 39 questions, 5 questions cannot be parsed due to unknown syntactic constructions or uncovered domain-independent expressions. This mainly concerns the noun phrase conjunction *as well as* and ordinals (*the 5th*, *the first*). These constructions will be added in the future; the only reason they were not implemented yet is that they require significant additional effort when specifying their compositional semantics.

Of the remaining 34 questions, 19 are answered exactly as required by the benchmark (i.e., with precision and recall 1.0), and another two are answered almost correctly (with precision and recall > 0.8). Listing 4 shows the four questions that could only be answered with the help of the BOA pattern library, thus leading to a 19% (4 of 21) increase in answered questions. The mean of all precision scores is therefore 0.61, and the mean of all recall scores is 0.63, leading to an *F*-measure  $[(2 \times \text{precision} \times \text{recall}) / (\text{precision} + \text{recall})]$  of 0.62. These results are comparable with those of systems such as FREyA and PowerAqua. The key advantage of our system is that the semantic structure of the natural language input is faithfully captured; thus, complex questions containing quantifiers, comparatives, and superlatives pose no problem, unlike in PowerAqua. Moreover, our system does not need any user feedback, as FREyA does.

## 6 Conclusion and Future Work

In this chapter, we presented BOA, a framework for the extraction of RDF from unstructured data. We presented the components of the BOA framework and applied it to English and German corpora. We showed in all cases that we can extract RDF from the data at hand with high precision. The precision of the extraction on German was lower than that on English because of the rich morphology and syntax of the German language as well as the limited availability of training data. Overall, the new version of BOA achieves a significantly higher recall by using surface forms to retrieve entities. We also showed that the BOA pattern library is beneficial for a variety of other use cases. We presented DeFacto, a framework for

---

<sup>9</sup><http://www.mpi-inf.mpg.de/yago-naga/yago/>.

fact validation, which verbalizes RDF triples to search for evidence on the Web. Additionally, we evaluated our approach on TBSL, a question answering system, which detects occurrences of formal relations in text with the help of the BOA pattern library. In future work, we want to implement pattern generalization in BOA to further increase recall. Moreover, we aim to extend our approach by including an analysis of dependency parse graphs. Additionally, we plan to use the BOA patterns to map natural language text to an ontology as presented by Bond et al. (this volume) and Unger et al. (2013).

## References

- Auer, S., Lehmann, J., & Ngomo, A.-C. N. (2011). Introduction to linked data and its lifecycle on the web. In *Reasoning Web* (pp. 1–75). Berlin: Springer.
- Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka, E., Jr., & Mitchell, T. M. (2010). Toward an architecture for never-ending language learning. In *AAAI*.
- Fader, A., Soderland, S., & Etzioni, O. (2011). Identifying relations for open information extraction. In *EMNLP* (pp. 1535–1545). Morristown: ACL.
- Finkel, J. R., & Manning, C. (2010). Hierarchical joint learning: Improving joint parsing and named entity recognition with non-jointly labeled data. In *ACL*.
- Gaag, A., Kohn, A., & Lindemann, U. (2009). Function-based solution retrieval and semantic search in mechanical engineering. In *IDEC '09* (pp. 147–158).
- Gerber, D., & Ngonga Ngomo, A.-C. (2011). Bootstrapping the linked data web. In *1st Workshop on Web Scale Knowledge Extraction*.
- Gerber, D., & Ngonga Ngomo, A.-C. (2012). Extracting multilingual natural-language patterns for RDF predicates. In *Proceedings of EKAW*.
- Jiang, J. J., & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *International Conference Research on Computational Linguistics (ROCLING X)* (pp. 9008+).
- Kim, S. N., Medelyan, O., Kan, M.-Y., & Baldwin, T. (2010). Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles. In *SemEval '10*.
- Lehmann, J., Gerber, D., Morsey, M., & Ngonga Ngomo, A.-C. (2012). DeFacto - Deep fact validation. In *11th International Semantic Web Conference*.
- Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., et al. (2013). DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web Journal* (in press).
- Mendes, P. N., Jakob, M., García-Silva, A., & Bizer, C. (2011). DBpedia spotlight: Shedding light on the Web of documents. In *Proceedings of I-SEMANTICS 2011*.
- Mintz, M., Bills, S., Snow, R., & Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *ACL* (pp. 1003–1011).
- Nakamura, S., Konishi, S., Jatowt, A., Ohshima, H., Kondo, H., Tezuka, T., et al. (2007). Trustworthiness analysis of web search results. In *ECDL* (Vol. 4675, pp. 38–49).
- Nakashole, N., Theobald, M., & Weikum, G. (2011). Scalable knowledge harvesting with high precision and high recall. In *WSDM* (pp. 227–236).
- Ngonga Ngomo, A.-C., Heino, N., Lyko, K., Speck, R., & Kaltenböck, M. (2011). SCMS - Semantifying content management systems. In *ISWC*.
- Seco, N., Veale, T., & Hayes, J. (2004). An intrinsic information content metric for semantic similarity in WordNet. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)* (Vol. 4, pp. 1089–1090).
- Unger, C., Bühmann, L., Lehmann, J., Ngonga Ngomo, A.-C., Gerber, D., & Cimiano, P. (2012). SPARQL template based question answering. In *Proceedings of ISWC*.

Unger, C., Mccrae, J., Walter, S., Winter, S., & Cimiano, P. (2013). A lemon lexicon for DBpedia. In *Proceedings of 1st International Workshop on NLP and DBpedia*, October 21–25, Sydney, Australia. NLP & DBpedia 2013 (Vol. 1064). Sydney, Australia: CEUR Workshop Proceedings.

# Multilingual Natural Language Interaction with Semantic Web Knowledge Bases and Linked Open Data

Mariana Damova, Dana Dannélls, Ramona Enache, Maria Mateva,  
and Aarne Ranta

**Abstract** This chapter presents a novel approach to Semantic Web technologies with the cultural heritage domain as a use case. Semantic Web technologies offer the technological backbone to meet the requirement of integrating heterogeneous data, but they are still more adapted to be consumed by computers rather than by humans. This chapter describes a method that allows interaction with semantic knowledge bases in natural language. The proposed method enables querying a semantic repository in natural language and obtaining results from it as a coherent text. The solution involves a conversion from natural language to SPARQL on one hand and from a set of Resource Description Framework (RDF) triples to coherent natural language descriptions in multiple languages on the other. The conversions are implemented in the Grammatical Framework (GF). The semantic knowledge infrastructure in RDF is based on OWLIM-SE and the data integration method reason-able view supplied with an ontological reference layer. The latter is connected via formal rules to a semantic representation layer and to a syntactic representation layer using GF. The resulting demonstration is a system that supports querying and text generation in 15 languages.

**Key Words** CIDOC-CRM • Grammatical Framework • Linked Open Data • Natural language generation • Ontology • OWLIM • Reason-able view • Semantic Web • SPARQL

---

M. Damova (✉) • M. Mateva  
Ontotext, AD, Sofia, Bulgaria  
e-mail: [mariana.damova@ontotext.com](mailto:mariana.damova@ontotext.com); [maria.mateva@ontotext.com](mailto:maria.mateva@ontotext.com)

D. Dannélls • R. Enache • A. Ranta  
University of Gothenburg, Göteborg, Sweden  
e-mail: [dana.dannells@chalmers.se](mailto:dana.dannells@chalmers.se); [ramona.enache@chalmers.se](mailto:ramona.enache@chalmers.se); [aarne.ranta@chalmers.se](mailto:aarne.ranta@chalmers.se)

## 1 Introduction

Cultural heritage is an excellent use case for Semantic Web technologies. Many applications in this domain require the integration and linking of different knowledge resources to ensure access to rich information and respond to the needs of different users who deal with cultural heritage content. The Semantic Web is an extension of the World Wide Web (WWW). It allows to structure information in a way that makes it possible for machines to understand the meaning of the content on the Web, interlink data and reason about it. Semantic Web technologies offer the technological backbone to meet the requirement of integrating and accessing heterogeneous data easily, but they are more adapted to be consumed by computers rather than by humans. As a result, the usability of such data among cultural heritage professionals and the general public is low.

To query a Semantic Web-based database, one has to be intimately familiar with the models according to which the data is represented and to hold good knowledge of SPARQL (Garlik and Andy 2013), the query language for Resource Description Framework (RDF) (Lassila and Swick 1999). Querying such a complex knowledge representation source is a difficult task for a non-technical person. Therefore, it is essential to find a mechanism that will allow to query semantic knowledge bases with queries formulated in natural language. Similarly, it can be preferable to convey the results returned from these knowledge bases, which are typically in the form of RDF triples, as a coherent natural language (NL) text. Since triples may be difficult to understand by non-engineers, presenting them in NL will considerably increase the usability of large semantic knowledge bases in the cultural heritage domain not only for experts but also for the general public.

This chapter presents a technique for interaction in NL with semantic knowledge bases. We describe a method that allows querying a semantic repository in natural language and obtaining results from it as a coherent natural language text. This unique solution includes several steps of transition from natural language to SPARQL and from RDF to coherent natural language descriptions in multiple languages by employing the Grammatical Framework (Ranta 2011).

The highlights of the approach and its realization are presented in the following order. Section 2 describes the technologies and the knowledge representation for query, retrieval and text generation. Section 3 describes the workflow from NL to SPARQL and from RDF to NL. Section 4 discusses some of the issues in building a multilingual grammar application from Semantic Web data. Section 5 comments on related work. Section 6 concludes with remarks about the approach and its novelty.

## 2 The Data and the Technologies

The technological infrastructure of the presented approach consists of (1) the knowledge resources structured according to the World Wide Web Consortium (W3C) standards to become interoperable with the semantic data on the

Web; (2) OWLIM, a commercial RDF database management system, developed by Ontotext; and (3) the Grammatical Framework, a free grammar resource which enables multilingual interaction with Semantic Web data in multiple languages.

## 2.1 *The Knowledge Representation*

The data layer of the Semantic Web is structured according to the Linked Data principles defined by Tim Berners-Lee as RDF graphs published on the WWW.<sup>1</sup> The idea is to explore large amount of data across servers by following the links in the graph in a manner similar to the way the Web is navigated across a multitude of distributed servers around the world. Linked data is a method for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web using Uniform Resource Identifiers (URIs) and RDF. The Linked Open Data (LOD) initiative began as an W3C project aiming to extend the Web by publishing open datasets as RDF and by creating RDF links between data items from different data sources. Currently, LOD provides more than 300 sets of referenceable, semantically interlinked resources with defined meaning. The central dataset of the LOD is DBpedia.<sup>2</sup>

Unfortunately, the distributed architecture offered by semantic web technologies does not allow to use one of its most powerful capabilities, namely, reasoning, especially because there are no mechanisms as of yet that provide streamed inference. An additional disadvantage of this distributed architecture is that it is impossible to guarantee 100 % availability of the resources, because of occasional downtimes of the servers where the resources are hosted. An approach overcoming these limitations is reason-able views (Kiryakov et al. 2010, 2009). It consists in the construction of a compound dataset from a collection of datasets performing inference on them on a single server and providing a reference layer with one unification ontology, mapped to the schemata of the single datasets constituting the reason-able view (Damova et al. 2012). This creates the conditions for efficient access and navigation of the data by allowing to formulate queries in terms of the unification ontology and retrieve data from all the datasets it contains.

This approach has been adopted for the knowledge representation infrastructure of this solution—the Museum Reason-able View (MRV). A complete description of the Semantic Web ontologies, how they were mapped and the cultural heritage data the Museum of Reason-able View gathers can be found in Damova and Dannélls (2011) as well as in Dannélls et al. (2011b). The ontologies that we provide natural language access to and that are relevant for this paper are outlined in the following sections.

---

<sup>1</sup><http://linkeddata.org>.

<sup>2</sup><http://dbpedia.org>.

### 2.1.1 Ontologies

(1) CIDOC-CRM,<sup>3</sup> an object-oriented ontology developed by the International Council of Museum's Committee for Documentation (ICOM-CIDOC), consisting of about 90 classes and 148 properties; (2) Museum Artefacts Ontology (MAO). It has about 10 classes and about 20 properties, developed to cover exhaustively the Gothenburg City Museum data; (3) painting ontology, developed to cover detailed information about paintings in the framework of the Semantic Web.<sup>4</sup> It contains 197 classes and 107 properties of which 24 classes are equivalent to classes from the CIDOC-CRM and 17 properties are sub-properties of the CIDOC-CRM properties.

To allow a unified access to cultural heritage data described according to the above conceptual models, they have been mapped to the painting ontology. The ontology is used as a reference unification ontology in order to support interoperability between natural language and ontology via SPARQL and generation of coherent natural language text (Dannélls 2011).

### 2.1.2 Data

The cultural heritage data which we made available through the MRV and that we provide NL access to are (1) 48 paintings from two collections from the Gothenburg City Museum database; (2) 614 paintings from DBpedia and (3) 167 paintings from Europeana Semantic Data.<sup>5</sup>

## 2.2 OWLIM: Semantic Data Storage

The MRV datasets are loaded into OWLIM-SE with inference performed on the data with respect to OWL Horst (ter Horst 2005). OWLIM is a family of semantic repositories<sup>6</sup> or RDF database management systems developed by Ontotext. It has the following characteristics: (a) native RDF engines, implemented in Java; (b) delivering full performance through both Sesame and Jena; (c) robust support for the semantics of RDFS, OWL 2 RL and OWL 2 QL; and (d) the best scalability, loading and query evaluation performance.

OWL Horst is an extension of RDFS (Brickley and Guha 2004) and is based on description logic (DL) (Baader et al. 2003). It is defined as an RDFS extension toward rule support as a dialect of OWL (i.e. OWL Lite, OWL DL, and OWL

---

<sup>3</sup><http://www.cidoc-crm.org/>.

<sup>4</sup><http://spraakdata.gu.se/svedd/painting-ontology/painting.owl>.

<sup>5</sup><http://europeana.ontotext.com>.

<sup>6</sup><http://www.ontotext.com/owlim>.

Full) (W3C OWL Working Group 2012), which makes use of rule entailment (R-entailment) of RDF graphs. Thus, the MRV loads the datasets and the ontologies with OWL Horst reasoning.

Implicit statements are then recorded in the repository by the process of full materialization during loading. As a result, the overall data available for query and retrieval counts 1,987,616 RDF triples derived from 460,367 explicit statements. These statistics reflect the number of triples formed by a selection of the paintings from DBpedia, Europeana Semantic Data and Gothenburg City Museum and generated triples ensuring the multilingual support of the objects. This selection ensures better quality and thorough curation of the data to be used for the experimentation of the method.

### 2.3 *The Grammatical Framework*

The grammar formalism we employ in order to support interoperability between natural language and Semantic Web ontologies is the Grammatical Framework (Ranta 2011).<sup>7</sup> It is a grammar formalism based on Martin-Löf's type theory (Martin-Löf 1984). The key feature of GF is the division between an abstract syntax, i.e. the semantic representation of the domain, and concrete syntaxes, representing linearizations in various target languages, either natural or formal.

GF comes with a resource library (Ranta 2009), covering the syntax of nearly 30 languages. The resource library aids the development of new grammars for specific domains by providing the operations for basic grammatical constructions. With GF it is possible to produce correct natural language rendering of content in all the languages that are covered in its library.

In a type-theory-based formalism such as GF, records and functions are used to describe data structures by means of features. Features of different objects can be encoded as records with record types. A record type is a tuple separated with a semicolon, e.g.:

```
Entity = {name : Str; isAnimate : Bool}
```

An object of type *Entity* is a record with two fields (separated by a semicolon), the first field with label *name* and type *Str* (string) and the second field with label *isAnimate* and type *Bool* (Boolean).

In addition to objects (such as records), GF has functions that build objects from arguments. An example of a function is

```
fun Pred : NP -> VP -> S
```

---

<sup>7</sup><http://www.grammaticalframework.org/>.



that is, the function *Pred*, which builds a sentence (*S*) from a noun phrase (*NP*) and a verb phrase (*VP*). This is an abstract syntax function, whose exact behaviour in different languages (word order, agreement, etc.) is defined by linearization rules in the concrete syntax.

The division into abstract and concrete syntaxes and the flexible use of records to encode different features of objects allow us to explore complex knowledge representation structures. These characteristics have been proved advantageous in the context of multilingual natural language generation from ontologies (Dannélls et al. 2011a, 2012; Dannélls 2012) and also in the context of multilingual semantic-based Wiki (Kaljurand and Kuhn 2013).

### 3 Multilingual Interactions with Ontologies

To enable multilingual interaction with Semantic Web data, it is necessary to provide mechanisms for mapping the syntactic analysis of the natural language input into the conceptual structure of the ontology. It is well known that one conceptual relation can be represented by multiple language realizations, e.g. declarative clauses, questions, and multiword entities (MWE) (Gromann and Declerk 2014), but their number is restricted by the semantics of the conceptual relation. That is why we argue that an ontology restricts the number of semantic queries that can be run against it, as it encompasses a logically organized semantic structure that represents a closed world defined by the concepts (ontology classes) and relations (ontology properties) that are included in it. Therefore, the number of possible semantic and hence natural language queries based on them is finite. This fact makes the ontology an excellent candidate for developing and implementing a controlled natural language application grammar that exhaustively covers all possible conceptual semantic queries (Kuhn 2013). The present approach conceives the technique for multilingual interaction with ontologies based on this assumption.

Most people who are using Web search engines usually formulate their queries with the help of keywords. However, Semantic Web data allow for more complex semantic-based queries, describing objects and their properties such as *Museum artefacts preserved in the museum since 2005, Where are the objects created by Anders Hafrin preserved, Paintings with length less than 1 m*. To retrieve results from a semantic repository, these queries have to be formulated in SPARQL, the query language of RDF. Furthermore, to allow users to interact with Semantic Web repositories in natural language, it is necessary to build translator modules that interpret and convert the natural language structures and semantics into the conceptual structure of the ontologies underlying the Semantic Web data. Below follows a step-by-step description of how the interoperability between ontologies via SPARQL and the natural language analysis and generation has been achieved.

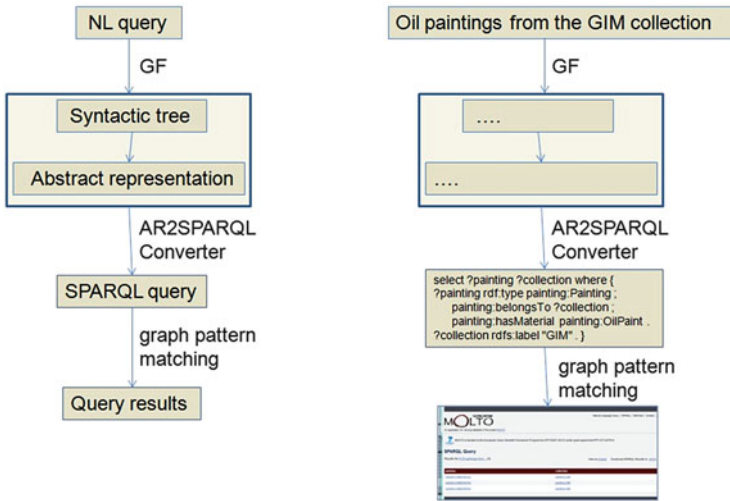


Fig. 1 NL to SPARQL processing flow

### 3.1 Querying: NL to SPARQL

The schema shown in Fig. 1 covers the flow for the analysis of natural language queries and the retrieval of the query results. The principle of the approach is illustrated on the left-hand side, and an example of the approach is provided on the right-hand side of the figure.

Our GF grammar for querying, i.e. the NL query module, uses a more general module for queries, i.e. the Yet Another Query Language (YAQL) module (Ranta 2012). YAQL provides the basis for query generation in a specific domain. For example, the grammar contains categories to allow us to describe things like names of objects, e.g. Leonard captured in *Term*; their types, e.g. political philosopher captured in *Kind* and *Property*; query statements, e.g. show, who, what captured in *Move* (the topmost category of YAQL); and functions from which statements can be linearized, such as *KProperty*, *TAll* and *MAllAbout*:

```
KProperty   : Kind -> Property -> Kind ;
TAll        : Kind -> Term ;
MAllAbout   : Term -> Move ;
```

We implemented an extra layer on top of YAQL. With this extra query layer, we gain support for expressing domain-specific queries, such as *who painted Mona Lisa* or *show everything about all oil paintings at the Louvre*:

```
PPainter     : Painter -> Property ;
PMuseum      : Museum -> Property ;
KPaintingType : PaintingType -> Kind ;
```

In addition, we implemented an extra SPARQL module that is specifically designed to map from NL to SPARQL representations. For example, to construct a SPARQL query, the one-place predicate *Property* has been redefined in the SPARQL module with additional parameters such as *material*, *museum*, *year*, etc. Some of them are associated with Boolean fields, e.g. *hasType*, *hasMaterial* to allow optionality, as illustrated below:

```
Property = {title : Str ; type : Str ; hasType : Bool ;
  material : Str ; hasMaterial : Bool ; museum : Str ;
  hasMuseum : Bool ; year : Str ; hasYear : Bool ;
  size : Str ; hasSize : Bool ; author : Str ;
  hasAuthor : Bool ; suffix : Str ; filter : Str } ;
```

The principle behind this implementation is to cover larger amount of SPARQL queries. For instance, in the concrete syntax, *type* is linearized with the default string *type* = “*?painting rdf:type painting:Painting;*” if the value of *hasType* is true. If the value is false, *type* is linearized with an empty string. Other parameters receive different linearizations depending on the type of query. As a result, the generated SPARQL query changes depending on the semantic information covered in NL query. For example, *suffix* is linearized with the string “*?museum rdfs:label ?loc.*”, and *filter* is linearized with the string “*FILTER (str(?loc)= “Musée\_du\_Louvre”)*” if the query contains a restriction of the museum, such as *show everything about all oil paintings at the Louvre*. The *MAllAbout* example below is an extract from the concrete syntax, showing the compositional approach for constructing SPARQL queries:

```
MAllAbout t =
"PREFIX painting:
<http://spraakbanken.gu.se/rdf/owl/painting.owl#> $n
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>$n
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#> $n
SELECT distinct ?painting ?title ?author ?year ?length
           ?height ?museum $n
WHERE $n { " ++ t.type ++ ";" ++ "$n" ++t.title ++ "$n"
  ++ t.museum ++ "$n" ++ t.year ++ "$n" ++ t.size ++
  "$n" ++ t.author ++ "$n" ++ t.suffix ++ "$n" ++
  t.filter ++"} $n LIMIT 200" ;
```

In the *where* statement, we can observe eight fields of type *Str* which are defined in *Property*. These fields are *title*, *type*, *author*, *year*, *size* (length and height), *museum*, *suffix* and *filter*.<sup>8</sup>

When a user formulates a query in NL, it is parsed in GF, and the results from the parser are linearized by the SPARQL module that generates the corresponding SPARQL query. For example, if the user asks

---

<sup>8</sup>The *\$n* stands for new line identifier for the back end to post-process.

*show everything about all oil paintings at the Louvre*

(or the same in any other language), then the abstract syntax returned by the parser is

```
MAllAbout (TAll (KProperty (KPaintingType PTPortrait)
(PMuseum MMus_e_du_Louvre)))
```

This tree is then linearized in GF by using the SPARQL module. The resulted query is

```
PREFIX painting:
    <http://spraakbanken.gu.se/rdf/owl/painting.owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT distinct ?painting ?title ?author ?year ?length
                ?height ?museum
WHERE {
    ?painting rdf:type painting:Portrait ;
              rdfs:label ?title ;
              painting:hasCurrentLocation ?museum;
              painting:hasCreationDate ?date;
              painting:hasDimension ?dim ;
              painting:createdBy ?author .
    ?author rdfs:label ?painter .
    ?date painting:toTimePeriodValue ?year .
    ?dim painting:lengthValue ?length ;
          painting:heightValue ?height .
    ?museum rdfs:label ?loc .
    FILTER (str(?loc)= "Mus\ '{e}e_du_Louvre" ) }
LIMIT 200
```

### 3.2 Answering: Multilingual Generation of the Query Results

The results retrieved from the above SPARQL query are returned in the form of RDF triples. In order to generate natural language descriptions from a selected set of the returned triples, these triples had to be defined in the grammar. Therefore, we implemented a *Text* module which maps from a set of RDF triples to multilingual natural language descriptions. The *Text* module captures eight classes that are most commonly used to describe a painting (Dannélls 2011), including *Title*, *Painter*, *Painting Type*, *Material*, *Colour*, *Year*, *Museum* and *Size*. Each of these classes is defined as a record and is captured in one function *DPainting* which has the following representation in the abstract syntax:

```

DPainting : Painting -> Painter -> PaintingType ->
  OptColours -> OptSize -> OptMaterial -> OptYear ->
  OptMuseum -> Description ;

```

Thus, the function *DPainting* takes eight arguments of which five are optional, i.e. *OptColour*, *OptSize*, *OptMaterial*, *OptYear* and *OptMuseum*. The advantage of this representation is that with only one function we are able to generate different descriptions depending on the information that is available about the retrieved painting.

Similar to the querying process, we have a mechanism to convert a set of triples returned from the ontology into a semantic representation. For example, one of the results returned from the query, *show everything about all oil paintings at the Louvre*, is a set of triples describing the painting *Grande Odalisque*. The set of triples covering the fields, *title*, *painter*, *size*, *year* and *museum*, is converted to the following semantic representation:

```

DPainting (PTitle TGrande_Odalisque)
  PJean_Auguste_Dominique_Ingres
  PTOilPainting_NoColours (MkSize (SIntInt 163 89))
  NoMaterial (MkYear (YInt 1814))
  (MkMuseum MMus_e_du_Louvre)

```

This semantic representation can be linearized in all of the 15 supported languages. Either a yes/no answer or a well-formed description (Dannélls 2012) is generated and returned to the user. The retrieved results are both available in the form of natural language text and RDF triples through the Web interface.<sup>9,10</sup> Here are the generated results in 10 languages:

- Cat:** Grande Odalisque fou pintat per Jean Auguste Dominique Ingres en 1814. Mesure 89 sobre 163 cm. Aquesta pintura està exposada al Museu del Louvre.
- Dut:** Grande Odalisque werd in 1814 door Jean Auguste Dominique Ingres geschilderd. Het werk is 89 bij 163 cm. Dit schilderij wordt in Musée du Louvre getoond.
- Eng:** Grande Odalisque was painted by Jean Auguste Dominique Ingres in 1814. It measures 89 by 163 cm. This painting is displayed at the Musée du Louvre.
- Fin:** maalauksen Grande Odalisque on maalannut Jean Auguste Dominique Ingres vuonna 1814. Se on kokoa 89 kertaa 163 cm. Tämä maalaus on esillä Louvressa.
- Fre:** Grande Odalisque a été peint par Jean Auguste Dominique Ingres en 1814. Il est de 89 sur 163 cm. Ce tableau est exposé au Musée du Louvre.
- Ger:** Grande Odalisque wurde in 1814 von Jean Auguste Dominique Ingres gemalt. Das Werk ist 89 mal 163 cm. Dieses Bild ist ausgestellt in der Der Louvre.
- Ita:** Grande Odalisque è dipinto da Jean Auguste Dominique Ingres in 1814. Misura di 89 su 163 cm. Questo dipinto è esposto al Museo del Louvre.
- Ron:** Grande Odalisque este pictat de catre Jean Auguste Dominique Ingres în 1814. Este din 89 pe 163 cm. Acest tablou este expus în Musée du Louvre.

<sup>9</sup>The MRV with the described natural language interface is available from <http://museum.ontotext.com>.

<sup>10</sup>The semantic data can be also extracted in JSON and XML formats.

**Spa:** Grande Odalisque fue pintado por Jean Auguste Dominique Ingres en 1814. Mide 89 por 163 cm. Esta pintura está expuesta en el Museo del Louvre.

**Swe:** Grande Odalisque målades av Jean Auguste Dominique Ingres år 1814. Den är 89 gånger 163 cm. Den här målningen är utställd på Louvren.

## 4 Multilingual Generation from Semantic Web

The current application supports interoperability between natural language and ontology models in 15 languages for querying and answering. These languages include: Bulgarian, Catalan, Danish, Dutch, English, Finnish, French, Hebrew, Italian, German, Norwegian, Romanian, Russian, Spanish, and Swedish. The specificity of the cultural heritage domain, the museum data and the amount of languages we cover in this application required certain adjustments concerning NL realizations and additions to the lexicons to support adequate translations. The multilingual issues we had to deal with are described in Dannélls et al. (2013) and summarized in this section.

### 4.1 *Lexicalizations of Ontology Content*

In the context of the semantic web, there are two ways to preserve lexical meanings across languages. A lexical unit can be either encoded directly in the ontology with the help of the *rdfs:label* predicate, i.e. *Painting rdfs:type owl:Class, Painting rdfs:label "pintura"@ep*, or indirectly through a lexicon model (Declerck et al. 2010; McCrae and Unger 2014).

Unfortunately, at the time of implementation, no multilingual translations were available. Therefore, a subset of the ontology classes and instances were translated manually by a native speaker of the language and were encoded directly in GF. We hoped we will be able to exploit some multilingual information from DBpedia to translate the ontology instances, but unfortunately there were no consistent translations of the data.

The manual work of translating the ontology classes, properties and some instances of the classes *Material* and *Colour* was estimated to less than an hour per language. The remaining translations comprise instances of the classes *Painter* and *Title* and *Museum*. Painters and painting titles remained untranslated. Translations of museum names were extracted automatically from Wikipedia.

**Table 1** The number of automatically translated museum names from Wikipedia

Language	Translated names
Bulgarian	26
Catalan	63
Danish	33
Dutch	81
Finnish	40
French	94
Hebrew	46
Italian	94
German	99
Norwegian	50
Romanian	27
Russian	87
Spanish	89
Swedish	58

## 4.2 Automatic Translations from Wikipedia

While the manual translation of the classes and the properties was an easy process in the context of this application, the translation of the instances was labour intensive. The most obvious problem we experienced is a mixture of translations such as descriptions in Italian with a museum name in English. To overcome this, we experimented with automatic translation of primarily museum names from Wikipedia. The approach of the translation process is described in detail in Dannélls et al. (2011b). Table 1 summarizes the results of the successfully translated names out of a total of 106. As can be seen in Table 1, the amount of translated names varies significantly for each language. French, Italian, German, Russian and Spanish are among the languages with the largest amount of translations with more than 90 % correct translations.

## 4.3 Linearizations from Ontology Content

To generate a coherent text from a set of RDF triples, we had to make different assumptions about how many sentences a description should consist of, how many ontology classes each sentence should convey and how to order the different classes in each sentence. We found that the most important issue to consider with respect to fluency and coherence on the sentence level is the order of the semantic information.

The first sentence of the description comprises four semantic classes: *Title*, *Material*, *Painter* and *Year*. In most languages, these classes are also realized in the listed order. Two noticeable exceptions were German and Russian whose linearizations required the following order: German: *Title*, *Year*, *Painter*, *Material*, and Russian: *Title*, *Painter*, *Material*, *Year*.

With respect to fluency and coherence on the discourse level, we observed differences in the use of reference between the languages. For example, while in most languages a pronoun is used to refer to a painting, languages such as Spanish, Italian, and Hebrew tend to use a noun or null reference.

## 5 Related Work

Natural language and ontology interoperability research area is rather new. The advances of mapping NL to ontology via SPARQL have been tested in three consequent question answering over linked data (QALD) challenges (Lopez et al. 2013; Walter et al. 2012). In these challenges, approaches to handle mapping between natural language and SPARQL differ from each other in the way the natural language input is interpreted and in the way the SPARQL query is produced.

The approach taken in Hakimov et al. (2013) is based on translating natural language questions to RDF triple patterns using the dependency tree of the question text and relational patterns extracted from the Web. Their system relies on processing the RDF predicates in a form that is comparable with the syntactic output, which makes it data-source dependent. As opposed to template-based question answering approaches over RDF data, where NL sentences are just a shortcut to formulate SPARQL sentences for non-expert users (Unger et al. 2012; Hakimov et al. 2013), our grammar-based query approach follows the WYSIWYM (what you see is what you meant) mechanism (Power et al. 1998); for example, the user can formulate queries by clicking on a proposed feedback text. In our approach, the interpretation of the formulated query derives a single semantic representation, which includes information about the sentence structure, the classes represented in it and the parts that are to be looked for. The SPARQL query is generated from the semantic representation of the sentence, similar to Gerber and Ngomo (2014). Further, our method differs from the ones presented in Ngonga Ngomo et al. (2013) and Unger et al. (2012) in that it realizes the ontology content rather than the ontology axioms.

With respect to multilinguality, many authors rely on a multi-layered ontology approach for generating multilingual descriptions (Androutsopoulos et al. 2001, 2005, 2007; O'Donnell et al. 2001; Bouayad-Agha et al. 2012). These approaches require extensive linguistic knowledge associated with the ontology classes and properties. Recently, there have been some attempts to generate descriptions in real time from a large set of ontologies (Demey and Heath 2014). In the context of cultural heritage, there have also been some attempts to generate natural language from ontologies using controlled natural language mechanism (Damljanovic and Bontcheva 2008).

Our approach differs from the above approaches as it offers mapping from abstract semantic representations to SPARQL by enabling cross-language interaction using GF. In addition, it constructs answers in the form of coherent texts, in contrast to other approaches which generate at most single grammatical sentences. Thus, the technique presented in this chapter is novel and unique in several



aspects: (1) It is a wholesome method capturing the entire cycle of interaction with the semantic knowledge base, from querying to result consumption; (2) both the analysis and the generation are based on a single interlingua semantic representation that ensures interoperability with the semantic knowledge base on the one hand and multilingual coverage on another. This becomes feasible because of the direct linking between the semantic representation and the GF resource grammars describing the syntactic structures of multiple languages.

## 6 Conclusions

This chapter presented a novel approach to natural language and ontology interoperability. The approach is used to interact with Semantic Web knowledge bases and LOD in multiple languages. It is based on the assumption that ontologies restrict the semantic queries that can be formulated over them. The grammar formalism chosen, GF provides the means to cover nearly 30 languages, which makes the interaction with the Semantic Web data in multiple languages inclusive.

The approach to GF and ontology interoperability for text analysis and generation is that the abstract syntax is driven by the ontology and the concrete syntax by the resource grammars. The grammar is successfully used by the cross-language retrieval system and supports querying and text generation. The chapter explained this approach with the cultural heritage domain as a use case. It showed the full cycle of natural language interaction, including both querying and generation over the Semantic Web knowledge infrastructure.

**Acknowledgements** This work is supported by MOLTO European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement FP7-ICT-247914. We are grateful to the anonymous referees for helpful and detailed comments.

## References

- Androutsopoulos, I., Kallonis, S., & Karkaletsis, V. (2005). Exploiting OWL ontologies in the multilingual generation of object descriptions. In *Proceedings of the 10th European Workshop on Natural Language Generation NLG*, Aberdeen, UK (pp. 150–155).
- Androutsopoulos, I., Kokkinaki, V., Dimitromanolaki, A., Calder, J., Oberl, J., & Not, E. (2001). Generating multilingual personalized descriptions of museum exhibits: The M-PIRO project. In *Proceedings of the International Conference on Computer Applications and Quantitative Methods in Archaeology*.
- Androutsopoulos, I., Oberlander, J., & Karkaletsis, V. (2007). Source authoring for multilingual generation of personalised object descriptions. *Natural Language Engineering*, 13(3):191–233.
- Baader, F., Calvanese, D., McGuinness, D. L., Nardi, D., & Patel-Schneider, P. F. (Eds.). (2003). *The description logic handbook: Theory, implementation, and applications*. Cambridge: Cambridge University Press.

- Bouayad-Agha, N., Casamayor, G., Mille, S., Rospocher, M., Saggion, H., Serafini, L., et al. (2012). From ontology to NL: Generation of multilingual user-oriented environmental reports. In *Lecture Notes in Computer Science* (Vol. 7337). Springer.
- Brickley, D., & Guha, R. (2004). *RDF Vocabulary Description Language 1.0: RDF Schema*. W3C. <http://www.w3.org/TR/rdf-schema/>.
- Damljanovic, D., & Bontcheva, K. (2008). Enhanced semantic access to software artefacts. In *Proceedings of Workshop on Semantic Web Enabled Software Engineering (SWESE) Held in Conjunction with ISWC*.
- Damova, M., & Dannélls, D. (2011). Reason-able view of linked data for cultural heritage. In *Proceedings of the 3rd International Conference on Software, Services and Semantic Technologies (S3T). Advances in Intelligent and Soft Computing* (Vol. 101, pp. 17–24). Berlin: Springer.
- Damova, M., Kiryakov, A., Grinberg, M., Bergman, M. K., Giasson, F., & Simov, K. (2012). Creation and integration of reference ontologies for efficient LOD management. In *Semi-automatic ontology development: Processes and resources* (pp. 162–199). Hershey PA, USA: IGI Global.
- Dannélls, D. (2011). *D.8.1 ontology and corpus study of the cultural heritage domain*. Deliverable of EU Project MOLTO Multilingual Online Translation.
- Dannélls, D. (2012). On generating coherent multilingual descriptions of museum objects from semantic web ontologies. In *Proceedings of the Seventh International Natural Language Generation Conference (INLG 2012)* (pp. 76–84). Utica, IL: Association for Computational Linguistics.
- Dannélls, D., Damova, M., Enache, R., & Chechev, M. (2011a). A framework for improved access to museum databases in the Semantic Web. In *Recent Advances in Natural Language Processing (RANLP). Language Technologies for Digital Humanities and Cultural Heritage (LaTeCH)*.
- Dannélls, D., Damova, M., Enache, R., & Chechev, M. (2012). Multilingual online generation from semantic web ontologies. In *Proceedings of the 21st World Wide Web Conference (WWW2012)*, Lyon, France.
- Dannélls, D., Ranta, A., Enache, R., Damova, M., & Mateva, M. (2011b). Multilingual access to cultural heritage content on the Semantic Web. In *Proceedings of the Association for Computational Linguistics (ACL) Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*.
- Dannélls, D., Ranta, A., Enache, R., Damova, M., & Mateva, M. (2013). Multilingual access to cultural heritage content on the semantic web. In *Proceedings of Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*.
- Declerck, T., Buitelaar, P., Wunner, T., McCrae, J., Montiel-Ponsoda, E., & de Cea, G. A. (2010). Lemon: An ontology-lexicon model for the multilingual Semantic Web. In *Proceedings of the World Wide Web Consortium W3C Workshop: The Multilingual Web – Where Are We?* Madrid, España: Universidad Politécnica de Madrid.
- Demey, Y. T., & Heath, C. (2014). Towards verbalizing multilingual N-ary relations. In P. Buitelaar & P. Cimiano (Eds.), *Towards the multilingual semantic web: Principles, methods and applications*. Berlin: Springer. doi:10.1007/978-3-662-43585-4.
- Garlik, S. H., & Andy, S. (2013). *SPARQL 1.1 Query Language*. <http://www.w3.org/TR/sparql11-query/>.
- Gerber, D., & Ngomo, A.-C. N. (2014). From RDF to natural language and back. In P. Buitelaar & P. Cimiano (Eds.), *Towards the multilingual semantic web: Principles, methods and applications*. Berlin: Springer. doi:10.1007/978-3-662-43585-4.
- Gromann, D., & Declerck, T. (2014). A cross-lingual correcting and complete method for multilingual ontology labels. In P. Buitelaar & P. Cimiano (Eds.), *Towards the multilingual semantic web: Principles, methods and applications*. Berlin: Springer. doi:10.1007/978-3-662-43585-4.
- Hakimov, S., Tunk, H., Akimaliev, M., & Doglu, E. (2013). Semantic question answering system over linked data using relational patterns. In *Proceedings of the 16th International Conference on Extending Database Technology (EDBT/ICDT)*, Genoa.

- Kaljurand, K., & Kuhn, T. (2013). A multilingual semantic wiki based on Attempto Controlled English and Grammatical Framework. In *Proceedings of the 10th Extended Semantic Web Conference (ESWC 2013)*. Berlin: Springer.
- Kiryakov, A., Ognyanoff, D., Velkov, R., Tashev, Z., & Peikov, I. (2009). LDSR: Materialized Reason-able view to the Web of linked data. In *Proceedings of the 5th International Workshop on OWL: Experiences and Directions*, Chantilly, USA.
- Kiryakov, A., Ognyanoff, D., Velkov, R., Tashev, Z., & Peikov, I. (2010). *LDSR: Materialized Reason-able view to the Web of linked data*. Sofia: Ontotext AD.
- Kuhn, T. (2013). A survey and classification of controlled natural languages. *Computational Linguistics*, 40(1):121–170
- Lassila, O., & Swick, R. R. (1999). *Resource Description Framework (RDF). Model and Syntax Specification*. <http://www.w3.org/TR/REC-rdf-syntax>.
- Lopez, V., Unger, C., Cimiano, P., & Motta, E. (2013). Evaluating question answering over linked data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 21.
- Martin-Löf, P. (1984). Intuitionistic type theory. Napoli: Bibliopolis.
- McCrae, J. P., & Unger, C. (2014). *Design patterns for engineering the Ontology-Lexicon interface*. In P. Buitelaar & P. Cimiano (Eds.), *Towards the multilingual semantic web: Principles, methods and applications*. Berlin: Springer. doi:10.1007/978-3-662-43585-4.
- Ngonga Ngomo, A.-C., Bühmann, L., Unger, C., Lehmann, J., & Gerber, D. (2013). Sorry, I don't speak SPARQL – Translating SPARQL queries into natural language. In *Proceedings of the 22nd World Wide Web Conference (WWW)*.
- O'Donnell, M. J., Mellish, C., Oberlander, J., & Knott, A. (2001). ILEX: An architecture for a dynamic hypertext generation system. *Natural Language Engineering*, 7(3):225–250.
- Power, R., Scott, D., & Evans, R. (1998). What You See Is What You Meant: Direct knowledge editings with natural language feedback. In *Proceedings of the 13th European Conference on Artificial Intelligence (ECAI 1998)* (pp. 677–681). Chichester: Wiley.
- Ranta, A. (2009). The GF resource grammar library. *The On-line Journal Linguistics in Language Technology (LiLT)*, 2(2):1–65.
- Ranta, A. (2011). *Grammatical framework: Programming with multilingual grammars*. *CSLI Studies in Computational Linguistics*. Stanford: CSLI.
- Ranta, A. (2012). *Implementing programming languages. An introduction to compilers and interpreters*. London: College Publications.
- ter Horst, H. J. (2005). Combining RDF and part of OWL with rules: Semantics, decidability, complexity. In *Proceedings of the 4th International Semantic Web Conference (ISWC). Lecture Notes in Computer Science* (Vol. 3729, pp. 668–684). Berlin: Springer.
- Unger, C., Buehmann, L., Lehmann, J., Ngomo, A.-C. N., Gerber, D., & Cimiano, P. (2012). Template-based question answering over RDF data. In *Proceedings of the 21st World Wide Web Conference – Ontology Representation and Querying: RDF and SPARQL*.
- W3C OWL Working Group. (2012). *OWL Web Ontology Language Overview*. <http://www.w3.org/TR/owl2-overview/>.
- Walter, S., Unger, C., Cimiano, P., & Baer, D. (2012). Evaluation of a layered approach to question answering over linked data. In *Proceedings of the 11th International Semantic Web Conference (ISWC)*, Boston.

# A Cross-Lingual Correcting and Completive Method for Multilingual Ontology Labels

Dagmar Gromann and Thierry Declerck

**Abstract** Multilingual content in ontologies has one of the highest potentials for bridging linguistic borders on the Semantic Web. Human readability and automated linguistic processing of Multilingual Semantic Web resources depend on natural language content represented in labels. As there are currently no standards or best practices for labeling ontologies, existing labels are frequently highly condensed up to the point of losing their domain-specific expressivity. For instance, ellipses often used in labels pose a challenge to linguistic processing. Elided domain-specific elements challenge human users and machines alike. Thus, the proposed method expands condensed labels in four main processing steps by resolving complex natural language phenomena. It heavily relies on a cross-lingual comparison and employs idiosyncratic benefits of one language to process other languages.

**Key Words** Cross-lingual patterns • Ontology-based NLP • Ontology design patterns • Ontology labels • Terms and subterms

## 1 Introduction

Natural language expressions are most frequently added to ontology elements by means of the annotation property `rdfs:label` (Ell et al. 2011), which supports multilinguality. Human users require that information in order to access, query, understand, and manipulate formal knowledge represented in the ontology (Garcia et al. 2012). Condensing or shortening complex labels, that is, multiword expressions, complicates their comprehension by humans as well as linguistically based processing. Automated natural language processing (NLP) tasks, such as

---

D. Gromann (✉)

Vienna University of Economics and Business, Welthandelsplatz 1, 1020 Vienna, Austria  
e-mail: [dgromann@wu.ac.at](mailto:dgromann@wu.ac.at)

T. Declerck

DFKI GmbH, LT-Lab, Stuhlsatzenhausweg 3, 66123 Saarbruecken, Germany

ICLTT, Austrian Academy of Sciences, Sonnenfelsgasse 19/8, 1010 Vienna, Austria  
e-mail: [declerck@dfki.de](mailto:declerck@dfki.de)

tokenization, lemmatization, decomposition, and part-of-speech (POS) tagging, frequently stumble over complex language phenomena, such as ellipses. Information extraction (IE) returns highly ambiguous results or even misses relevant information in texts when based on shortened ontology labels. Omitting contextual complements renders labels more difficult to comprehend for human users as well. Furthermore, most ontology matching approaches still rely on a string-based comparison of entity labels to estimate the initial likelihood of two elements being equivalent (Trojahn et al., this volume). Expanding labels by inserting the elided content eases those language-based processes as well as human understanding. The detection and supplementation of elided content strongly depend on language-specific features. That is why we propose a cross-lingual method based on correcting-completive patterns (CCPs) for performing expansions of ontology labels.

In four main steps, the proposed cross-lingual method automatically resolves ellipses, adds elided domain-specific complements, and identifies comprised subterms. Firstly, nonlexical symbols are replaced by their lexical equivalents. Secondly, a label expansion is performed by a cross-lingual resolution of ellipses employed in existing labels. This step also entails a classification of different types of ellipses. Thirdly, labels that are more general than the concept they designate are expanded by a context-determining complement. Detecting this complement requires a cross-lingual analysis of labels and definitions attached to them. Finally, an external repository of identified subterms of expanded labels is generated to ease information extraction and multilingual label alignment. A total of seven languages is used to evaluate the method—English, German, Spanish, French, Italian, Russian, and Chinese—contained in four industry classification ontologies described in Sect. 3. All analyses and processing steps of language phenomena herein are based on multilingual language data in labels of those ontologies.

While the first step is comparatively straightforward, for example, replacing ampersands, colon, and semicolons by their lexical representation in the respective language, the other three require a thorough linguistic analysis and pattern-based implementation. The proposed CCPs formalize recurrences of linguistic phenomena, similar to lexico-syntactic patterns (Gangemi and Presutti 2009), but the latter focus more on acquiring logical elements from text. CCPs are the basis for automating the expansion of ontology labels. As the process is triggered in one language but applied to several other languages, it has to be considered cross-lingual rather than multilingual. The language that triggers the pattern can vary, depending on the problems to be solved and the actual language coverage of the ontology under consideration.

Each step of the proposed method will be detailed and exemplified in Sect. 2. Subsequently, the data sets utilized for evaluating the method are introduced, and the results of that evaluation are presented. Similar approaches regarding the content of labels and linguistic patterns are discussed prior to some concluding remarks.

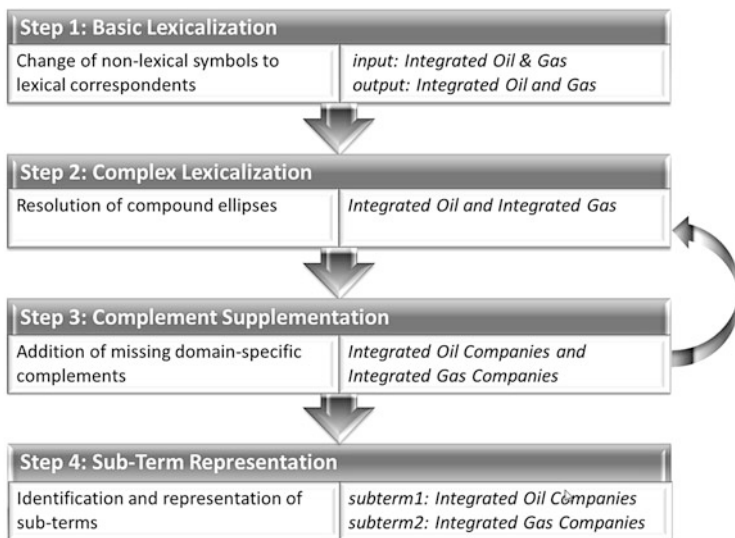


Fig. 1 Overview of four-tiered label expansion method

## 2 Correcting-Compleitive Method

To remedy linguistically based processing issues arising from the surface realization of ontology labels, they are processed by a four-tiered correcting and compleitive method we have partially introduced previously (see Declerck and Gromann 2012). The proposed method is correcting as it detects misalignment and missing complements across languages and replaces non-lexical with lexical elements. Complement is defined herein as a domain-specific, disambiguating element that alters the meaning of the expression when entirely omitted, for example, “integrated oil” without the complement “companies” attached to it. It is compleitive as it expands labels by missing context-determining complements. The four main processing steps of the proposed method, each relying on its own set of CCPs, are depicted and exemplified in Fig. 1. While the right side details the processing step, the left exemplifies its output.

In the example of Fig. 1, a first preprocessing step replaces the ampersand prior to resolving the ellipsis that is clearly indicated in German by a hyphen. The domain-specific complement “companies” in the third step are taken from the natural language definition associated with the English label of the ontology class. As this expansion might lead to new ellipses, step three reiterates the second resolution step before continuing to step four. Finally, potential subterm structures of domain-specific labels are analyzed and represented.

**Table 1** Nonlexical items processed by the first step of the proposed method

Symbol	Lexical equivalent
& (ampersand)	English and
	German und
	French et
	Italian e or ed
	Spanish y or e
	Chinese 和 or 與
Russian и	
: (colon)	German preposition or premodifying adjective
	Italian conjunction or premodifying adjective
; (semicolon)	Equivalent to &

In order to load, extract, and later add the processed labels to the ontology, the OWL API,<sup>1</sup> object-oriented programming components, and the linguistic development environment NooJ<sup>2</sup> have been used. Resulting processed labels are added to the original ontology concept by means of our subproperty of `rdfs:label` called “expanded,” while the respective subterm structure is added as external resource. Each of the above steps is described in detail in the following subsections.

## 2.1 Basic Lexicalization

Many linguistic analyzers, such as POS taggers or decomposers, classify nonlexical items without further investigating their meaning. Thus, an expression using such an item is frequently not processed correctly. This is why this preprocessing step lexicalizes the nonlexical items listed in Table 1 by means of a pattern-based approach. This first step of our method is considered correcting, since it facilitates the correct linguistic processing of labels by NLP and IE tasks.

A detailed analysis of linguistic recurrences specific to each language is vital to this task. While the ampersand can globally be replaced for English, German, French, and Russian, other languages require specific patterns. In front of vowels, the Italian “e” needs to be “ed,” and when followed by “(h)i” the Spanish “y” turns into an “e.” As regards Chinese, “和” and “與” are frequently used to connect nouns and noun phrases. The decision which one to use is currently based on the frequency of its occurrence in the resource.

Semantics of a colon as utilized within the labels of the analyzed ontologies strongly depend on the language in which it is used. In our data set, colons

<sup>1</sup><http://owlapi.sourceforge.net/>.

<sup>2</sup><http://www.nooj4nlp.net>.

were only utilized in German and Italian. In German, the colon is used to avoid complex ellipses or lengthy compounds. Its replacement requires a preposition that is obtained from analyzing synonymous labels in other languages. For instance, “Erdöl und Erdgas: Ausrüstung und Dienste”@de (Oil and Gas: Equipment and Services) requires the German preposition “für”@de (for) which can be derived from the available “pour”@fr and “per”@it. The CCP output “Ausrüstung und Dienste für Erdöl und Erdgas” for this example also involves a syntactic reordering, which is derived from the ordering of other labels with languages similar to German. The Italian use of the colon substitutes a coordinating conjunction, which can be replaced equivalent to the ampersand. Both languages require a different processing when the colon is succeeded by an adjective, in which case the adjective is affixed to the expression before the colon.

Finally, the semicolon usually joins sentences in a coordinating or adversative function. Considering the function of labels, namely, to designate a specific ontology element, semicolons in labels are most unlikely adversative. This is why they are replaced in line with the ampersand.

## 2.2 *Complex Lexicalization*

In linguistics, anaphora, cataphora, and ellipsis resolutions focus on elided content relying on previous or succeeding utterances or statements. However, labels of domain ontologies provide concise domain-specific multiword expressions with a minimum of syntagmatic structure. Nevertheless, the linguistic environment and especially the differences in this environment across languages provide the basis for resolving existing ellipses.

Compound ellipses result from a deletion process of identical constituents. In order to structure the proposed pattern-based and language-specific ellipsis grammars, we categorized types of ellipses relevant with respect to the purpose at hand. **Syntactic ellipses** feature a clear indication of elided content by means of syntactic elements, for example, the hyphen in “Metall- und Glasbehälter”@de (Metal and Glass Containers). **Structural ellipses** depend on the linguistic structure or pattern for the detection and resolution of the elided element(s), such as “Equipos e Instrumentos Electrónicos”@es (Electronic Equipment and Instruments), where “Electrónico(s)” is elided after the first noun. **Contextual ellipses** refer to the elision of domain-specific complements, for example, “Aluminum”@en missing the reference to “Producers of.” This last type of ellipsis is more complicated and thus is attributed a more detailed discussion in Sect. 2.3. While this last type is completed in step three of our correcting-compleitive method, the first two are resolved as part of step two. Each set of labels is analyzed consecutively as to the presence of any of these types of ellipsis.

Syntactic ellipses in our data set refer to up to four elements. Furthermore, the ellipsis might be on the left or right side of a coordinating conjunction. As illustrated in Table 2, the Penn Treebank tag-set is used to formalize the linguistic content of



**Table 2** Selected formalized pattern for four languages and examples

DE: <NN1>hyphen und <NN2+NNS>resolved to <NN1+NNS>und <NN2+NNS>
EN: <NN1>and <NN2><NNS>resolved to <NN1><NNS>and <NN2><NNS>
ES: <NNS><IN><NN1>y <NN2>resolved to <NNS><IN><NN1>y <NNS><IN><NN2>
RU: <NN><JJ1>и <JJ2><NNS>resolved to <NN><JJ1><NNS>и <NN><JJ2><NNS>
DE: <ELL="Metall#behälter und Glasbehälter">Metall- und Glasbehälter</>
EN: <ELL="Metal #Containers and Glass Containers">Metal and Glass Containers</>
ES: <ELL="Contenedores de Metal y #Contenedores #de Cristal">Contenedores de Metal y Cristal</>
RU: <ELL="Производство металлической #тары и #Производство стеклянной тары">Производство металлической и стеклянной тары</>

an ellipsis. To differentiate elements with identical POS tags, we numbered them. Table 2 exemplifies a left-side syntactic ellipsis detection and resolution by means of a German hyphen across four languages utilizing actual labels from the data set. “ELL” is the short form for annotating an ellipsis in the textual analysis whereby the preceding hash sign indicates the supplemented elements. The expression between the two sets of angle brackets is the input to our method.

A cross-lingual comparison confirmed that the use of hyphenation as an indication of ellipses is most consistently used in German. Moreover, a German version of ontology labels is more frequently available than other languages with clear hyphenation indicators, such as Scandinavian languages, Finnish, or Icelandic, and German has the highest frequency of ellipses within the resources analyzed.

A clearly indicated elided content in German might require a substantially different resolution pattern in another language. The English pattern in Table 2 is equivalent to the German one, apart from missing the hyphen and an open noun compound separated by a space. Spanish uses a preposition to join the noun with the (plural) complement, which needs to be inserted with the noun. Russian even requires a twofold resolution, appending and prefixing a noun phrase to the adjective. As regards cross-lingual similarities, it could be observed that German, English, and Scandinavian languages are dominated by nominal structures. Baltic languages, Estonian, Spanish, Italian, French, Icelandic, and Russian more frequently require prepositional and adjectival complement supplementation. In case there is no syntactic element to trigger the process, the label is searched for structural ellipses.

Structural ellipses within the context of domain ontology labels most frequently rely on adjectival patterns and prepositional modifying phrases to elide content. The first set of labels of Table 3 provides an example of an adjective supplementation to resolve the ellipses, while the second set illustrates a premodifying prepositional phrase. The preposition is only present in German and Spanish in this example but nevertheless correctly triggers the resolution of English and Russian as well.

The second set of labels in Table 3 exemplifies a domain-specific expression that can equally be utilized to resolve ellipses, that is, “Extraction of.” Such references

**Table 3** Examples of structural ellipses

DE: <ELL="Elektronische Geräte und #Elektronische Instrumente">Elektronische Geräte und Instrumente</>
EN: <ELL="Electronic Equipment and #Electronic Instruments">Electronic Equipment and Instruments</>
ES: <ELL="Equipo #Electrónico e Instrumentos Electrónicos">Equipo e Instrumentos Electrónicos</>
RU: <ELL="Производство электронного оборудования и #Производство #электронных приборов">Производство электронного оборудования и приборов</>
DE: <ELL="Gewinnung von Erdöl und #Gewinnung #von Erdgas">Gewinnung von Erdöl und Erdgas</>
EN: <ELL="Extraction of crude petroleum and #extraction #of natural gas">Extraction of crude petroleum and natural gas</>
IT: <ELL="Estrazione di petrolio greggio ed #estrazione di gas naturale">Estrazione di petrolio greggio e di gas naturale</>

frequently point to an ellipsis. Encoding these recurrences as CCPs supports the resolution of structural ellipses. Furthermore, it is essential to always consider language-specific features, such as the need of “ed” instead of “e” in the Italian resolution in Table 3.

Automating the expansion of elliptical labels with or without German hyphens as trigger is nontrivial, since it requires both the analysis of German compounds and the resolution of a compound ellipsis. Subsequently, both steps need to be extended to all other languages featured in the ontologies. Thus, there is a need to use and adapt a morphological analysis component and write idiosyncratic ellipsis grammars for each language. Both have been implemented in NooJ. Furthermore, a grammar to adapt the moved complements to correspond in gender and case to its environment is needed. In Table 3, this is shown by adapting the Spanish “Electrónico” and the Russian “ЭЛЕКТРОННЫХ” to the pertaining noun.

### 2.3 Cross-Lingual Supplementation of Domain-Specific Complements

Labels attached to a concept can be assumed to designate this ontology concept and to be synonymous. However, at times they provide different levels of granularity, that is, one label provides more domain-specific information than another. Within the proposed typology, this omission of domain-specific content is called contextual ellipsis. In addition, ontology concepts are occasionally not only labeled, but also defined in natural language. These definitions frequently contain such a domain-specific constituent elided in the label. Herein, these constituents are called complements as they complete the domain-specific meaning of an expression and are usually adjectives, nouns, noun phrases, or modifying phrases. Adding these

**Table 4** Example listing of cross-lingual complement extraction

English	German	Italian	Spanish	Russian	Chinese
Producer	Hersteller	Produttori	Productores	производители	生商
Stores	Geschäfte	Negozi	Tiendas	торговля	店
Retail	Einzelhandel	Negozi	Venta	Розничная торговля	零售
Retailer	Einzelhändler	Vendita al dettaglio	Minoristas		零售商
Wholesalers	Großhändler	Vendita all'ingrosso	Mayoristas		批商
Distributors	Vertriebs- unternehmen	Distributori	Distribución	Деятельность дистрибьюторов	分銷商
Distributors	Vertrieb	Distributori	Distribuidores	Дистрибьюторы	經銷商
Providers	Anbieter	Fornitori	Proveedores		供商

meaningful components to the label as step three of our method is vital to ensure its transparency, that is, the meaning of the concept can at least partially be inferred from the label without any further logical or natural language definition.

To initiate the process, a list of commonly used domain-specific complements is extracted and aligned across all available languages. An excerpt of such a list for the domain of industry classifications is illustrated in Table 4. This set of aligned terms is vital to automatically identifying the elision of the domain-specific reference in case it is omitted in all languages. Should the domain-specific reference be available in one language, the established list accelerates the addition of omitted complements across languages.

If no label contains any indication of a domain-specific reference, the definition is analyzed. If available, the definition frequently points to one of the terms in the previously established list. In case no natural language definition is available, the hierarchical structure is traversed to see whether the superordinate concept features a natural language definition. Should this superordinate definition be available, it is assumed that the same complement can be applied to the subordinate class. Thus, the domain-specific complement of the superordinate label is added to the subordinate label.

If the complement were added based on English as a pivotal language, the subtle difference in granularity of, for example, “distribución”@es and “distribuidores”@es, both being “distributors”@en in English, would have been lost. However, dynamically extracting complements from natural language labels and definitions and a cross-lingual comparison ensures that these differences are retained.

For instance, for the concept “Health Care Providers,” all domain-specific complements marked in bold below were originally part of the taxonomy, while the ones marked in italics could be added due to the cross-lingual comparison and aligned list of complements: Health Care **Providers**@en, 生保健供商@zh, *Fornitori*

*di Servizi Sanitari@it*, *Terveyspalveluyritykset@fi*, *Anbieter von Medizinischen Leistungen@de*, and so on.

Once the appropriate domain-specific complement has been identified, attaching it to the label also requires thorough linguistic analysis. Frequently, additional prepositions or a change of gender or case is needed. For this purpose, a number of lexico-syntactic patterns and grammar rules encoding the language-specific behavior are applied. Apart from some exceptions, domain-specific references together with the appropriate preposition are added as premodifier to the label.

Adding the complement to a label with two major components, for example, two noun phrases separated by a conjunction, requires the addition of two complements to avoid creating another ellipsis. This is why this step refers back to step two in Fig. 1. Having added the complement, the label is tested regarding existing ellipses. If required, these ellipses are resolved utilizing the previous processing step described in Sect. 2.2.

At this point, the expanded labels are added to the ontology by means of the annotation property “expanded” to clearly differentiate them from the original labels. They have undergone the process of ellipsis resolution and complement supplementation and represent the starting point for the external subterm repository.

## 2.4 Subterm Structures

Current labeling practices on the Semantic Web represent labels and terms without any information on their internal semantics or structure. Several representation models related to ontologies, such as the ontology-lexicon format *lemon* by McCrae et al. (this volume), allow for a more fine-grained representation. In the *lemon* model, multiword expressions can be represented as words, phrases, or parts of words, and the decomposition of a phrase can be clearly indicated. Decomposition refers to the process of separating a multiword expression into its component words. The final step of the proposed method equally decomposes labels into subterms. In some contexts, the expression subterm is used to indicate that the term is hierarchically subordinated to another term. Here, we use it in the sense that the subterm is equal to a substring of the term, that is, is contained in the term. In contrast to lexical units, the focus here clearly is on preserving term transparency, that is, each subterm maintains the domain context and the multilingual alignment of each subterm set. Thus, components of a term differ from a subterm, as a component need not be domain-specific or transparent on its own.

While there usually is one ontology-lexicon for each language, the proposed representation of subterms relies on the terminological practice of aligning synonymous designations of one concept across languages. The domain-specific references, which if possible are added in the previous processing step, and the conjunction are taken as separation markers. The number of complements corresponds to the number of subterms. The alignment of complements in the previous step eases the establishment of equivalences of the remainder of the label. If there is no

complement, only the conjunction and possible commas are taken as separation markers. During the process of alignment, lemmata of individual words are compared to ensure equivalence.

This cross-lingual comparison to achieve a multilingual alignment may uncover a number of variations across languages within one resource. On the one hand, equivalents may change with the context, such that “Products”@en is “Artikel”@de once and “Produkte”@de for another label. As identical subterms in one language might not be duplicated in another language, identical subterms may be kept and marked as equivalent. On the other hand, variation can originate from different levels of granularity in the conceptualization. While the Italian label “Esplorazione e Produzione di Petrolio e Gas Naturale”@it clearly references oil and gas as the object to be explored and produced, the other labels only refer to exploration and production.

During the automated alignment of subterms across languages, a list of strongly diverging labels, such as in terms of number of elements, is created. A subsequent manual inspection allows for the supplementation of omitted details, for example, adding “Oil and Gas” to all other subterms based on the one language containing it. This addition relies on other labels or attached natural language definitions to identify the equivalents of “Oil and Gas”@en in all other languages. Other types of varying conceptualization, such as verbose paraphrases instead of a term in Russian, have been aligned without further processing.

Although most variants reveal contextual differences within a domain, some help uncovering erroneous alignments and conceptualizations in the original resource. For instance, “Groß- und Einzelhandel”@de (wholesaling and retailing) is conceptually not equivalent to “Retailing”@en. Due to the difference in number of complements and a comparison with the list of complements of step three, in which both German terms have different English equivalents, the inconsistency can be detected automatically and corrected manually.

By creating an additional terminological resource, we derive an easily (re-)usable repository of subterms for Information Extraction and similar ontology-based linguistic processing activities. Stored in an external OWL module, the subterms reference the expanded labels in the original resource. One set of subterms in different languages is grouped by means of the data category *terminological entry*. This data category is part of the ISOcat data category repository,<sup>3</sup> a point of reference for providing an easily comprehensible and reproducible model. Each such entry uses the ID of the ontology concept with an integer as ascending suffix, for example, GICS20201040-1. Thereby, possible alignments with other ontologies are facilitated as synonyms or quasi-synonyms can be added to the subterm resource and clearly indicated by means of data categories. Should an equivalence with all subterms be identified, the label of the other ontology can be regarded and modeled as equivalent to the original label.

---

<sup>3</sup><http://www.isocat.org/atacat/>.

**Table 5** Overview of languages contained in ontology repository

	GICS	ICB	DAX	NACE
English	✓	✓	✓	✓
German	✓	✓	✓	✓
Italian	✓	✓	✗	✗
Spanish	✓	✓	✗	✗
Chinese	✓	✓	✗	✗
French	✓	✗	✗	✗
Russian	✓	✗	✗	✗

### 3 Data Sets

In order to exemplify and evaluate the proposed method, we apply it to multilingual labels of industry classification ontologies derived from structured company-related information from the Web and written in the Web Ontology Language (OWL) (Hitzler et al. 2012). Three of them were adapted as part of the Monnet Federated Financial ontology (Krieger et al. 2012), namely, the German Stock Index (DAX), Industry Classification Benchmark (ICB), and the Statistical Classification of Economic Activities in the European Community (NACE) ontology. The Global Industry Classification Standard (GICS<sup>4</sup>) ontology created by the authors is used additionally. All four combine input from international research teams and domain expertise as regards portfolio and investment analyses. We note also that ICB is used, for example, in four languages at the Euronext page (<https://europeanequities.nyx.com/icb>) and GICS for the S&P indices (<http://www.spindices.com/>).

Although each resource is multilingual, the comprised languages vary. The number of languages in labels for the ontologies we used is depicted in Table 5. The ICB ontology originally only contained English, Spanish, and German, to which we added Chinese and Italian. ICB offers its taxonomy also in these two languages online.

The basic subsumption hierarchy of each ontology is derived from the hierarchical structure of the original taxonomy ranging from industry sector to subindustries. Due to its substantial tool support and the fact that it is the most widely used ontology language, all four ontologies are represented in OWL. The proposed method, however, is not limited to the use of OWL as it focuses on the natural language and not logical content. Each ontology class of ICB, GICS, and NACE is identified by a unique integer derived from the original taxonomy, which is higher the lower its position in the hierarchy is. Natural language definitions from the leaf nodes of the taxonomic structures are assigned to the corresponding ontology class by means of the annotation property `rdfs:comment`. Since the DAX ontology

<sup>4</sup>Developed and issued by Standard & Poor's and MSCI <http://www.msci.com/products/indices/sector/gics/>.

has no numerical identifier, natural language designations identify each concept in camel case notation.

## 4 Results

Evaluating the method of expanding and segmenting ontology labels is based on the data sets introduced in Sect. 3. Results of the evaluation are illustrated in Table 6 and are based on a manual evaluation by researchers native to the respective language. Evaluators were provided with the original label and definition as well as the processed label for comparison. Idiomatic, grammatical, or basically any issues originating from the original labels were ignored.

For the present repository of ontologies, the basic lexicalization step of this method was applied to all resources, but only English, Italian, and German contained any nonlexical symbols. Out of the 470 comprised symbols, 462 could be replaced without any issues, corresponding to 98.3 %. The eight problematic labels originate from a more complex resolution of the colon, which lead to grammatically incorrect constructs.

Table 6 provides the total number of original labels for each resource. “Single,” “double,” and “triple++” refer to the count of syntactic indicators for an ellipses and thus the category of syntactic ellipses, while “structural” refers to the correspondent category of ellipses. The first “resolved” refers to the number of successfully expanded labels provided as count and percentage. “Missing complement” denominates the number of identified elided domain-specific complements. The second “resolved” in Table 6 indicates how many of these complements could be added successfully to the labels as a count and a percentage.

**Table 6** Results of ellipsis resolution and complement supplementation

	English	German	Spanish	French	Italian	Russian	Chinese	Total
No. of labels	1,575	1575	453	269	1449	269	453	6,043
Single	69	63	39	35	39	28	39	312
Double	76	82	6	4	6	3	6	183
Triple++	75	75	3	2	3	2	3	163
Structural	238	238	31	16	31	17	31	602
Total	458	458	79	57	79	57	79	1,267
Resolved	422	423	71	52	71	35	63	1,137
Percentage (%)	92.14	92.36	89.87	91.23	89.87	61.40	79.75	89.74
Missing complement	511	501	332	193	330	194	333	2,394
Resolved	388	391	234	132	289	78	294	1,806
Percentage (%)	75.93	78.04	70.48	68.39	87.57	40.21	88.29	75.44

The reason that the ellipsis resolution process seems substantially more successful than the complement supplementation in Table 6 can partially be attributed to the fact that not all labels have a definition from which to derive a complement. The extraordinarily low result for Russian is due to a high number of grammatical (e.g., case) and semantic (e.g., producer of the production of) errors. As the utilized NACE ontology does not feature any definitions at all and only three languages, it has not been considered in this step. Issues for the ellipsis resolution particularly arise from the paraphrasing of labels in specific languages, where other languages use short and precise terms. Furthermore, complex ellipses with synchronously left-hand and right-hand elisions could not be resolved automatically, for example, “Elektrizitätsverteilungs- und -schalteinrichtungen”@de (electricity distribution and control apparatus).

Creating a subterm repository depends on two individual processes: subterm extraction and alignment. Each label is separated into domain-specific subterms. Quantifying the generation of subcomponents, that is, labels separated at the coordination or comma, results in a 100 % success factor for all 3,394 labels concerned. However, these subcomponents do not necessarily comply with the criterion of domain specificity of subterms, even if a contextual complement is provided. For instance, the German “Herstellung von Erzeugnissen daraus”@de (manufacture of articles made thereof) depends on “Vliesstoff”@de (nonwovens) in another sub-term of the label, while the English equivalent subterm “articles made from nonwovens”@en explicitly mentions the industry domain. Both, creation and alignment of subterms require these domain references.

Subterms are aligned to equivalent subterms originating from the same ontology concept. Thus, the alignment process faces the same issue of domain specificity. Matching the subcomponents by their count and position in the original label can be automated and quantified. However, their content differs across languages. This difference originates from references to other parts of the label and complements omitted in specific languages or differing conceptualizations. Nevertheless, a total count of 3,197 domain-specific subterms (37 %) could be aligned in seven languages.

## 5 Related Work

Research at the intersection of ontologies and natural language spans a range of different fields. In general, approaches reconciling linguistic data and ontologies can be classified as either (1) underpinning linguistic data with ontological modeling techniques (e.g., Bond et al., this volume) or (2) associating linguistic (e.g., Declerck and Lendvai 2010) and/or lexical (e.g., McCrae et al., this volume) data with ontologies. Linguistic patterns are frequently applied to the acquisition of labels or logical elements (Gherasim et al. 2013), but their use for the expansion of existing labels seems to be a novel approach. Similarly, ellipsis resolution has a long tradition in linguistics and relies on the discursive context. However,



domain-specific expressions in ontology labels are devoid of such context and thus require quite a different approach.

## ***5.1 Ontology Labels***

Entities of Semantic Web resources need to be expressed in natural language in addition to logic, to be meaningful to human users. Ell et al. (2011) found `rdfs:label` to be the most frequently used labeling property, whereas the most frequent language across all properties is English (Garcia et al. 2012). If a resource lacks human-readable labels, many approaches, such as an automated natural language representation of inferences drawn from the ontology (Nguyen et al. 2013), use fragment identifiers of URIs to produce natural language expressions. Alternatively, labels might be extracted from structured Web resources, such as DBpedia, or unstructured text. Although individual endeavors at standardizing labels exist (e.g., Fliedl et al. 2007; Montiel-Ponsoda et al. 2011), there are no general best practices or guidelines regarding their representation or their internal semantics. Fliedl et al. (2007) investigate and exemplify the substantial heterogeneity of term usage not only in annotation properties but also in URIs. They identify the heterogeneity and different linguistic styles as one of the core problems of ontology interpretation and reuse.

## ***5.2 Cross-Lingual or Multilingual***

The terms cross-lingual and multilingual are often confusingly used interchangeably in literature. While multilingual refers to entities being described in different natural languages, cross-lingual links one entity in one natural language to an entity in another language. The same differentiation applies to current ontology matching approaches, a method to access semantics across natural languages and resources (Trojahn et al., this volume). For instance, Fu et al. (2012) achieve a cross-lingual matching by translating the URI fragments of one resource in one natural language to the natural language of another resource. Also extracting information from (un)structured resources based on ontologies can either be multilingual (Federmann et al. 2012), that is, using multiple languages, or cross-lingual, that is, starting with an ontology in one language but extracting information in another language (Wimalasuriya and Dou 2010). The method presented herein can be considered cross-lingual in that it uses one language to trigger the correcting and complete process in other languages.

### 5.3 *Linguistic Patterns*

Current results of our work might best be compared to state-of-the-art research in the field of lexico-syntactic patterns, which are part of ontology design patterns<sup>5</sup> and mostly used for learning ontologies from natural language text (e.g., Gherasim et al. 2013). Instead of learning ontologies, we develop and use linguistic patterns in order to expand existing cross-lingual content of ontologies. The major problem of such patterns is low precision and overgeneralization, which Maynard and Peters (2009) try to overcome by restricting their main approach to three sets of patterns. Similarly, the patterns presented herein are limited to ellipsis resolution and complement supplementation. Most linguistic and lexico-syntactic patterns are language-specific due to idiosyncratic characteristics of natural language.

## 6 Conclusion

Natural language strings in labels represent the input to a substantial range of linguistically based applications processing knowledge resources. The proposed correcting-compleitive method seeks to facilitate the processing of multilingual Semantic Web content by expanding shortened labels. Creating subterm structures of the processed labels specifically targets information extraction and string-based ontology alignment. Each of the four steps benefits from comparing content across languages and utilizing linguistic patterns. Replacing nonlexical symbols by their lexical equivalents and resolving compound ellipses especially when triggered by explicit syntactic markers achieved satisfactory results. Supplementing contextual complements and establishing subterm structures, however, strongly depend on the initial input and available additional natural language information, such as definitions. While the proposed method focuses on NLP activities, it also improves human readability, particularly by adding context to substantially shortened labels. As future work, a generalization of our methodology to other domains and its application to more languages could uncover new and interesting divergences as well as convergences of languages and contribute to an improvement in the representation and standardization of natural language content in ontology labels.

**Acknowledgments** The DFKI part of this work has been supported by the Monnet project (Multilingual Ontologies for Networked knowledge), cofunded by the European Commission with Grant No. 248458, and by the TrendMiner project, co-funded by the European Commission with Grant No. 287863.

---

<sup>5</sup><http://ontologydesignpatterns.org>.

## References

- Declerck, T., & Gromann, D. (2012). Towards the generation of semantically enriched multilingual components of ontology labels. In P. Buitelaar, P. Cimiano, D. Lewis, J. Pustejovsky, & F. Sasaki (Eds.), *Proceedings of the 3rd International Workshop on the Multilingual Semantic Web (MSW3)* (Vol. 936, pp. 11–23). Boston: CEUR.
- Declerck, T., & Lendvai, P. (2010). Towards a standardized linguistic annotation of the textual content of labels in knowledge representation system. In N. Calzolari et al. (Eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)* (pp. 3836–3839). Valletta, Malta: European Language Resources Association (ELRA).
- Ell, B., Vrandečić, D., & Simperl, E. (2011). Labels in the web of data. In L. Aroyo et al. (Eds.), *The Semantic Web: ISWC 2011* (Vol. 7031, pp. 162–176). Berlin/Heidelberg: Springer.
- Federmann, C., Gromann, D., Declerck, T., Hunsicker, S., Krieger, H.-U., & Budin, G. (2012). Multilingual terminology acquisition for ontology-based information extraction. In G. A. de Cea, M. C. Suárez-Figueroa, R. García-Castro, & E. Montiel-Ponsoda (Eds.), *Proceedings of the 10th Terminology and Knowledge Engineering Conference (TKE 2012)* (pp. 166–175). Madrid: TKE.
- Fliedl, G., Kop, C., & Vöhringer, J. (2007). From OWL class and property labels to human understandable natural language. In Z. Kedad, N. Lammari, E. Métais, F. Meziane & Y. Rezgui (Eds.), *Natural Language Processing and Information Systems* (Vol. 4592, pp. 156–167). Berlin/Heidelberg: Springer.
- Fu, B., Brennan, R., & O'Sullivan, D. (2012). A configurable translation-based cross-lingual ontology mapping system to adjust mapping outcome. *Web Semantics: Science, Services and Agents on the World Wide Web*, 15, 15–36.
- Gangemi, A., & Presutti, V. (2009). Ontology design patterns. In S. Staab & R. Studer (Eds.), *Handbook on Ontologies* (pp. 221–243). Berlin/Heidelberg: Springer.
- García, J., Montiel-Ponsoda, E., Cimiano, P., Gómez-Pérez, A., Buitelaar, P., & McCrae, J. (2012). Challenges for the multilingual web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 11, 63–71.
- Gherasim, T., Harzallah, M., Berio, G., & Kuntz, P. (2013). Methods and tools for automatic construction of ontologies from textual resources: A framework for comparison and its application. In F. Guillet, B. Pinaud, G. Venturini & D. A. Zighed (Eds.), *Advances in Knowledge Discovery and Management* (Vol. 471, pp. 177–201). Berlin/Heidelberg: Springer.
- Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P. F., & Rudolph, S. (2012, October). *OWL Web Ontology Language Primer* (Vol. 27). Recommendation. World Wide Web Consortium (W3C).
- Krieger, H.-U., Declerck, T., & Nedunchezian, A. K. (2012). MFO - The federated financial ontology for the MONNET project. In *Proceedings of the 4th International Conference on Knowledge Engineering and Ontology Development (KEOD-2012)* (pp. 327–330). Barcelona: SciTePress.
- Maynard, D., Funk, A., & Peters, W. (2009). Using lexico-syntactic ontology design patterns for ontology creation and population. In E. Blomqvist, K. S. F. Scharffe, & V. Svatek (Eds.), *Proceedings of the Workshop on Ontology Patterns (WOP2009)* (Vol. 516). CEUR-WS.org.
- Montiel-Ponsoda, E., Vila-Suero, D., Villazón-Terrazas, B., Dunsire, G., Escolano, E., & Gómez-Pérez, A. (2011). Style guidelines for naming and labeling ontologies in the multilingual web. In S. A. Sutton, T. Baker, M. Dekkers, D. I. Hillmann, M. Lauruhn, & J. Park (Eds.), *Proceedings of International Conference on Dublin Core and Metadata Applications (DC-2011)* (pp. 105–115). Dublin Core Metadata Initiative.
- Nguyen, T.-A., Power, R., Piwek, P., & Williams, S. (2013). Predicting the understandability of OWL inferences. In P. Cimiano, O. Corcho, V. Presutti, L. Hollink, & S. Rudolph (Eds.), *The Semantic Web: Semantics and Big Data* (Vol. 7882, pp. 109–123). Berlin/Heidelberg: Springer.
- Wimalasuriya, D. C., & Dou, D. (2010). Ontology-based information extraction: An introduction and a survey of current approaches. *Journal of Information Science*, 36, 306–323.

# A Multilingual Lexico-Semantic Database and Ontology

Francis Bond, Christiane Fellbaum, Shu-Kai Hsieh, Chu-Ren Huang, Adam Pease, and Piek Vossen

**Abstract** We discuss the development of a multilingual lexicon linked to the Suggested Upper Merged Ontology (SUMO) formal ontology. The ontology as well as the lexicon have been expressed in Web Ontology Language (OWL), as well as their original formats, for use on the semantic web and in linked data. We describe the Open Multilingual Wordnet (OMW), a multilingual wordnet with 22 languages and a rich structure of semantic relations. It is made by exploiting links from various monolingual wordnets to the English Wordnet. Currently, it contains 118,337 concepts expressed in 1,643,260 senses in 22 languages. It is available as simple tab-separated files, Wordnet-Lexical Markup Framework (LMF) or lemon and had been used by many projects including BabelNet and Google Translate. We discuss some issues in extending the wordnets and improving the multilingual representation to cover concepts not lexicalized in English and how concepts are stated in the formal ontology.

**Key Words** Multilingual • Ontology • Open data • Semantic lexicon • Wordnet

---

F. Bond (✉)

Nanyang Technological University, Singapore, Singapore  
e-mail: [bond@ieee.org](mailto:bond@ieee.org)

C. Fellbaum

Princeton University, Princeton, NJ, USA  
e-mail: [fellbaum@princeton.edu](mailto:fellbaum@princeton.edu)

S.-K. Hsieh

National Taiwan University, Taipei, Taiwan  
e-mail: [shukaihsieh@ntu.edu.tw](mailto:shukaihsieh@ntu.edu.tw)

C.-R. Huang

Hong Kong Polytechnic University, Hung Hom, Hong Kong  
e-mail: [churen.huang@polyu.edu.hk](mailto:churen.huang@polyu.edu.hk)

A. Pease

Articulate Software, San Francisco, CA, USA  
e-mail: [apease@articulatesoftware.com](mailto:apease@articulatesoftware.com)

P. Vossen

Vrije Universiteit, Amsterdam, The Netherlands  
e-mail: [piek.vossen@vu.nl](mailto:piek.vossen@vu.nl)

© Springer-Verlag Berlin Heidelberg 2014

P. Buitelaar, P. Cimiano (eds.), *Towards the Multilingual Semantic Web*,  
DOI 10.1007/978-3-662-43585-4\_15

## 1 Introduction

What do words mean and how are the words in different languages related? We make a start at answering these questions with a large multilingual lexical database and formal ontology. Each formalism captures knowledge about words and language in a different way. Linked together, they form a unified representation of knowledge suitable for language processing and logical reasoning.

An electronic lexicon is a fundamental resource for computational linguistics in any language, and Princeton English WordNet (PWN) (Fellbaum 1998) has become a de facto standard in English computational linguistics. WordNet represents meanings in terms of lexical and conceptual links between concepts and word senses. This allows us to model how concepts are represented in various languages. Ontologies offer a complementary representation where concepts are defined more axiomatically and can be formally reasoned with. The Suggested Upper Merged Ontology (SUMO) model of meaning (Pease 2011) addresses language-independent concepts, formalized in first- and higher-order logic. Bringing these two models together (Niles and Pease 2003) has resulted in a uniquely powerful resource for multilingual computational processes.

There have been a number of efforts to create wordnets in other languages than English. The EuroWordNet (EWN) project provided a first solution for also connecting these wordnets to each other by introducing a shared Interlingual Index (ILI) (Vossen 1998). The ILI was based on the English Wordnet (mainly for pragmatic reasons) and was considered as an unstructured fund of concepts for linking synsets across wordnets.

Most wordnets developed since EWN have used PWN as a common pivot to which each new wordnet is linked. This has the drawback of making English a privileged language and creating a certain linguistic bias. Since all languages have a different set of lexicalized concepts, it is not possible to have an interlingua where everything is lexicalized in all languages. A solution to this was proposed in the ILI using the union of synsets from all languages, arranged and related via the semantic links of PWN (Laparra et al. 2012). In this case, wordnets in the individual languages do not have to lexicalize all synsets but can still be linked together.

Another approach is to use a language-independent formal ontology—SUMO (Pease 2006)—as the common hub, which allows for the creation of arbitrary new concepts that can eventually encompass the union of lexicalized concepts in all languages. This has additional advantages such as a logical language for creating definitions of concepts that can be checked automatically for logical consistency and a much larger inventory of possible relations among concepts. Using the ILI as an intermediate approach collects and arranges synsets that are in need of formalization while deferring that effort to a later time. It is hoped that by cataloging these synsets, it should be possible to have some of the benefits of a common hub while speeding construction. This will likely be used as input to full SUMO-based formalizations in the future.

Currently, we are exploring both approaches in parallel—creating an ILI (not yet released) and extending SUMO (which has been released and is regularly updated).

A key organizational challenge for a true multilingual lexico-semantic database has been the large-scale nature of the effort needed. Each wordnet project has generally had its own funding and processes, even when coordinated in a broad sense with the original PWN. A variety of formats have proliferated. Wordnets do not all link to one another or a central ontology. Another challenge has been that some wordnets have not been released under open licenses and thus cannot be legally redistributed. This has greatly improved since the initial survey in Bond and Paik (2012) with many more wordnets being made open (Bond and Foster 2013). Some years ago, we introduced the idea of combining wordnets in a single resource<sup>1</sup> (Pease et al. 2008). This original vision has now been realized in the Open Multilingual Wordnet (OMW) described in Sect. 4. At the time of this writing, there are 22 wordnets that have been put into a common database format and linked to SUMO.

In the next section, we describe the Princeton Wordnet in more detail. We then introduce the linked ontology, SUMO (Sect. 3). In the next section, we describe how we built and made accessible the OMW: the main new resource described here (Sect. 4). Finally, we discuss how it can be extended to cover more languages better (Sect. 5).

## 2 Princeton English WordNet

Princeton WordNet (PWN: Fellbaum 1998) is a large lexical database comprising nouns, verbs, adjectives, and adverbs. Cognitively synonymous word forms are grouped into **synsets**, each expressing a distinct concept. Within each synset, words are linked by synonymy. Synsets are interlinked by means of lexical relations (among specific word forms) and conceptual relations (among synsets). Examples of the former are antonymy and the morphosemantic relation; examples of the latter are hyponymy, meronymy, and a set of entailment relations. The resulting network can be navigated to explore semantic similarity among words and synsets. PWN's graph structure allows one to measure and quantify semantic similarity by simple edge counting; this makes PWN a useful tool for computational linguistics and natural language processing.

The main relation among words in PWN is synonymy, as between the words *shut* and *close* or *car* and *automobile*. A group of synonyms—words that denote the same concept and are interchangeable in many contexts—is grouped into an unordered set. Synsets are linked to other synsets by means of a small number of **conceptual relations**, such as **hyperonymy**, **meronymy**, and **entailment**. Additionally, each synset contains a brief definition and, in most cases, one or more short sentences illustrating the use of the synset members. Word forms with several distinct

---

<sup>1</sup>[http://www.globalwordnet.org/gwa/gwa\\_grid.html](http://www.globalwordnet.org/gwa/gwa_grid.html).

meanings are represented by appearing in as many distinct synsets as there are meanings. Thus, each form-meaning pair (or **sense**) in PWN is unique.

### 3 Suggested Upper Merged Ontology

The SUMO<sup>2</sup> (Niles and Pease 2001; Pease 2011) began as just an upper-level ontology encoded in first-order logic. The logic has expanded to include higher-order elements. SUMO itself is now a bit of a misnomer as it refers to a combined set of theories: (1) The original upper level, consisting of roughly 1,000 terms, 4,000 axioms, and some 750 rules; (2) A Mid-Level Ontology (MILO) of several thousand additional terms and axioms that define them, covering knowledge that is less general than those in the upper level. We should note that there is no objective standard for what should be considered upper level or not. (3) There are also a few dozen domain ontologies on various topics including theories of economy, geography, finance, and computing. Together, all ontologies total roughly 22,000 terms and 90,000 axioms. There are also an increasing group of ontologies which are theories that consist largely of ground facts, semiautomatically created from other sources and aligned with SUMO. These include Yet Another Giant Ontology (YAGO) (de Melo et al. 2008), which is the largest of these sorts of resources and has millions of facts.

SUMO is defined in the Suggested upper Ontology-Knowledge Interchange Format (SUO-KIF) language,<sup>3</sup> which is a derivative of the original KIF (Genesereth 1991). It has been translated automatically, although in what is a necessarily very lossy translation into the W3C Web Ontology Language (OWL).<sup>4</sup> The translation also includes a version of PWN in OWL<sup>5</sup> and the mappings between them.<sup>6</sup>

SUMO proper has a significant set of manually created language display templates that allow terms and definitions to be paraphrased in various natural languages. These include Arabic, French, English, Czech, Tagalog, German, Italian, Hindi, Romanian, and Chinese (traditional and simplified characters).

SUMO has been mapped by hand to the entire PWN lexicon (Niles and Pease 2003). The mapping statistics are given in Table 1. There are a number of other approaches for mapping ontologies to wordnets (Fellbaum and Vossen 2012; Vossen and Rigau 2010). However, these have not involved ontologies that are either comparable in size or degree of formalization to SUMO.

<sup>2</sup>[www.ontologyportal.org](http://www.ontologyportal.org).

<sup>3</sup><http://sigmakee.cvs.sourceforge.net/viewvc/sigmakee/sigma/suo-kif.pdf>.

<sup>4</sup><http://www.ontologyportal.org/SUMO.owl>.

<sup>5</sup><http://www.ontologyportal.org/WordNet.owl>.

<sup>6</sup><http://sigma-01.cim3.net:8080/sigma/OWL.jsp?kb=SUMO> also provides a “live” generation of OWL one term at a time, where “&term=name” can be appended to the URL and the desired term name substituted for “name.”

**Table 1** SUMO WordNet mappings (115,261 total)

	Instance	Equivalence	Subsuming
Noun	9,837	3,329	68,919
Verb	0	600	13,150
Adj	724	540	14,771
Adverb	57	99	3,235
Total	10,618	4,568	100,075

## 4 Open Multilingual Wordnet

Wordnets have now been made for many languages. The Global Wordnet Association currently lists over 60 wordnets.<sup>7</sup> The individual wordnets are the result of many different projects and vary greatly in size and accuracy. The OMW (Bond and Paik 2012)<sup>8</sup> provides access to some of these, all linked to the PWN and SUMO. The goal is to make it easy to access lexical meaning in multiple languages. OMW has (1) extracted and normalized the data, (2) linked it to PWN 3.0, and (3) put it in one place. It includes a simple search interface that uses the SQL database developed by the Japanese Wordnet.

In order to make the wordnets more **accessible**, we have built a simple server with information from those wordnets whose licenses allow us to do so. It is based on a single shared database with all the languages in it. We only include data that is open: “anyone is free to use, reuse, and redistribute it—subject only, at most, to the requirement to attribute and/or share-alike.”<sup>9</sup>

The accessibility of the data means that it is becoming widely used. BabelNet 2.0,<sup>10</sup> a very large multilingual encyclopedic dictionary and semantic network, is made by combining the OMW, PWN, Wikipedia, and OmegaWiki (a large collaborative multilingual dictionary). Google Translate<sup>11</sup> also uses the OMW data.

The majority of freely available wordnets have been based on the **expand** approach, basically adding lemmas in new languages to existing PWN synsets (Vossen 1998, p. 11). These wordnets can easily be combined by using the PWN as a pivot. We realize that this is an incomplete solution, and a better one is discussed in Sect. 5.2. Some wordnets are based on the **merge** approach, where independent language-specific structures are built first and then some synsets linked to the PWN. For those merged wordnets in the OMW (Danish and Polish), only a small subset are actually linked, due more to lack of resources to link them than semantic incompatibility.

<sup>7</sup><http://globalwordnet.org/>.

<sup>8</sup><http://compling.ntu.edu.sg/omw>.

<sup>9</sup>Definition from the Open Knowledge Foundation: <http://opendefinition.org/>.

<sup>10</sup><http://babelnet.org/about.jsp>.

<sup>11</sup>[http://translate.google.com/about/intl/en\\_ALL/](http://translate.google.com/about/intl/en_ALL/).



Adding a new language to the OMW turned out to be difficult for two reasons. The first problem was that the wordnets were linked to various versions of PWN. In order to combine them into a single multilingual structure, we had to map to a common version. The second problem was the incredible variety of formats that the wordnets are distributed in. Almost every project used a different format and thus required a new script to convert it. In fact, different releases from the same project often had slightly different formats. These two problems mean that, even if a wordnet is legally available, there is still a technical hurdle before it becomes easily accessible.

The first problem can largely be overcome using the mappings from Daude et al. (2003). Mapping introduces some distortions. In particular, when a synset is split, we chose to only map the translations to the most probable mapping, so some new synsets will have no translations. For example, the synset `pwn16-legn:8` “a section or portion of a journey or course” in PWN 1.6 maps to two senses in PWN 3.0: `pwn30-legn:9` “a section or portion of a journey or course” and `pwn30-legn:8` “the distance traveled by a sailing vessel on a single tack”. `pwn16-legn:8` to `pwn30-legn:9` is the most probable mapping, so any lemmas associated with `pwn16-legn:8` will be associated only with `pwn30-legn:9`.

The second problem we have currently solved through brute force, writing a new script for every new wordnet we add. We discuss better possible solutions in Sect. 5.2. In the future, we hope people will move to a common standard for exchange, with Wordnet-LMF being the strongest contender (Vossen et al. 2013).

The server currently includes English (Fellbaum 1998); Albanian (Ruci 2008); Arabic (Black et al. 2006); Chinese (Huang et al. 2010; Wang and Bond 2013); Danish (Pedersen et al. 2009); Finnish (Lindén and Carlson 2010); French (Sagot and Fišer 2008); Hebrew (Ordan and Wintner 2007); Indonesian and Malaysian (Nurril Hirfana et al. 2011); Italian (Pianta et al. 2002); Japanese (Isahara et al. 2008); Norwegian (Bokmål and Nynorsk: Lars Nygaard 2012, p.c.); Persian (Montazery and Faili 2010); Polish (Piasecki et al. 2009); Portuguese (de Paiva and Rademaker 2012); Thai (Thoongsup et al. 2009); and Basque, Catalan, Galician, and Spanish from the Multilingual Common Repository (Gonzalez-Agirre et al. 2012).

The wordnets are all in a shared `sqlite` database with either Python or PERL CGI clients using the wordnet module produced by the Japanese Wordnet project (Isahara et al. 2008). The database is based on the logical structure of the PWN, with an additional language attribute for lemmas, examples, definitions, and senses. It is thus effectively a single open multilingual resource. We summarize the size of the wordnets and their coverage of **core concepts** in Table 2. Core concepts are the 5,000 synsets proposed as a core lexicon based on the frequency of the word forms in the British National Corpus (Burnard 2000) and an intuitive sense of salience (Boyd-Graber et al. 2006). That is, the core concepts are frequently occurring concepts (at least in British English).

**Table 2** Available wordnets

Wordnet	Lang	Synsets	Words	Senses	Core (%)	License
Albanet	als	4,676	5,990	9,602	31	CC BY 3.0
Arabic WordNet (AWN)	arb	10,165	14,595	21,751	48	CC BY SA 3.0
Chinese Wordnet (Taiwan)	cmn	4,913	3,206	8,069	28	wordnet
Chinese Open Wordnet	cmn	42,316	61,536	79,812	99	wordnet
DanNet	dan	4,476	4,468	5,859	81	wordnet
Princeton WordNet	eng	117,659	148,730	206,978	100	wordnet
Persian Wordnet	fas	17,759	17,560	30,461	41	Free to use
FinnWordNet	fin	116,763	129,839	189,227	100	CC BY 3.0
WOLF	fra	59,091	55,373	102,671	92	CeCILL-C
Hebrew Wordnet	heb	5,448	5,325	6,872	27	wordnet
MultiWordNet	ita	34,728	40,343	61,558	83	CC BY 3.0
Japanese Wordnet	jpn	57,179	91,959	158,064	95	wordnet
Multilingual Central Repository (MCR)	cat	45,826	46,531	70,622	81	CC BY 3.0
	eus	29,413	26,240	48,934	71	CC BY 3.0
	glg	19,312	23,124	27,138	36	CC BY 3.0
	spa	38,512	36,681	57,764	76	CC BY 3.0
Wordnet Bahasa	ind	51,755	64,948	142,488	99	MIT
	zsm	42,615	51,339	119,152	99	MIT
Norwegian Wordnet	nno	3,671	3,387	4,762	66	wordnet
	nob	4,455	4,186	5,586	81	wordnet
plWordNet	pol	14,008	18,860	21,001	30	wordnet
OpenWN-PT	por	41,810	52,220	68,285	79	CC BY SA 3.0
Thai Wordnet	tha	73,350	82,504	95,517	81	wordnet

We make available the synset-lemma pairs as tab-separated files, where they can be used by the Natural Language Toolkit<sup>12</sup> (Bird et al. 2009) as well as WordNet-LMF (Lexical Markup Framework: Vossen et al. 2013) and lemon (McCrae et al. 2011).<sup>13</sup>

Finally, we also make the SQL database available (with all languages except French and Basque, whose licenses are incompatible with the others). We use a simple database schema extended from the schema for the Japanese wordnet (Bond et al. 2009). When we use the combined database in applications, we typically use the database directly or through the Perl interface. Licenses that allow redistribution of derivative works allow people to make the entire lexicons available in any format, thus greatly improving their usefulness. There are also APIs for the database

<sup>12</sup>With the extensions that were added with the Japanese translation by Masato Hagiwara (Bird et al. 2010).

<sup>13</sup>Thanks to John P. McCrae for help in adding this.

produced by other researchers in Python, Java, Ruby, Objective-C, Gauche, and an alternative Perl module.<sup>14</sup>

There has been much research on making Wordnets available to the Semantic Web, including formatting as RDF (van Assem et al. 2006; Koide et al. 2006), serving LMF directly (Savas et al. 2010), or serving them through the lemon format (McCrae et al. 2011). Typically, these do not involve any changes in the actual content; the emphasis is instead on making it more easily accessible as Linked Open Data (Berners-Lee 2009). The proliferation of these approaches suggests that there is still some way to go until we will have an agreed-upon universal standard. Therefore, our approach has been to make our data open, clearly documented, well formatted, and validated in a simple format we use ourselves (tab-separated text) and some standard formats for exchange (LMF and lemon). This can then be straight-forwardly converted to whatever format is desired by those who want it in that format. Currently, in most of our use scenarios (principally word sense disambiguation and semantic processing), the latency of a Web interface is problematic—we expect that most of the users of our data will want to download the entire lexicon, and this is what we offer.

#### ***4.1 Possible Wordnet Structural Enhancements***

In this section, we will discuss some extensions people have suggested to the structure of the original PWN: these are not currently part of the open wordnet. One advantage of having many language-specific projects loosely coordinated is that there can be a wide variety of experimentation.

Our conversion scripts basically reduce each wordnet to a list of synset-lemma pairs, plus frequency, definitions, and examples if available. Everything is mapped to PWN 3.0 synsets. Therefore, the current version loses any synsets not in the English 3.0 wordnet. Many of the wordnets have such synsets, as well as metadata, definitions, examples, and other useful information. One of the ongoing goals of the OMW project is to make this information more easily accessible between projects.

We do not consider wordnets with licenses that do not allow redistribution, as we cannot legally include them. This includes some very well-constructed wordnets with excellent coverage, such as the Dutch,<sup>15</sup> German, and Korean wordnets (Vossen et al. 2008; Kunze and Lemnitzer 2002; Yoon et al. 2009). It is unfortunate that they cannot be integrated into the Open Wordnet. Some wordnets are built with their own structure and do not link to the PWN. These also cannot be included. Finally, some wordnets were not included even though they were open as the quality was still too

---

<sup>14</sup><http://nlpwww.nict.go.jp/wn-ja/index.en.html>.

<sup>15</sup>We are delighted to see that an Open Dutch Wordnet will be released soon (Vossen and Postma 2014) and will integrate it as soon the data is available.

poor due to the fact that they had been automatically made, with very little quality control.

Many of the wordnet projects extend the PWN relations in some way. For example, EWN defined many cross-part-of-speech links: *hammer*<sub>n:1</sub> is an *involved-role* of *hammer*<sub>v:1</sub> (Vossen 1998, pp. 97–110). Another instance of extensions is the Chinese Wordnet (Taiwan) which takes a different approach in representing lexical meanings. Unlike most models of lexical ambiguity resolution that assume only one meaning is chosen in a given context, it allows more than one (related) meanings to coexist in the same context. A lexical item is **actively complex** if it allows simultaneous multiple readings.<sup>16</sup> Meaning extensions thus are proposed to be distinguished between two types: **sense** and **meaning facet** (Ahrens et al. 1998). These can be distinguished as follows: given multiple possible meanings of a lemma, if a sentence that allows coexisting multiple readings for that lemma can be found, the distinction of these meanings is recognized as **meaning facet** distinction; otherwise, they are **sense** distinctions. The **coexistence test** for sense/meaning facet distinction can be illustrated in (1)–(4). The lemma *kànbìng* “seeing-sickness” in (1) allows two readings (“seeing the doctor” or “examining the patient”). The ambiguity can be resolved given more contextual information, and we cannot find a sentence that allows the coexistence of these two readings. Therefore, it is treated as two senses of that lemma. However, for the lemma *zázhì* “magazine,” it can refer to the physical object in (2) or the information contained in (3); more specifically, we can find a sentence like (4) in which the meaning of the lemma can refer to both the physical object and the information contained in that object. We therefore consider this meaning distinction of *zázhì* “magazine” is a **meaning facet** rather than a **sense**. Interestingly, among the 5,890 meaning facets being identified in Chinese Wordnet, 9 regular systematic patterns are extracted, which are similar to the regular polysemy (Apresjan 1973) (of complex types) proposed by Pustejovsky (1995). This fine-grained distinction is implemented by extending the types of semantic relations within the Chinese wordnet. Many (perhaps most) of these relations are not specific to Chinese. One of the advantages of the OMW is that we can look at research like this being done for one language and easily test its applicability to other languages:

- (1) 他正在 看病  
*tā zhèngzài kànbìng*  
 He PROG seeing-sickness  
 “He is seeing the doctor./He is examining the patient.”

<sup>16</sup>Note that according to psycholinguistic studies from Ahrens et al. (1998), there are two types of active complexity in natural language. The first is “triggered complexity” initiated by the speaker that involves puns; the second is “latent complexity” in which no pun or vagueness is intended. The Chinese Wordnet’s model focuses only on latent complexity.

- (2) 他手 上 拿 了 本 雜誌  
 tā shǒu shàng ná le běn zázhì  
 He hand on hold asp. CL magazine  
 “He is holding one magazine in his hand.”
- (3) 他 在 讀 那 一 本 雜誌  
 tā zài dú nà yī běn zázhì.  
 He PROG read that one CL magazine.  
 “He is turning the pages of the magazine and reading it.”
- (4) 他 拿 一 本 雜誌 給 我 看  
 tā ná yī běn zázhì gěi wǒ kàn  
 He takes one CL magazine give me read  
 “He passed me a magazine (to read).”

## 5 Extending the Multilingual Wordnet

In this section, we discuss the immediate plans to extend the wordnets to deal with multilingual issues. As was demonstrated in EWN, we can expect most languages to have concepts that are not lexicalized in English. In addition, there are still many concepts lexicalized in English, but not in PWN. Thus, different wordnets will have synsets that do not appear in most or even any other existing wordnet (this was the case for seven of the wordnets in the OMW). Consider the example of the Tagalog word *hilamos*—*to wash one’s face* (Borra et al. 2010).

Words such as this form part of the motivation for using a formal ontology. While some wordnets have used English as an interlingua and created phrases to stand in the place of otherwise unlexicalized concepts, another approach is to use SUMO as an interlingua which can contain concepts which stand for the lexicalized concepts of any particular language.

Exactly what counts as lexicalized can be hard to determine. Consider the following example: *foal* is lexicalized in English so must be in the English Wordnet. In Malay, the closest equivalent is a phrase: *anak kuda* “horse child” which can be produced compositionally by fully productive syntactic rules. In Japanese, it is *ko-uma* “child+horse” a word produced by a semiproductive process. So it is not clear whether the Malay wordnet should have an entry here. On the one hand, it is produced by a fully productive process. On the other, it is useful to have an entry, even if fully compositional, for completeness. We suggest that it should be entered but marked as syntagmatic using metadata, following the example of Italian, Basque, and Hungarian wordnets (Pianta et al. 2002; Pociello et al. 2011). Vincze and Almázi (2014) show how it is possible to exploit this metadata to automatically make two versions of the monolingual wordnets—one showing translation equivalents and one only showing concepts lexicalized in a particular language.

EWN distinguished a few types of nonuniversal lexicalizations and expressions, which call for different methods of handling:

**Cultural concepts:** Concepts that exist in some cultures and not in others, for example, Dutch *klunen*=*to walk on skates*.

**Pragmatic lexicalizations:** Concepts that are known in all/most cultures but are not considered lexicalized in all of these, for example, we all know the concept of a small fish, but Spanish happens to have a separate word for it *alevin*.

**Morphosyntactic mismatches:** Concepts that are lexicalized through words with different morphosyntactic properties across languages, for example, Dutch has no equivalence for *like* but uses the adjective *aardig*.

**Differences in perspective:** Some languages distinguish things depending on who is doing what to whom in ways that other languages don't, for example, *teach* and *learn* in English, whereas French uses *apprendre* for both.

A pertinent question is what defines a word and what defines a concept. Commonly occurring collocations may have transparent, compositional semantics, yet we may still consider these words. For example, noun compounds such as *sailing boat* are so common and ready-made that we consider them to be one word. Another point is that the relation between the components cannot be predicted from the structure: who is doing the sailing, who has the sail, and what is being sailed? A classical Dutch example is *kindermeel*: *meal for children* and *tarwemeel*: *flour made of oats*. From the structure, we cannot infer the relation. It needs to be learned or inferred, but Dutch speakers are probably not deriving them over and over again.

We are also extending the wordnets in terms of their size and coverage both within individual projects and by exploiting the disambiguating power of multilingual data to link to other open resources such as Wiktionary (Bond and Foster 2013). The core idea is that by looking at multiple translations of a concept, we can pinpoint the meaning exactly: *bat* in English is ambiguous between the sporting equipment and the flying mammal, but adding, for example, French, removes the ambiguity (*batte* vs. *chauve-souris*).

We are investigating two (compatible) methods of dealing with these new concepts. One is to create a concept in an external ontology and use this to link languages. In this approach, as *hilamos* is not lexicalized in English, it is not linked directly to English *wash* in the English wordnet. The fundamental value of the ontology is to define meaning using axioms in an expressive logic so that the meanings can then be manipulated without recourse to a human's intuition about the meaning of a word.

The second approach is to have a shared group of synsets for all languages, but not have them lexicalized in all languages. In this model, English *wash* and Filipino *hugas* are both lexicalizations of the same synset, and the synset for *hilamos* "wash one's face" inherits from this but would be marked as unlexicalized in English. Most **expand** style wordnets take this approach with nonlexicalized synsets being either just left blank or explicitly marked as nonlexicalized (as in, e.g., the MCR (Gonzalez-Agirre et al. 2012)).

## 5.1 Wordnets Linked to External Ontologies

Using ontologies<sup>17</sup> to link words (the first approach) is more labor intensive but offers other advantages.

Consider the notion of *earlier*. PWN has a synset for this word, but not a way to use it in temporal inference. SUMO however has a relation for earlier and a formal rule (among others) that allows an automated inference system such as those available with Sigma (Pease and Benz Müller 2013; Pease et al. 2010) to conclude that an interval that is earlier than another has an endpoint that precedes the start point of the following interval. This is a necessary and sufficient definition for *earlier* and uses the bi-implication or equivalence sign  $\Leftrightarrow$ :

```
(=>
  (earlier ?INTERVAL1 ?INTERVAL2)
  (before
    (EndFn ?INTERVAL1)
    (BeginFn ?INTERVAL2)))
```

Another example is the SUMO-based content developed to represent Muslim cultural concepts in Arabic Wordnet (Black et al. 2006). The *Udhiyah* ritual is performed during the period of Eid al-Adha and involves slaughtering a lamb by a Muslim. If a lamb has the attribute of being Udhiyah, then there necessarily exists an UdhiyahRitual in which it is the subject of the ritual:

```
(=>
  (instance ?UR UdhiyahRitual)
  (exists (?S ?EA ?P)
    (and
      (instance ?EA EidAladha)
      (during ?UR ?EA)
      (attribute ?S Udhiyah)
      (agent ?UR ?P)
      (attribute ?P Muslim)
      (patient ?UR ?S))))

(=>
  (attribute ?S Udhiyah)
  (exists (?UR)
    (and
      (instance ?S Lamb)
      (instance ?UR UdhiyahRitual)
      (patient ?UR ?S))))
```

Each of these symbols is further formalized, allowing them to be checked for logical consistency by automated theorem provers. This is also a key advantage for formal logic representation. The more expressive the representation and the more extensive the set of formalizations for each concept, the more things that can be checked automatically. A conventional dictionary must be checked by humans to ensure correctness of definitions. This is true with a conventional data dictionary, in which concepts in a database are defined in natural language in hopes of ensuring their correct usage. But when such a corpus of definitions grows large,

---

<sup>17</sup>It would be possible to link ontologies other than SUMO. There are other ontologies with at least partial links to wordnet, including DOLCE (Gangemi et al. 2003) and the Kyoto Ontology (Laparra et al. 2012). We only discuss SUMO here, as it is both the largest ontology and the most fully integrated with the OMW.

into the thousands or more, it is not likely that a human or even many humans will be able to find all inconsistencies. Automated means are needed. At that point, expressiveness also matters. In a taxonomy, the only error that can be caught automatically is the presence of a cycle in the graph. With a description logic, many more checks can be performed. In a higher-order language such as that used by SUMO, theorem proving (Benzmüller and Pease 2010) can find much more deep and subtle errors, leading to definitions of considerable depth and consistency.

Because SUMO terms are mathematical symbols, with a semantics given solely by their logical axioms, and unlike taxonomies or semantic networks, the symbol names can be changed without altering their meaning. In fact, the current Sigma browser can display terms with their names in different languages in order to emphasize this point and make them more accessible to logicians who may not speak English.

## 5.2 *Interlingual Index*

The second approach is basically that of the **Interlingual Index** (ILI: Peters et al. 1998). The variety of approaches in the EWN initially resulted in wordnets that were mapped to very different sets of concepts in the ILI. Likewise, only a small set of synsets could be traced to other languages through the ILI. To harmonize the output, EWN took two measures: (1) the definition of a shared set of (1,000 up to 5,000) Base Concepts that were manually aligned and (2) the classification of these Base Concepts using a small top ontology of 63 terms. Base Concepts (not to be confused with the “Basic Level Categories” of Rosch (1978)) represent synsets that have the highest connectivity to the other synsets. The top-ontology classification of these synsets provided a shared semantic framework. Each wordnet made sure the Base Concepts were presented properly in their language and manually mapped to the ILI. The minimal intersection across these wordnets through the ILI is thus the set of Base Concepts, but in practice the intersection is much larger. During the EWN project, it became clear that there are many problems with the ILI being based on PWN and that there are many possibilities to improve the ILI for linking wordnets (Vossen et al. 1999).

## 6 Conclusion

Several goals are being pursued in parallel: (1) research on building wordnets for individual languages, (2) research on building a more formal upper ontology, and (3) research on linking wordnets in many languages to make a multilingual resource. The ontology as well as some of the lexicons have been expressed in OWL, as well as their original formats, for use on the Semantic Web and in Linked Data. This effort builds on WordNet, Global Wordnet, and SUMO to create a rich Web of linguistic data and mathematically specified world knowledge.



## References

- Ahrens, K., Chang, L. L., Chen, K. J., & Huang, C.-R. (1998). Meaning representation and meaning instantiation for Chinese nominals. *International Journal of Computational Linguistics and Chinese Language Processing*, 3, 45–60.
- Apresjan, J. (1973). Regular polysemy. *Linguistics*, 142(5), 5–32.
- Benzmüller, C., & Pease, A. (2010). Progress in automating higher-order ontology reasoning. In B. Konev, R. Schmidt, & S. Schulz (Eds.), *Workshop on Practical Aspects of Automated Reasoning (PAAR-2010)*. Edinburgh, UK: CEUR Workshop Proceedings.
- Berners-Lee, T. (2009). Linked data-the story so far. *International Journal on Semantic Web and Information Systems*, 5(3), 1–22.
- Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python*. O'Reilly. [www.nltk.org/book](http://www.nltk.org/book).
- Bird, S., Klein, E., & Loper, E. (2010). *Nyumon Shizen Gengo Shori [Introduction to natural language processing]* (Hagiwara, Nakamura, & Mizuno, Trans.). Sebastopol: O'Reilly, Beijing, China.
- Black, W., Elkateb, S., Rodriguez, H., Alkhalifa, M., Vossen, P., Pease, A., et al. (2006). Introducing the Arabic WordNet project. In P. Sojka, K.-S. Choi, C. Fellbaum, & P. Vossen (Eds.), *Proceedings of the Third International WordNet Conference*, Jeju, Korea, 295–299.
- Bond, F., & Foster, R. (2013). Linking and extending an open multilingual wordnet. In *51st Annual Meeting of the Association for Computational Linguistics: ACL-2013*, Sofia (pp. 1352–1362). <http://aclweb.org/anthology/P13-1133>
- Bond, F., Isahara, H., Fujita, S., Uchimoto, K., Kuribayashi, T., & Kanzaki, K. (2009). Enhancing the Japanese WordNet. In *The 7th Workshop on Asian Language Resources* (pp. 1–8). Singapore: ACL-IJCNLP 2009.
- Bond, F., & Paik, K. (2012). A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, Matsue (pp. 64–71).
- Borra, A., Pease, A., Roxas, R., & Dita, S. (2010). Introducing Filipino WordNet. In P. Bhattacharyya, C. Fellbaum, & P. Vossen (Eds.), *Principles of Construction and Application of Multilingual Wordnets: Proceedings of the 5th Global WordNet Conference* (pp. 306–310). Mumbai, India: Narosa Pub.
- Boyd-Graber, J., Fellbaum, C., Osherson, D., & Schapire, R. (2006). Adding dense, weighted connections to WordNet. In *Proceedings of the Third Global WordNet Meeting*, Jeju.
- Burnard, L. (2000). *The British national corpus users reference guide*. Oxford: Oxford University Computing Services.
- Daude, J., Padro, L., & Rigau, G. (2003). Validation and tuning of Wordnet mapping techniques. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'03)*, Borovets, Bulgaria.
- de Melo, G., Suchanek, F., & Pease, A. (2008). Integrating YAGO into the suggested upper merged ontology. In *Proceedings of the 20th IEEE International Conference on Tools with Artificial Intelligence*.
- de Paiva, V., & Rademaker, A. (2012). Revisiting a Brazilian wordnet. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, Matsue.
- Fellbaum, C. (Ed.). (1998). *WordNet: An electronic Lexical database*. Cambridge: MIT Press.
- Fellbaum, C., & Vossen, P. (2012). Challenges for a multilingual wordnet. *Language Resources and Evaluation*, 46(2), 313–326. [Doi=10.1007/s10579-012-9186-z](https://doi.org/10.1007/s10579-012-9186-z).
- Gangemi, A., Guarino, N., Masolo, C., & Oltramari, A. (2003). Sweetening WordNet with DOLCE. *AI Magazine*, 24(3), 13–24.
- Genesereth, M. (1991). Knowledge interchange format. In J. Allen, R. Fikes, & E. Sandewall (Eds.), *Proceedings of the Second International Conference on the Principles of Knowledge Representation and Reasoning* (pp. 238–249). Los Altos: Morgan Kaufman.

- Gonzalez-Agirre, A., Laparra, E., & Rigau, G. (2012). Multilingual central repository version 3.0: Upgrading a very large lexical knowledge base. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, Matsue.
- Huang, C.-R., Hsieh, S.-K., Hong, J.-F., Chen, Y.-Z., Su, I.-L., Chen, Y.-X., et al. (2010). Chinese wordnet: Design and implementation of a cross-lingual knowledge processing infrastructure. *Journal of Chinese Information Processing*, 24(2), 14–23 (in Chinese).
- Isahara, H., Bond, F., Uchimoto, K., Utiyama, M., & Kanzaki, K. (2008). Development of the Japanese WordNet. In *Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech.
- Koide, S., Morita, T., Yamaguchi, T., Muljadi, H., & Takeda, H. (2006). OWL expressions on WordNet and EDR. In *AI Society Semantic Web Ontology SIG 13*, SIG-SWO-A601-03 (in Japanese). <http://www.jaist.ac.jp/ks/labs/kbs-lab/sig-swo/fpapers.htm>
- Kunze, C., & Lemnitzer, L. (2002). Germanet — Representation, visualization, application. In *LREC* (pp. 1485–1491).
- Laparra, E., Rigau, G., & Vossen, P. (2012). Mapping wordnet to the KYOTO ontology. In N. Calzolari, K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odiijk, & S. Piperidis (Eds.), *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC2012)* (pp. 2584–2589). Luxembourg: Publ. European Language Resources Association (ELRA).
- Lindén, K., & Carlson, L. (2010). Finnwordnet — wordnet påfinska via översättning. *LexicoNordica — Nordic Journal of Lexicography*, 17, 119–140. In Swedish with an English abstract.
- McCrae, J., Spohr, D., & Cimiano, P. (2011). Linking lexical resources and ontologies on the semantic web with lemon. In *The Semantic Web: Research and applications*, Springer Berlin Heidelberg, (pp. 245–259).
- Montazery, M., & Faili, H. (2010). Automatic Persian wordnet construction. In *23rd International Conference on Computational Linguistics* (pp. 846–850).
- Niles, I., & Pease, A. (2001). Toward a standard upper ontology. In C. Welty & B. Smith (Eds.), *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)* (pp. 2–9).
- Niles, I., & Pease, A. (2003). Linking lexicons and ontologies: Mapping WordNet to the suggested upper merged ontology. In *Proceedings of the IEEE International Conference on Information and Knowledge Engineering* (pp. 412–416).
- Nurril Hirfana Mohamed Noor, Sapuan, S., & Bond, F. (2011). Creating the open Wordnet Bahasa. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC 25)*, Singapore (pp. 258–267).
- Ordan, N., & Wintner, S. (2007). Hebrew wordnet: A test case of aligning lexical databases across languages. *International Journal of Translation*, 19(1), 39–58.
- Pease, A. (2006). Formal representation of concepts: The Suggested Upper Merged Ontology and its use in linguistics. In *Ontolinguistics: How ontological status shapes the linguistic coding of concepts*. New York: Mouton de Gruyter.
- Pease, A. (2011). *Ontology: A practical guide*. Angwin, CA: Articulate Software Press.
- Pease, A., & Benz Müller, C. (2013). Sigma: An integrated development environment for logical theories. *AI Communications*, 26, 9–97.
- Pease, A., Fellbaum, C., & Vossen, P. (2008). Building the global WordNet grid. In *Proceedings of the CIL-18 Workshop on Linguistic Studies of Ontology*, Seoul, South Korea.
- Pease, A., Sutcliffe, G., Siegel, N., & Trac, S. (2010). Large theory reasoning with SUMO at CASC. *AI Communications, Special Issue on Practical Aspects of Automated Reasoning*, 23(2–3), 137–144.
- Pedersen, B.S., Nimb, S., Asmussen, J., Sørensen, N.H., Trap-Jensen, L., & Lorentzen, H. (2009). DanNet: the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation*, 43(3), 269–299.

- Peters, W., Vossen, P., Díez-Orzas, P., & Adriens, G. (1998). Cross-linguistic alignment of wordnets with an inter-lingual-index. In P. Vossen (Ed.), *Euro WordNet* (pp. 149–251). Dordrecht: Kluwer
- Pianta, E., Bentivogli, L., & Girardi, C. (2002). Multiwordnet: Developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, Mysore, India (pp. 293–302).
- Piasecki, M., Szpakowicz, S., & Broda, B. (2009). *A Wordnet from the Ground Up*. Wrocław University of Technology Press. ISBN 978-83-7493-476-3. [http://www.plwordnet.pwr.wroc.pl/main/content/files/publications/A\\_Wordnet\\_from\\_the\\_Ground\\_Up.pdf](http://www.plwordnet.pwr.wroc.pl/main/content/files/publications/A_Wordnet_from_the_Ground_Up.pdf)
- Pociello, E., Agirre, E., & Aldezabal, I. (2011). Methodology and construction of the Basque wordnet. *Language Resources and Evaluation*, 45(2), 121–142.
- Pustejovsky, J. (1995). *The generative lexicon*. Cambridge, MA: MIT Press.
- Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization*, (pp. 27–48). Hillsdale, NJ, USA: Lawrence Erlbaum Associates. Reprinted in *Readings in Cognitive Science. A Perspective from Psychology and Artificial Intelligence*, A. Collins and E.E. Smith, editors, Morgan Kaufmann Publishers, Los Altos (CA), USA, 1991.
- Ruci, E. (2008). On the current state of Albanian and related applications. Tech. Rep., University of Vlora. <http://fjalnet.com/technicalreportalbanet.pdf>.
- Sagot, B., & Fišer, D. (2008). Building a free French wordnet from multilingual resources. In ELRA (Ed.), *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.
- Savas, B., Hayashi, Y., Monachini, M., Soria, C., & Calzolari, N. (2010). An LMF-based web service for accessing wordnet-type semantic lexicons. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias (Eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA).
- Thoongsup, S., Charoenporn, T., Robkop, K., Sinthurahat, T., Mokrat, C., Somlertlamvanich, V., et al. (2009). Thai wordnet construction. In *Proceedings of The 7th Workshop on Asian Language Resources (ALR7), Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics (ACL) and the 4th International Joint Conference on Natural Language Processing (IJCNLP)*, Suntec, Singapore.
- van Assem, M., Gangemi, A., & Schreiber, G. (2006). Conversion of wordnet to a standard RDF/OWL representation. In *Proceedings of The Fifth International Conference on Language Resources and Evaluation (LREC 2006)*.
- Vincze, V., & Almázi, A. (2014). Non-lexicalized concepts in wordnets: A case study of English and Hungarian. In *Proceedings of the 7th Global WordNet Conference (GWC 2014)*, Tartu (pp. 118–126).
- Vossen, P. (Ed.). (1998). *Euro WordNet*. Dordrecht: Kluwer.
- Vossen, P., Maks, I., Segers, R., & Van der Vliet, H. (2008). Integrating lexical units, synsets and ontology in the Cornetto database. In *LREC 2008*. Marrakech, Morocco: European Language Resources Association (ELRA).
- Vossen, P., Peters, W., & Gonzalo, J. (1999). Towards a universal index of meaning. In *Proceedings of ACL-99 Workshop, Siglex-99, Standardizing Lexical Resources*, Maryland (pp. 81–90).
- Vossen, P., & Postma, M. (2014). Open Dutch wordnet. In *Proceedings of the 7th Global WordNet Conference (GWC 2014)*, Tartu (presentation only).
- Vossen, P., & Rigau, G. (2010). Division of semantic labor in the global wordnet grid. In P. Bhattacharyya, C. Fellbaum, & P. Vossen (Eds.), *5th Global Wordnet Conference: GWC-2010*. Mumbai: Narosa Pub.
- Vossen, P., Soria, C., & Monachini, M. (2013). LMF - Lexical markup framework. In G. Francopoulo (Ed.), *LMF - Lexical markup framework*, Chap. 4. New York: ISTE Ltd + Wiley.
- Wang, S., & Bond, F. (2013). Building a Chinese wordnet: Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources*, Nagoya.
- Yoon, A., Hwang, S., Lee, E., & Kwon, H.-C. (2009). Construction of Korean wordnet KorLex 1.5. *Journal of KHISE: Software and Applications*, 36(1), 92–108.

# Multilingual Lexicalisation and Population of Event Ontologies: A Case Study for Social Media

Hristo Tanev and Vanni Zavarella

**Abstract** We describe a semi-automatic method for ontology-driven lexical acquisition and ontology population. Our method is language independent and weakly supervised and encompasses three distributional semantics sub-algorithms to learn semantic classes, modifiers and event patterns from an unannotated text corpus. The distributional features which our algorithms use are linear contexts, extracted without any language-specific resources, apart from a list of stop words. This makes our method applicable across different languages and domains. To illustrate the feasibility of our approach, we learned lexicalizations of concepts from the domain of natural disasters in Spanish and English. Then, we populated an event micro-ontology by performing event extraction from tweets published during several big tropical storms. The evaluation showed quite promising precision, while the event extraction recall could be improved further.

**Key Words** Concept learning • Event extraction • Event ontologies • Multilinguality • Ontology lexicalisation • Ontology population • Semantic class learning • Social media • Terminology extraction • Twitter

## 1 Introduction

The real-world Semantic Web needs to encompass many languages in order to reflect adequately the multilingual nature of the Web and to provide means for the development of the next generation of semantic information services. Multilinguality in this context can be achieved by linking monolingual ontologies across the Web and by building multilingual ontologies where concepts are aligned across languages.

Building the lexical layer of the Multilingual Semantic Web poses an important problem—how to acquire effectively and efficiently relevant lexica in many languages and domains. We think that the development of multilingual lexical learning

---

H. Tanev (✉) • V. Zavarella  
European Commission, Joint Research Centre, Ispra, Italy  
e-mail: [hristo.tanev@jrc.ec.europa.eu](mailto:hristo.tanev@jrc.ec.europa.eu); [vanni.zavarella@jrc.ec.europa.eu](mailto:vanni.zavarella@jrc.ec.europa.eu)

algorithms can provide a viable solution to this problem. The importance of lexical learning algorithms is underlined further by the fact that the language in use on the Web and in particular scientific and technical terminology is constantly evolving; thus, most online lexical resources need to be frequently updated.

Automatic knowledge acquisition techniques are important for a series of information extraction-related applications and also for automatic construction and population of ontologies. In this clue, these technologies are important for automatic building of Semantic Web services (Buitelaar and Cimiano 2008). In order to apply such approaches in the context of the Multilingual Semantic Web, efforts must be taken to make them work for more languages. Our vision is that knowledge extraction with domain-specific lexicalized surface grammars and heuristics, which is backed up by multilingual lexical learning, has more potential to be expanded across languages, rather than approaches based on generic parsers. One reason for this is that modelling syntactic structures requires linguistic knowledge, while the development of domain-specific dictionaries and surface parsing rules is less elaborated and can be done by more people, and consequently it can be carried out on a larger scale and even through crowdsourcing.

State-of-the-art ontology learning and population approaches have reached a certain level of maturity (Buitelaar and Cimiano 2008). However, they are still not widely applied to the vast majority of multilingual online data. The main reason is that the existing methods strongly depend on language-processing tools, such as part-of-speech taggers and parsers.

On the other hand, the importance of social media has been growing in recent years. Social networking sites, blogs, wikis, video sharing sites and folksonomies, commonly referred to as Web 2.0, gave birth to a new type of Internet culture in which the user becomes an active player rather than a passive consumer. People use Twitter, Facebook, LinkedIn, Pinterest, blogs and Web forums to give and get advice and share information on products, opinions and real-time information about ongoing and future events. In particular, Twitter with about 232 million registered active users as of the end of 2013 (Edwards 2013) was established as an important platform for the exchange of ideas, sharing of links to online content, updates about ongoing events, etc. It is noteworthy that Twitter was used during natural and man-made disasters and political crises for exchange of real-time information about situation developments. There is a significant amount of research regarding knowledge extraction and classification of tweets (Breslin et al. 2012). However, most of these approaches are not very relevant to the Semantic Web, since they ignore the linguistic and semantic structure of the text.

The Semantic Web is seen by some authors (Gruber 2008) as an important technology to bring more integration into Web 2.0, to increase the value of user-generated content and to establish the Web 2.0 sites as hubs of real collective intelligence. The approach proposed in this chapter automatically extracts event metadata from user-generated content. Such automatic metadata annotation can be regarded as a step towards the development of the Social Semantic Web, where the user-generated content becomes machine readable.

In order to bridge the gap between ontology learning and user-generated, multilingual online content, we propose a new language-independent algorithm for weakly supervised lexical acquisition. It comprises three sub-algorithms based on distributional semantics: a sub-algorithm for expansion of semantic classes, a sub-algorithm for learning modifiers and a sub-algorithm for learning event patterns. The novelty of our approach compared to the ones proposed by Carlson et al. (2010) and Riloff and Jones (2002) lies in the complete lack of any language-specific text processing. To compensate for the absence of language processing, we use a set of language-independent techniques which guarantee reasonable levels of accuracy. This makes our algorithm more relevant for use in the context of the Multilingual Semantic Web than other similar approaches.

In order to demonstrate the feasibility of our method, we used it for acquiring lexical knowledge in two languages and mapping this knowledge to an event-based micro-ontology from the domain of disaster management. Then, we perform event extraction from a corpus of tweets and populate the ontology with event instances detected in these tweets.

## 2 Related Work

Recently, different approaches for ontology learning and population have been proposed (for an overview, see Buitelaar and Cimiano 2008 and Drumond and Girardi 2008, among others). In particular, the Class-Example ontology population approach presented by Tanev and Magnini (2008) is relevant to our work. This method is based on the distributional similarity paradigm. The distributional similarity methods use Harris' distributional hypothesis, which states that words that occur in the same contexts tend to be semantically similar. It was shown that the Class-Example approach outperforms other state-of-the-art algorithms, such as the Hearst pattern approach. In another distributional similarity method Almuhabeb and Poesio (2008), the semantics of the distributional features was exploited. Völker et al. (2008) presented two approaches for learning of expressive ontologies: a lexical approach to generate complex class descriptions from definition sentences and a logical approach to generate general-purpose ontology constructs such as disjointness axioms.

Concept and pattern learning are strongly related to ontology learning. However, they have been used also outside of the ontology context. Relevant to our work is the concept learning algorithm described by Pantel and Lin (2002). This algorithm finds concepts as sets of semantically similar words. It uses distributional clustering by applying a novel clustering algorithm, called CBC (Clustering by Committees). A good example for information-extraction-related concept and pattern learning is presented in Riloff and Jones (2002). In this work, bootstrapping is introduced: as input the system obtains a handful of lexicalisations for each concept, and in every iteration it learns context patterns which are used in turn to obtain new concept lexicalisations. It was shown that this method succeeds in harvesting patterns

and semantic classes with good levels of accuracy. Recently, a new concept and pattern learning architecture for Never-Ending Language Learning (NELL) was developed and described by Carlson et al. (2010). The NELL system uses the Web to learn concepts and patterns. It exploits a bootstrapping algorithm which is running continuously as a Web-based learning agent.

A common feature of the approaches mentioned so far is that they rely on language-specific parsers and part-of-speech taggers and all of them work only for English. In contrast, we use statistical language-independent algorithms, based on surface features.

Integration between Web 2.0 and the Semantic Web was discussed in different papers: For example, Gruber (2008) suggests that Semantic Web technologies can help to extract new knowledge from the user-generated content.

There are many approaches for text mining from Twitter data (Breslin et al. 2012). Relevant to our approach are the methods for automatic event detection from Twitter like the one described by Reuter and Cimiano (2002). In contrast to the already existing approaches, we perform structured event extraction from the tweets and not only event detection.

### 3 Weakly Supervised Lexical Acquisition and Information Extraction for Building Event Ontologies

We propose an ontology of events where the root classes are *Event* and *ParticipatingEntity*. The class *Event* describes an event. It may have different subclasses. In this book chapter, we refer to event subclasses like *BuildingDamage*, *InterruptionOfCityService* and others, related to the domain of disaster management. Each instance of *Event* or one of its subclasses may be related via the *has-a-participant* property to one or more instances of *ParticipatingEntity* or one of its subclasses. For example, a *BuildingDamage* instance may be related to a *Building* instance. In the next section, we will describe in more detail an event micro-ontology for the domain of disaster management.

In order to build this type of event ontology, we adhere to the following procedure:

1. Manually define the top classes and the relations between them.
2. Learn terms which refer to the subclasses of *ParticipatingEntity*, e.g. *Building*, *EmergencyCrew*, etc. For this task, we use our multilingual semantic class learning algorithm. For example, for English, the algorithm will learn that *building*, *home*, *homes*, *houses* and others refer to the concept *Building* (including its subclasses). For Spanish, the algorithm will learn words like *edificio* and *casa*. The learned terms are used to discover mentions of participating entities (e.g. buildings) in the text. This lexical learning process can be viewed as related to concept learning, since it is the basis for the acquisition of subclasses of already defined concepts. In our example, *home* and *house* are two new concepts that can

be added to the ontology. However, additional processing, which goes beyond the scope of this book chapter, is needed to form concepts from the acquired lexica and to link concepts across languages.

3. Learn modifiers for the participating entities to recognize phrases describing these entities. For example, this algorithm will learn that *luxury*, *commercial*, *residential*, etc. are premodifiers of lexical items, belonging to the class *Building*. It will also learn postmodifiers like *in the city*, *of flats*, etc. In this way, the system can recognize phrases like *residential building* and *house in the city*.
4. Learn linear patterns which refer to each event subclass, defined by the ontology. As an example, consider the pattern *[BUILDING] was destroyed* for the event class *Building Damage*.
5. Using the lexical classes and patterns learned in steps 2, 3 and 4, we create a finite-state grammar to detect event reports and extract the entities participating in the reported events. We run this grammar on a text corpus and populate our event ontology with instances of the *Event* class and the related participating entities.

The suggested procedure learns in step 2 lexicalizations of subclasses of *ParticipatingEntity*. In step 4, it learns lexicalizations of *Event* subclasses in the form of one-slot linear patterns, e.g. *[BUILDING] was damaged*. While this is out of the scope of this chapter, the learned lexicalizations can be used to obtain new concepts via manual clustering and selection. As an example, consider some of the learned terms with highest score for the category *Building*—*hotel*, *mosque* and *church*; each of these terms represents a new subclass of *Building*. Moreover, *mosque* and *church* can be put under a new category *ReligiousBuilding*. The modifier lexicalizations, learned in step 3, can be used to manually acquire new attributes and possible values for these attributes. For example, our system has learned that *police*, *military* and *navy* are modifiers of the *EmergencyCrew* class. Considering these terms, a domain expert can define a new property for this class, *Institution-Of-Origin*, and define the previously mentioned terms as possible values for the new property. In this way, our approach provides means for expanding the event ontology structure and also to populate it (in step 5) with event instances.

The learning algorithms that are used in steps 2, 3 and 4 are implemented in the context of the multilingual lexical learning system *Ontopopulis++*, which is an extension of the *Ontopopulis* system (Tanev et al. 2009). We will describe in detail these algorithms in the following subsections.

The rules of the grammar built in step 5 are manually created. However, these rules are not language-specific, neither domain-specific. They just combine linearly the semantic classes and patterns learned in the previous steps.



### 3.1 *Learning of Lexicalisations of Semantic Classes*

The algorithm we describe here accepts as input a list of small seed sets of terms, one for each semantic class under consideration in addition to an unannotated text corpus. Then, the algorithm learns other terms that are likely to belong to each of the input semantic classes. Consider the following English language example: As an input, we give two semantic classes: *Building* and *Vehicle*. For *Building*, seed terms are *home*, *house*, *houses* and *shop*. For *Vehicle*, seed terms are *bus*, *train* and *truck*. On the output, the algorithm will return extended classes which contain additional terms like *cottage*, *mosque*, *property*, etc. for *Building* and *taxi*, *lorry*, *minibus*, *boat*, etc. for *Vehicle*.

Learning of several classes simultaneously requires more time and memory. However, our class learning algorithm uses the knowledge, gained for each class to boost the learning for the other classes: First, it downgrades the score of the learned features when they appear in more than one class. In this way, ambiguous features are downgraded. Second, when acquiring new terms, if a term is assigned to two or more classes and the score in one of the classes is much higher, then we delete the assignment of the term to the other classes.

The semantic class expansion algorithm has two main steps: (a) seed set expansion and (b) cluster-based term selection. The seed set expansion learns new terms which have similar distributional features to the words in the seed set. This is similar to the semantic lexicon expansion (Riloff and Jones 2002). However, our experiments show that this procedure alone does not guarantee good precision, especially when no language-processing tools are used. In order to improve the precision, we introduce a second term selection procedure. It uses clustering in the following way: the learned terms and the seed terms are clustered, based on their distributional similarity. Then, we consider only the terms which appear in a cluster, where at least one seed term is present. We call these clusters *good clusters*. This step is motivated by our observation that correct terms tend to form clusters in the distributional semantic space in which the seed terms are included. On the other hand, the irrelevant terms either do not enter into clusters or participate in clusters with other irrelevant terms.

The increased precision introduced by the cluster-based term selection allows for introducing bootstrapping in our process. The typical problem with bootstrapping in semantic class learning is the propagation of errors across iterations, called by some authors “semantic drifting.” We have two means to mitigate the effect of this phenomenon: First, the cluster-based term selection makes the semantic class learning relatively precise in each iteration. Second, during bootstrapping we pass the output of the algorithm as a new input for the seed set expansion. However, when performing cluster-based selection, we use the original seed set given by the user in order to check if a term belongs to a good cluster. In this way, we guarantee that the learning process will not explore areas of the semantic space which are too far from the original seed set. More formally, our semantic set expansion bootstrapping learning procedure is the following:

**input** : OriginalSeedSets- a set of seed sets of words for each considered class; Corpus- non-annotated text corpus; NumberIterations- number of the bootstrapping iterations  
**output** : Expanded semantic classes

```

CurrentSets ← OriginalSeedSets;
for i ← 1 to NumberIterations do
  CurrentSets ← SeedSetExpansion (CurrentSets, Corpus) ;
  CurrentSets ← ClusterBasedTermSelection (CurrentSets,
  OriginalSeedSets, Corpus) ;
end
return CurrentSets

```

In this algorithm, two sub-algorithms are called *SeedSetExpansion* and *ClusterBasedTermSelection*. The first one is the seed set expansion, and the second one implements the cluster-based term selection.

In the following paragraphs, we will describe these sub-algorithms in more detail.

### 3.1.1 Seed Set Expansion

Our algorithm is weakly supervised (Tanev and Magnini 2008). It accepts as input one or more sets of terms *CurrentSets*, each representing a semantic class as well as an unannotated text corpus. Then, it learns for each input set of terms additional terms which tend to appear in similar contexts as the terms from the corresponding input set. More formally, let's denote the list of semantic categories with  $(c_1, c_2, \dots, c_N)$ . For example, they can be *(Vehicle, Building)*. For each category  $c_i$ , a proper seed set is provided which we will denote with  $seed(c_i)$ . As an example, consider  $seed(Vehicle) = (bus, train, truck)$ .

Our algorithm has two main steps: (a) finding contextual features and (b) extracting new terms using contextual features extracted in step (a).

#### Learning Contextual Features

For each semantic class  $c_i$ , we consider as a *contextual feature* each uni-gram or bigram  $n$  which co-occurs at least three times in the corpus with any of its seed terms  $seed(c_i)$  (we have co-occurrence only when  $n$  is adjacent to a seed term on the left or on the right). A contextual feature cannot be composed only of stop words; we also do not consider words beginning with capitalized letters and numbers. These restrictions were introduced on the basis of empirical observations: For example, words with capitalized letters are usually names which tend to co-occur with particular terms, rather than term classes. Each contextual feature is assigned a score which shows how well it co-occurs with the seed terms. For example, some of the top-scoring left contextual features for *Vehicle* are *driver of*

*the, driving a, collided with a, and travelling in a*; some of the top-scoring right context features are *collided, parked, was stolen, and with registration*. Contextual features are assigned a score according to an algorithm described earlier by Tanev et al. (2009). We take for each category the top-scoring features and merge them into a contextual feature pool, which constitutes a semantic space where the categories are represented. The contextual features for each category  $c$  form a *context vector*  $v^{\text{context}}(c)$  which represents the semantics of  $c$  through its typical contexts. Each dimension in this vector is a contextual feature, extracted for any of the considered categories. If a feature does not co-occur with a category, then the corresponding coordinate will be 0; otherwise, it is equal to  $\text{score}(f, c)$ , which is calculated with our scoring algorithm.

### Learning New Terms

After contextual features are learned for each semantic category, our approach can extract new terms which tend to co-occur with these contextual features. To do this, we search in the text corpus the occurrences of the contextual features for each category. Then, we extract as a potential new term each word or bigram that does not contain a stop word and which is preceded immediately by a left contextual feature or followed immediately by a right contextual feature. By considering immediate contexts, we avoid using morphological analysis or parsing, unlike other approaches searching for nouns or noun phrases. Instead, the contextual features will ensure in most cases that the co-occurring term belongs to the right part of speech. For example, for the category *Vehicle* one of the left feature contexts is *driving a*; clearly, immediately after such a feature only a noun phrase can appear. The fact that we do not use any morphological analysis during term selection makes our algorithm applicable to non-standard languages, such as the ones used in Twitter and other social media.

Our algorithm represents each term  $t$  as a vector  $v^{\text{context}}(t)$  in the space of contextual features. The dimensions of this semantic space are all the extracted contextual features. The coordinate of a term  $t$  with respect to the dimension  $f$  reflects the co-occurrence trend between  $t$  and the contextual feature  $f$ :  $\text{cooccurrence}(f, t) = \frac{\text{freq}(f, t)}{\text{freq}(f, t) + 3} \cdot \text{PMI}(f, t)$ . The rationale behind this formula is similar to the one about the co-occurrence estimation between a feature and seed terms.

Our term learning approach calculates the relevance of a term  $t$  for a category  $c$ , using the following formula:

$\text{termscore}(t, c) = \frac{v^{\text{context}}(t) \cdot v^{\text{context}}(c)}{|v^{\text{context}}(c)|}$ . This relevance value is actually the projection of the term vector on the category vector. We found that in this case the projection works better than cosine similarity. It should be noted that the term vector  $v^{\text{context}}(t)$  does not represent the typical contexts of the term  $t$ , but it rather shows how the term co-occurs with the contextual features for the considered semantic classes. The algorithm completely ignores contextual features of  $t$  that are not related to

any of the categories. This makes the estimation of the term vector length quite unrealistic. This is the reason why cosine similarity deteriorates the results in this case. The cluster-based term selection algorithm, however, builds a less biased term vector representation for a subset of the terms.

### 3.1.2 Cluster-Based Term Selection

The algorithm clusters the top-scoring terms from the output of the previous algorithm and selects those which are in the same clusters as the original seed set given as input to the main algorithm for semantic class expansion. Let us denote the original seed set for a semantic category  $c$  as *OriginalSeedSet(c)*:

1. The highest scoring terms from the previous algorithm are considered for each semantic category. It is not possible to consider all the extracted terms, because of efficiency constraints.
2. The algorithm searches for each of these terms in the text corpus and extracts their left and right contextual features. Here, we run the already described feature learning algorithm. In this case, each term is considered to be in a separate category; therefore, we obtain an unbiased list of contextual features for each term.
3. Then, each term is represented as a binary vector of all the contextual features co-occurring with all the terms. A contextual feature is set to 1 for a term if it co-occurs with this term; otherwise, it is set to 0.
4. Term vectors are clustered using average-link agglomerative clustering with a cosine similarity function. The number of clusters is selected in such a way that the average similarity between the term vectors in each cluster is bigger than a certain threshold.
5. For each term  $t$  which was assigned to a class  $c$  by the seed set expansion algorithm, we check if it appears in a cluster with at least one member of *OriginalSeedSet(c)*. If so, then  $t$  is accepted as a term for the category  $c$ ; otherwise, not.

The term selection algorithm provides means to restrict the term sets, learned by the seed set expansion algorithm. In order to use the output of the seed set expansion algorithm without this selection, one needs to define a score threshold under which the terms should be ignored, because the terms with low score may be completely unrelated to the input semantic classes. Unfortunately, our experience with the seed set expansion algorithm shows that it is not possible to define a good threshold which works well in all cases.

Moreover, the score assigned to the terms by the seed set expansion is not always indicative about the relevance of these terms. Therefore, a term with high score may be irrelevant to the input seed set, while some of the terms with low score could be relevant. This is due to the fact that the seed set expansion algorithm does not consider all the contextual features of the terms. The cluster-based term selection makes more comprehensive feature extraction and puts together similar terms. This

compensates for the inconsistencies of the term scoring, introduced by the seed set expansion algorithm.

### 3.2 Learning Modifiers

A modifier is a phrase which is syntactically attached to another phrase and modifies its meaning. The scope of our modifier learning algorithm is, given a semantic category expressed as a set of terms, to find phrases which tend to modify the instances of this category. Modifiers usually express values of properties of the semantic classes. In this sense, modifier detection can be considered as part of the ontology learning process.

In order to model the modifier concept in the language of distributional semantics, we assume the following hypothesis:

*A phrase  $m$  is a modifier for the semantic class  $S$  if the context vector for  $S$  is similar to the context vector for the lexicalisations of  $S$  modified with  $m$ .*

For example, if class *Building* is represented through the terms *house*, *home*, *building*, *church* and *shop*, then *modern* can be considered a premodifier, since *modern home*, *modern shop*, etc. will probably share similar contextual features with the *Building* class.

To check the correctness of this hypothesis, we implemented a modifier detection algorithm based on this assumption, and we used it to extract modifiers for the semantic classes learned with the semantic class expansion algorithm.

The algorithm accepts as its input a set of semantic classes, represented as lists of terms and an unannotated text corpus. Then it performs three main processing steps:

1. Learning of potential modifiers. For this purpose, we extract the top 10,000 contextual features for the considered semantic classes, using the contextual feature learning algorithm.
2. In the text corpus, the system searches for sequences of the type

$$\textit{LeftContextualFeature}_1 \textit{LeftContextualFeature}_2 \textit{Term} \textit{and} \\ \textit{Term} \textit{RightContextualFeature}_2 \textit{RightContextualFeature}_1$$

where *Term* is a term from any of the input semantic classes and *LeftContextualFeature<sub>i</sub>* and *RightContextualFeature<sub>i</sub>* are left and right contextual features for the class to which *Term* belongs. All the contextual features that appear in this sequence at the place of *LeftContextualFeature<sub>2</sub>* or *RightContextualFeature<sub>2</sub>* are collected in a list of candidate modifiers where for each modifier we also memorize the contextual features (designated in the above formulae with index 1) which co-occur with it. These last features are considered contextual features of the modifier candidates.

3. For each candidate modifier, a weight is calculated which shows the similarity of the context vector of the semantic class lexicalisations modified by this candidate to the context vector of the class lexicalisations without this modifier. If this similarity is above a certain threshold, the candidate is accepted as a modifier. The weight of the contextual features in the vectors is calculated in a way similar to the weighting in the class expansion algorithm.

### 3.3 Learning Patterns

The objective of the pattern learning algorithm is to provide the user with means to acquire automatically in a weakly supervised manner a list of patterns which describe certain actions or situations. We use these patterns to detect event reports. Each pattern has a slot which should match a phrase referring to a semantic category. For example, the pattern *damaged a [BUILDING]* will match phrases like *damaged a house* and *damaged a primary school*. In the event extraction context, this pattern can be used to detect building-damage events, where the reference to the damaged building will match the slot of the pattern.

The algorithm accepts as its input (a) a list of action words, e.g. *damaged*, *damaging*, etc.; (b) representation of the semantic category for the slot as a term list, e.g. *house*, *town hall*, etc.; and (c) an unannotated text corpus.

As output, the user obtains a list of patterns like *[BUILDING] was destroyed*.

The main idea of the algorithm is the following: It finds patterns which are semantically related to the action, specified through the input set of action words, and on the other hand it will co-occur with words which belong to the semantic class of the slot. For example, *destroyed a [BUILDING]* is semantically similar to *damage* and also tends to co-occur with terms from the category *Building*. On the other hand, *built a [BUILDING]* is a co-occurrence pattern for *Building*, but it is not related semantically to *damaged*, and *injured* is semantically related to *damage*, but does not co-occur with *Building*. The algorithm performs the following three steps:

1. It finds terms similar to the list of action words, e.g. *destroyed*, *inflicted damage*, etc. We use the semantic class expansion algorithm to expand the seed set of action words into a bigger list.
2. Learns pattern candidates which co-occur with the slot semantic category (e.g. *Building*). We use the contextual feature extraction sub-algorithm of the class expansion. Each contextual feature of the slot class is considered a candidate pattern.
3. The algorithm keeps only the pattern candidates from the second step, which contain terms similar to the action words (discovered in the first step), and discards the others. In this way, only contextual patterns like *inflicted damage on a [BUILDING]* will be left. This is the output of the algorithm.

### 3.4 Detection of Event Reports Through a Finite-State Cascaded Grammar

We built a finite-state cascaded grammar to detect event reports, which combines semantic classes, patterns and modifiers learned with the previously described algorithms. The grammar is based on the Ex-PRESS formalism (Piskorski 2008). On the left-hand side of the Ex-PRESS grammar, rules are regular expressions over term classes or strings, recognized at the previous levels. On the right-hand side, each rule generates a new structure. Since the Ex-PRESS engine does not support unification, one can impose constraints on the structures via functional and logical operators, given on the right side. Our grammar has two levels: First, it detects the participating entity from a semantic category  $C$ , e.g. *Building*, through the rule

$$\begin{aligned} & (PreModifier(class : C, surface : S1)) * Term(class : C, surface : S2) \\ & (PostModifier(class : C, surface : S3))* \rightarrow ParticipatingEntity(class : C, surface : S), \\ & S = concatenation(S1, S2, S3) \end{aligned}$$

where the *Term*, *PreModifier* and *PostModifier* match terms, premodifiers or postmodifiers from category  $C$ , learned with the semantic class expansion and modifier learning algorithms described in the previous paragraphs. Second, event-specific actions or situations are detected with the second-level rules:

$$\begin{aligned} & LeftPattern(class : A)ParticipatingEntity(class : C, surface : S) \\ & \rightarrow ActionOrSituation(class : A, participant : S), C \in PossibleSlotFor(A) \\ & ParticipatingEntity(class : C, surface : S)RightPattern(class : A) \\ & \rightarrow ActionOrSituation(class : A, participant : S), C \in PossibleSlotFor(A) \end{aligned}$$

In these two rules, a structure representing an action or situation of type  $A$ , e.g. *BuildingDamage*, is generated when the action/situation pattern is detected next to a participating entity. *PossibleSlotFor(A)* returns a list of all the possible subclasses of *ParticipatingEntity* which may fill the slots of a pattern from class  $A$ . Then, event instances are generated from the detected action/situations. This may be done at different levels of complexity—from generating an event instance for each detected action/situation to clustering of actions and situations referring to the same event. However, in the context of short social media messages, usually an event description consists of one action/situation. We leave for future work the problem of integrating actions and situations into event instances in the context of larger texts or in clusters of texts.

## 4 Building a Disaster Management Micro-Ontology from Twitter Messages

To test the effectiveness of our machine learning algorithms, we created a micro-ontology related to disaster impact and management. Further, we acquired lexica for this ontology using our algorithms and populated the ontology from Twitter messages (tweets). In particular, the ontology models the following *ParticipatingEntity* subclasses: *Building*, *CityService* and *EmergencyCrew*. The class *Building* represents buildings, especially the ones whose damage may affect the normal life of the people. The class *CityService* is a broad class that encompasses transport services, health care, schools, public offices and shops. We assume that interruption or restoration of such services is of importance for the population of a city. Another class we model is *EmergencyCrew*—we consider emergency crews all the teams of professionals and volunteers who act during a disaster, usually in risky conditions with the purpose to mitigate the effects of this disaster. During a disaster, emergency crews include fire brigade, different types of technicians, ambulance crews, policemen and others.

In our micro-ontology, we also included some subclasses of *Event*, namely, *BuildingDamage*, *InterruptionOfCityService*, *RestorationOfCityService* and *DeploymentOfEmergencyCrew*. Each object from these event types is related to exactly one object from a subclass of *ParticipatingEntity* via *has-a-participant* relation. In particular, each *BuildingDamage* object is related to an object of class *Building*; each instance of *InterruptionOfCityService* or *RestorationOfCityService* is related to one instance of *CityService*; and each object of class *DeploymentOfEmergencyCrew* is related to only one participant in this micro-ontology. Consequently, the detection of an event report can be done by detecting an action or situation with one participant. We recognize these actions and situations through the action/situation extraction grammar discussed in the previous section.

It was out of the scope of the presented experiments to detect time and location of the events and to aggregate different event reports about the same event. However, these tasks are important for the automatic construction of event ontologies, and we consider them as a future direction for the development of our ontology learning and population approach.

## 5 Evaluation and Discussion

We ran our learning algorithms on two corpora of 1 million Spanish and English news titles, and we added to each corpus 220,000 tweets related to disasters. The tweets were both in Spanish (10%) and English (90%). Tweets were obtained using the Twitter Streaming API and disaster-related keywords in Spanish and English. In particular, we used the names “Bopha” and “Sandy” which were



**Table 1** Accuracy of lexical learning

Class	Seed set	Number learned	Accuracy (strict) (%)	Accuracy (lenient) (%)	Top modifiers (strict) (%)	Top modifiers (lenient) (%)
<b>English</b>						
Building	10	82	82	93	60	94
CityService	7	92	43	48	42	88
EmergencyCrew	6	52	86	86	42	88
<b>Spanish</b>						
Building	10	261	48	69	70	86
CityService	7	200	42	43	72	82
EmergencyCrew	8	80	68	68	50	98

names of typhoons that happened during our experiments; “flood” and “snow storm” are their derivatives and translations in Spanish. We also captured tweets which contained the hashtag #NJ and #NYC, which at the time of experiments mostly referred to news and situation reports from New Jersey and New York in the aftermath of the Sandy storms. Most of the tweets with these hashtags were in English, but there were also some Spanish. Using this corpus, we acquired lexicalizations for the classes of the disaster-related micro-ontology described in the previous section. We used two bootstrapping iterations. The following table shows the accuracy of the acquisition of the three subclasses of *ParticipatingEntity*, *Building*, *CityService* and *EmergencyCrew*, and of the top-scoring modifiers for each of these classes (Table 1).

The column *Number learned* reports the number of learned terms for each class. The column *Accuracy strict* reports the percentage of the terms which really belong to this class. *Accuracy lenient* reports the percent of the terms which belong to the class, or there is *part-of* or other strong semantic relation, so that the term can be used as a metonym to refer to the class in an indirect way. For example, if the text says that *a residential complex was destroyed*, this implies that some buildings were destroyed. In the same way, *the roof was damaged* means that the building was damaged. Regarding the modifiers, we evaluated the top 50 learned modifiers for each class. A modifier is considered as correct in the strict evaluation when it is an adjective, prepositional phrase or another phrase which has the role of syntactic modifier. On the other hand, some patterns are not modifiers, but they can be used to obtain phrases which belong to the same class. For example, the pattern *X and other buildings*, when filled at the position of *X* with a term referring to *Building*, will produce phrases like *villa and other buildings* which still refer to buildings. We consider such patterns relevant in the lenient modifier evaluation, together with the syntactic modifiers. For each event type, a series of action/situation patterns were also learned.

Using these resources, we constructed an event detection grammar and ran it on a mixed-language English and Spanish corpus of 270,000 test tweets which were obtained in a similar way as the training corpus. The system detected 544 events

**Table 2** Accuracy of event extraction

	Precision (%)	Recall (%)
English	85	23
Spanish	95	15

in English and 232 events in Spanish. We also created manually a small collection of event-reporting tweets to evaluate the recall of the grammars. It contained 30 English tweets and 60 Spanish ones. The results are shown in Table 2.

The results so far are encouraging regarding the precision. The errors in the precision evaluation came mostly from ambiguous action patterns which could be eliminated manually with minimal effort. The low recall was due to missing patterns; we believe it can be significantly improved with more bootstrapping iterations.

## 6 Conclusions

Our approach provides multilingual domain-independent means for ontology lexicalisation and population. In particular, we experimented with event-based annotation of Twitter messages related to disasters. The semantic layer, provided by the event annotation, can be used in the context of a Semantic Web service which provides real-time information about effects and ongoing recovery work from disasters. The presented results are encouraging and can be improved further. In our future work, we plan to experiment with integration of language-specific knowledge in our algorithms. We also think to experiment with multilingual lexicalisation and population of more complex ontologies in the area of disaster management and in other areas.

## References

- Almuhareb, A., & Poesio, M. (2008). Extracting concept descriptions from the web: The importance of attributes and values. In P. Buitelaar & P. Cimiano (Eds.), *Ontology learning and population. Bridging the gap between text and knowledge* (pp. 29–44). Berlin: Springer.
- Breslin, J., Ellison, N., Shanahan, J., & Tufekci, Z. (Eds.). (2012). *Proceedings of the 6th International Conference on Weblogs and Social Media (ICWSM - 12)*. Dublin: AAAI Press.
- Buitelaar, P., & Cimiano, P. (Eds.). (2008). *Ontology learning and population. Bridging the gap between text and knowledge*. Berlin: Springer.
- Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Estevam, R., Hruschka, J., & Mitchell, T. (2010). Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10)* (pp. 1306–1313). Atlanta, Georgia.
- Drumond, L., & Girardi, G. (2008). A survey of ontology learning procedures. In *The 3rd Workshop on Ontologies and Their Applications*, Salvador, Brasil (pp. 13–25).

- Edwards, J. (2013). Twitter is surprisingly small compared to a bunch of other apps and online companies. Retrieved from <http://www.businessinsider.com/twitter-user-base-compared-to-other-apps-and-online-companies-2013-11>
- Gruber, T. (2008). Collective knowledge systems: Where the social web meets the semantic web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6, 4–13.
- Pantel, P., & Lin, D. (2002). Discovering Word senses from text, In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 613–619). Edmonton.
- Piskorski, J. (2008). Ex-press - Extraction pattern recognition engine and specification suite. In *Finite State Methods and Natural Language Processing: 6th International Workshop, FSMNLP 2007* (pp. 166–183). Potsdam, Germany: Univesitätsverlag.
- Reuter, T., & Cimiano, P. (2002). Event-based classification of social media streams. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, (pp. 22:1–22:8). Hong Kong.
- Riloff, E., & Jones, R. (2002). Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI 99)*, Orlando, FL (pp. 474–479).
- Tanev, H., & Magnini, B. (2008). Weakly supervised approaches for ontology population. In *Ontology learning and population. Bridging the gap between text and knowledge* (pp. 129–144). Berlin: Springer.
- Tanev, H., Zavarella, V., Kabadjov, M., Piskorski, J., Atkinson, M., & Steinberger, R. (2009). Exploiting machine learning techniques to build an event extraction system for Portuguese and Spanish. *Linguamatica*, 2, 55–66.
- Völker, J., Haase, P., & Hitzler, P. (2008). Learning expressive ontologies. In *Proceedings of the 2008 conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge* (pp. 45–69). Amsterdam: IOS Press.

# **Part III**

## **Applications**

# Semantically Assisted XBRL-Taxonomy Alignment Across Languages

Susan Marie Thomas, Xichuan Wu, Yue Ma, and Sean O’Riain

**Abstract** The eXtensible Business Reporting Language (XBRL) has standardized the generation of and the access to financial statements like balance sheets, but language and XBRL-taxonomy diversity makes financial data integration across national borders and jurisdictions problematic. Integrating financial data in these circumstances requires that different multilingual jurisdictional taxonomies be aligned by finding correspondences between concepts. In this chapter, we outline a logic-based approach to this important alignment problem. The approach centers around the construction of an Accounting Ontology which, acting as a common denominator, is first used to enrich the semantics of ontologized XBRL taxonomies before reasoning is applied for alignment. Initial alignment experiments conducted on the French and Spanish balance sheets yielded 73.9% recall and 36.6% precision, but 100% precision, if redundant mappings are ignored.

**Key Words** Financial reporting • Multilingual • Protégé OWL • XBRL

## 1 Introduction

Financial reports, which contain various types of statements like balance sheets and earnings, inform interested parties about the current financial position of a company and the results of operations for a reporting period. The volume of such reports has become so enormous that automated processing has become a necessity. To meet this need, the eXtensible Business Reporting Language (XBRL) (Hoffman and Watson 2009) was developed and has been adopted worldwide by regulatory and

---

S.M. Thomas (✉) • X. Wu  
SAP AG, Karlsruhe, Germany  
e-mail: [susan.marie.thomas@sap.com](mailto:susan.marie.thomas@sap.com); [xichuan.wu@sap.com](mailto:xichuan.wu@sap.com)

Y. Ma  
Institute of Theoretical Computer Science, Technische Universität Dresden, Dresden, Germany  
e-mail: [mayue@tcs.inf.tu-dresden.de](mailto:mayue@tcs.inf.tu-dresden.de)

S. O’Riain  
Digital Enterprise Research Institute (DERI), NUI, Galway, Ireland  
e-mail: [sean.oriain@deri.org](mailto:sean.oriain@deri.org)

governmental organizations such as the US SEC,<sup>1</sup> the UK revenues and customs,<sup>2</sup> the European Financial Reporting authority,<sup>3</sup> and the individual European Business Registries.<sup>4</sup> Such authorities use XBRL to define taxonomies for the financial and business data that they are legally authorized to collect from the organizations or companies under their jurisdiction. An XBRL taxonomy specifies the content (concepts in XBRL terms) and structure of financial reports, which are created according to specific accounting regulations and conventions, which vary across country, jurisdiction, industry, time, etc. An XBRL taxonomy functions like an XML schema, in that its concepts are used to tag data in reports, so that the data can be automatically processed by software.

Much XBRL-based financial data is already available on the Web in multiple languages,<sup>5</sup> ready for use by interested parties, such as regulators, potential investors, creditors, competitors, and the general public. One Wired article<sup>6</sup> even suggests that automated monitoring of financial data by the public could perform a very important social service, namely, prevention of the waves of corporate fraud experienced in recent years. In order for authorities, financial analysts, and the public to fully benefit from the masses of XBRL data being made available, they will need to compare data from multiple jurisdictions, but language barriers and the diversity of XBRL taxonomies make it operationally difficult to do so. In Europe, for example, each member state has jurisdiction-specific rules for registering a company, publishing its bylaws, its annual financial statements, and other official documents. Accordingly, each national business register has defined its own sets of local taxonomies to be used by companies when filing and publishing their data as XBRL-instance documents. To achieve some cross-border comparability, the xEBR WG, with European Registers as members, has created an XBRL taxonomy of concepts widely shared in Europe and has aligned some national taxonomies to that (Verdin et al. 2012). But the alignment process is tedious, and verification of correctness is difficult.

Frankel (2009) warns that the lack of comparability threatens to undermine the very goals of XBRL. He analyzes the basic problem as a lack of semantic clarity, a problem which, in general, accounts for the bulk of software integration costs. The XBRL organization also acknowledges the problem and, in response, has recently formed the Comparability Task Force, which is collecting requirements around comparability.<sup>7</sup> This task force envisions a solution to the problem through the provision of cross-taxonomy correspondences, in the form of XBRL assertions

---

<sup>1</sup> See <http://www.sec.gov/>.

<sup>2</sup> See <http://www.hmrc.gov.uk/>.

<sup>3</sup> See <http://www.eba.europa.eu/Supervisory-Reporting/FINER.aspx>.

<sup>4</sup> See <http://www.ebr.org/>.

<sup>5</sup> See [http://www.xbrl.org/knowledge\\_centre/projects/list](http://www.xbrl.org/knowledge_centre/projects/list).

<sup>6</sup> See [http://www.wired.com/techbiz/it/magazine/17-03/wp\\_reboot?currentPage=all](http://www.wired.com/techbiz/it/magazine/17-03/wp_reboot?currentPage=all).

<sup>7</sup> See <http://www.xbrl.org/comparability-task-force>.

about relationships between comparable sets of elements in different taxonomies. Assertion creation is dependent, however, on taxonomy alignment.

This chapter proposes to align XBRL taxonomies using a four-phased process: (1) conversion of XBRL taxonomies to ontologies, (2) description of ontology concepts, (3) reasoning to infer mappings, and (4) mapping verification. Phase 1 automatically converts each XBRL taxonomy into an Ontology Web Language (OWL)<sup>8</sup> ontology. In Phase 2, the OWL ontologies are manually enriched and clarified by describing the concepts in each ontology using an *Accounting Ontology* which contains widely used accounting concepts. These enriched ontologies are the input to Phase 3, which automatically computes cross-taxonomy equality and subsumption relationships by means of logical reasoning services. Phase 4 presents the computed relationships to an expert for confirmation. The process is tested on pairs of taxonomies but could also be used for more than two.

The rest of this chapter is organized as follows. Section 2 introduces related work and compares our approach to it. Section 3 explains our alignment methodology and process and the taxonomies and ontologies used in the evaluation. Section 4 evaluates the proposed method and discusses some of its pros and cons. Finally, Sect. 5 summarizes the work and indicates the direction of future work.

## 2 Related Work

The proposed approach addresses a problem of ontology alignment by means of formalizing the accounting knowledge implicit in financial statements. In this section, the approach is first compared to existing work on alignment and then to work on formalization in the accounting and financial reporting domain.

### 2.1 Ontology Alignment

Ontology alignment, also called matching or mapping, is generally performed to integrate knowledge bases described by different ontologies. In our case, the different ontologies correspond to different XBRL taxonomies, and the knowledge bases correspond to collections of taxonomy instances, that is, financial statements, to be integrated or compared. The approach we take was developed specifically for this use case and takes advantage of some special features it has.

Our problem is a special case of multilingual ontology alignment, a field still in its infancy, as described in the survey by Trojahn, Fu, Zamazal, and Ritze (this volume). The solution we propose is unlike any of the approaches covered

---

<sup>8</sup>See <http://www.w3.org/TR/owl-features/>.

by that survey. Rather than using translation or corpus-based methods, it relies on logic and domain-specific features of the problem.

Alignment, as defined by Shvaiko and Euzenat (2013) in a recent analytical survey of the state of the art, is the operation of computing a set of correspondences, also called mappings, between two ontologies  $\mathcal{O}_1$  and  $\mathcal{O}_2$ . Each correspondence relates entities from  $\mathcal{O}_1$  to entities from  $\mathcal{O}_2$  and can be represented as a four-tuple  $\langle id, e_1, e_2, r \rangle$ , where  $id$  is an identifier,  $e_1$  is an entity from  $\mathcal{O}_1$ ,  $e_2$  is an entity from  $\mathcal{O}_2$ , and  $r$  is the relationship between  $e_1$  and  $e_2$ . Entities can be classes and properties, in general. However, the ontological form of an XBRL taxonomy, which is created in Phase 1 of our process, has no properties that need to be aligned, so that  $e_1$  and  $e_2$  are always classes and  $r$  is one of the OWL axiomatic relations applicable to classes: `subClassOf` or `equivalentClass`. This enables correspondences to be turned into axioms, like  $e_1 r e_2$ . The set of correspondence axioms constitutes a mapping ontology, which can be the input or output of a reasoner.

Our approach is entirely based on logic. By contrast, most of the state-of-the-art systems perform alignment by means of programmatic procedures called matchers (Shvaiko and Euzenat 2013), which generate correspondences by computing the similarity of an entity in  $\mathcal{O}_1$  to an entity in  $\mathcal{O}_2$ . Entity pairs which meet predefined criteria, for example, high similarity, are considered for inclusion in the alignment, that is, the set of correspondences between  $\mathcal{O}_1$  and  $\mathcal{O}_2$ . Current systems have three main kinds of matchers: (1) terminological, which compute similarity based on text associated with entities in the ontologies, for example, labels; (2) structural, which compute similarity based on relationships between entities, for example, the class hierarchies; and (3) extensional, which compute similarity based on individuals, which may be obtained from knowledge bases.

All of the systems compared in Shvaiko and Euzenat (2013) utilize different kinds of matchers, as well as different varieties of each kind. Most compute an overall match score as a weighted average of the outputs from multiple matchers. This score typically ranges between zero and one and can also be interpreted as the confidence that the match is correct. With some systems, the user is burdened with the task of specifying the weights for the matchers. Normally, the user also sets a threshold above which an entity pair is added to the alignment. A few systems have additional functionality that tests the generated correspondences to decide whether to discard or retain them. These tests may be ad hoc rules, as in the DSsim system (Nagy et al. 2006), or logic based as in the ASMOV system (Jean-Mary et al. 2009), which performs consistency checking on each correspondence as it is computed.

In contrast to current state-of-the-art systems, which primarily generate correspondences via matchers and sometimes refine them via logic, Phase 3 of our approach generates the correspondences entirely via logic, deducing them from the merge of  $\mathcal{O}_1$ ,  $\mathcal{O}_2$  and the Accounting Ontology. Assuming the concepts in  $\mathcal{O}_1$  and  $\mathcal{O}_2$  have been correctly defined in terms of the Accounting Ontology during Phase 2, the confidence score of every correspondence is one. Another novelty of our approach is that it generates subclass and superclass relations, in addition to the usual equivalent class relations generated by most systems today.



Noy (2004) takes a slightly different perspective on alignment by discussing the use of more than two ontologies as input to the alignment process. This paper categorizes alignment into two kinds of approaches: direct alignment, essentially the type already discussed above, and indirect. Indirect alignment uses a shared ontology as a common grounding for the two ontologies to be aligned. Ideally, these ontologies are extensions of the shared ontology, so that alignment profits considerably from the fact that they share common vocabulary with the shared ontology. Our approach is closer to the indirect category, with the Accounting Ontology serving as the shared ontology, which, in our case, is used to enrich the shallow semantics conveyed by XBRL taxonomies. Aleksovski et al. (2006) have shown that, in general, a shared ontology is helpful for aligning ontologies whose semantics is shallow. However, the approach there is quite different from ours in the way it builds connections between the shared ontology and the ontologies to be aligned; alignment is based on terminological matchers rather than precise logical definitions that enable mappings to be inferred.

Jiménez-Ruiz et al. (2012) take yet another perspective on alignment. Their system, LogMap, divides alignment into two phases: generation of mappings, followed by their refinement, that is, the elimination of logically implausible mappings. They generate mappings via terminological matchers and refine them by first detecting logical inconsistencies, if any, and then deleting or repairing mappings which have been identified as the cause of the inconsistencies. As discussed in Sect. 4, LogMap yielded very poor results when applied to our problem.

The YAM++ system of Ngo and Bellahsene (2012), like LogMap, has a mapping-generation phase and a logic-based refinement phase. In addition to terminological matchers, it has structural matchers based upon the well-known Similarity Flooding Algorithm. Furthermore, if training data are available, it can utilize supervised machine-learning algorithms to learn weights for matchers. Like most systems, it only generates equivalent class relations. Since it performed best overall on the financial benchmark in the 2012 Ontology Alignment Evaluation Initiative,<sup>9</sup> we decided to apply it to our use case. However, as reported in Sect. 4, like LogMap, it yielded poor results.

Spohr et al. (2011) describe a novel system which takes advantage of multilingual ontologies and which employs supervised machine learning to learn weights for matchers. It suits our particular problem well, since it yielded reasonably good results in the financial domain, and we have access to multilingual XBRL taxonomies and mappings with which to train it. Moreover, it is possible to train it to generate not only equivalent class relations but also subclass relations. Given all these points in its favor, we decided to evaluate our system against it. Indeed, as discussed in Sect. 4, it gave good results, although not quite as good as our logic-based method.

---

<sup>9</sup>See <http://oaei.ontologymatching.org/2012/>.

## 2.2 Accounting Ontologies

Recent work related to the formalization of accounting concepts can be divided into two categories. First, there has been a lot of work which converts XBRL reports into Semantic Web Representations: Declerck and Krieger (2006), García and Gil (2009), and Bao et al. (2010). While these efforts are useful for linking XBRL data to other data on the Web of linked data as discussed in O’Riain et al. (2011), there is little further semantic addition during the conversion process. This is also true of the XBRL-related ontologies listed in a recent survey of financial ontologies in O’Riain (2012).

In contrast, the second category of work focuses on a direct ontological specification of fundamental accounting concepts and processes. Krahel (2012) proposed the formalization of accounting standards as a means to discover and resolve inconsistencies and ambiguities in the standards. Gailly and Poels (2007) redesigned the resource, event, agent (REA) model, popular in the accounting literature, and formalized it in OWL. Chou and Chi (2010) proposed the EPA model (event, principle and account) as a way to model the correct accounting classification of business transactions. And Gerber and Gerber (2011) built a small OWL ontology as an experiment in formalization. Also, a financial ontology<sup>10</sup> has been designed to serve as the backbone of securities-trading and risk-management software, but it is too specific to these applications to be of use for alignment of financial statements.

Our research has a similar departure point as the second category. But, unlike existing work, which attempts to model the accounting process or the securities-trading process, we aim at a detailed characterization of the concepts in XBRL taxonomies in order to perform cross-taxonomy alignment. In spite of cultural and linguistic diversity, there are many concepts common to the XBRL taxonomies used in different countries. In general, these common concepts are finer-grained than the XBRL concepts, so that each XBRL concept can be described by means of multiple Accounting Ontology concepts, which it often shares with other XBRL concepts, even in the same taxonomy. Thus, our approach extends the semantics of each XBRL taxonomy. It makes explicit the fine-grained shared semantics which is only implicit in an XBRL taxonomy, often visible in labels or textual descriptions, but not available for machine processing. Moreover, our approach encodes this fine-grained semantics in such a way that logical reasoners can be used to infer mappings between taxonomies represented as ontologies. Although Li and Min (2009) propose the use of ontologies to extend the semantics of XBRL, they do not propose to do so in a methodical way for the purpose of enabling alignment.

Work which complements the Accounting Ontology is being done by Hoffman, often called the father of XBRL. He is developing an ontology that encompasses

---

<sup>10</sup>See <http://fadyart.com/en/>.

the big picture of financial reporting.<sup>11</sup> By contrast, the Accounting Ontology takes a microscope and reveals fine-grained concepts inherent in the concepts commonly found in financial statements.

### 3 Alignment Process and Evaluation Data Set

Figure 1 illustrates the four phases of the logic-based alignment process: conversion of XBRL taxonomies to ontologies, description of ontology concepts, reasoning to infer mappings, and mapping verification. Boxes with bold lines are phases that have been automated, whereas boxes with dotted lines normally need human intervention. Ideally, the manual work would be performed by an accountant, but for the experiment described in this chapter, the concept-description phase was done by a person proficient in ontologies, with the help of an accountant.

As indicated in the figure, the linchpin of the process is the Accounting Ontology, which we created (also with accounting advice) to add more fine-grained semantics to the very coarse-grained semantics of XBRL taxonomies. The ontology was developed in line with Semantic Web best practices, making heavy use of the value partition pattern.<sup>12</sup> Currently, it consists of 189 accounting concepts (classes), 36 object properties, and 5 data properties. An extract from its class hierarchy appears in the Protégé screenshot in Fig. 2. The subsections which follow give details about the process. The last phase is not described, because for our experiments it was automated as explained in Sect. 4. Each of the other phases is illustrated using examples drawn from the data set used for evaluation. Thus, this section explicates both the process and the data used in the evaluation.

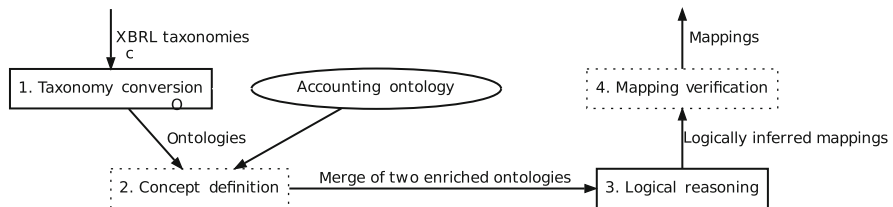
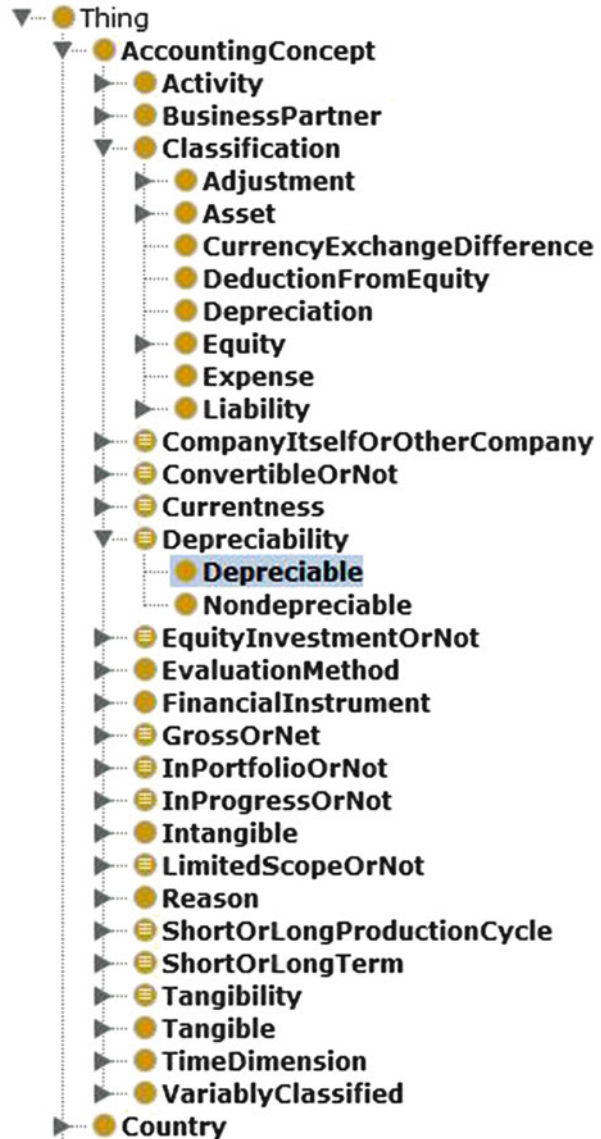


Fig. 1 Four phases of the proposed alignment process

<sup>11</sup>See <http://xbrl.squarespace.com/financial-report-ontology/>.

<sup>12</sup>See <http://www.w3.org/TR/swbp-specified-values/>.

Fig. 2 Accounting ontology: Protégé screenshot



### 3.1 Taxonomy Conversion

The inputs to *Phase 1* of the alignment process are the two XBRL taxonomies to be aligned. The process focuses on the XBRL monetary concepts, for example, concepts like Assets, which an XBRL taxonomy defines and which are used in XBRL instance files to tag an actual monetary value, for example, `<Assets>2000(/Assets)`. In addition, an XBRL taxonomy specifies calculation

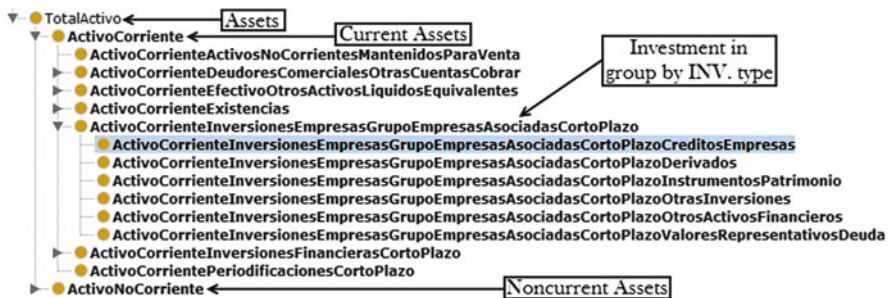


Fig. 3 Protégé screenshot of calculation hierarchy of Spanish balance sheet assets; parent calculated from children. “INV.” is short for investment

relationships between concepts, for example, the value of Assets is the sum of the value of Current Assets and Noncurrent Assets. These numeric relationships are usually expressed in what XBRL calls *calculation hierarchies*, in which a parent concept like Assets is computed by adding up, also called rolling up, its direct children, like Current Assets and Noncurrent Assets in the previous example.

In Phase 1, each taxonomy is first converted into RDF using the MONNET<sup>13</sup> *xblr2rdf* converter (Declerck et al. 2010), which preserves the calculation hierarchies. In the conversion process, each XBRL concept in a calculation hierarchy becomes an OWL class with the same URI as the XBRL concept. Given this one-to-one relationship, these classes are often referred to as XBRL concepts in the following explanations. In the next step of Phase 1, the calculation hierarchies are converted into OWL subclass relationships using SPARQL.<sup>14</sup> This conversion is done to facilitate the rapid addition of semantics to concepts in Phase 2.

Phase 1 was applied to the balance sheets of the French TCA and Spanish PGC2007 taxonomies. Before converting the taxonomies to ontologies, it was necessary to decide which entry point of each taxonomy to use. It is common for XBRL taxonomies to have multiple so-called entry points, each of which contains a different set of standard financial statements like balance sheets or income statements. The Spanish taxonomy has four such entry points. An accounting expert advised us to use the one called *Normal*. Similarly, the French has three independent entry points, of which we were advised to use the *Extended*, the idea being that these would be most comparable. Multiple entry points to taxonomies pose a well-known challenge to taxonomy processing, for which there is a recently proposed solution, in the form of a format for describing entry points (Allen 2012).

Figure 3 is a Protégé screenshot showing an extract of the output from Phase 1, namely, part of the class hierarchy corresponding to the calculation hierarchy for assets in the Spanish balance sheet. One of its numeric relationships, as indicated by the screenshot, is  $Assets = Current\ Assets + Noncurrent\ Assets$ .

<sup>13</sup>See <http://www.monnet-project.eu/>.

<sup>14</sup>See <http://www.w3.org/TR/rdf-sparql-query/>.

**Table 1** Properties for the Spanish concept highlighted in Fig. 3

pgc07:ActivoCorrienteInversionesEmpresasGrupo- EmpresasAsociadasCortoPlazoCreditosEmpresas
rdfs:subClassOf
(hasGrossOrNet some Net)
(hasClassification some Asset)
(hasCurrentness some Current)
(hasClassification some FinancialInvestmentAsset)
(investmentIn some GroupCompanyOrAssociate)
(hasFinancialInstrument some Loan)

### 3.2 Concept Description Using the Accounting Ontology

The purpose of the class hierarchies created in Phase 1 is to speed up *Phase 2*, which is a manual process in which each XBRL concept is described using concepts from the Accounting Ontology. These descriptions rectify the lack of semantic clarity, discussed in Sect. 1, by adding more fine-grained semantics to concepts. As mentioned, each class with its direct subclasses usually represents a computational roll-up (sum) in the XBRL world. This roll-up works by virtue of the fact that the concepts being rolled up share certain properties. For instance, the subclasses of Assets (*TotalActivo*) in Fig. 3 all share the property of being classified as assets. Rather than editing each concept individually to add this property, it is given just once to Assets, and is then inherited by all its subclasses, thus speeding up the process of enrichment. The procedure of adding shared properties is repeated for each class (roll-up). Moreover, care is taken that each sibling in a roll-up is differentiated from the others by means of properties. Siblings must be mutually disjoint (nonoverlapping); otherwise, they could not be added up to create a total. For example, Current Assets and Noncurrent Assets must be disjoint; otherwise, the total Assets would be incorrect, having double counted the overlap between the two addends. In this instance and in general, this disjointness is deduced from disjointness in the Accounting Ontology, for example, Current and Noncurrent are disjoint.

An example of the result of Phase 2 can be seen in Table 1, which shows the semantics added to the Spanish concept highlighted in Fig. 3. The additional semantics takes the form of property restrictions, which are represented in the figure using the Manchester OWL syntax.<sup>15</sup> Most of the restrictions for the Spanish concept under discussion are inherited from its superclasses. From Assets, it inherits the property restrictions (*hasClassification some Asset*) and (*hasGrossOrNet some Net*), as it happens that these are Net Assets only. Similarly, (*hasCurrentness some Current*)

<sup>15</sup>See [http://www.co-ode.org/resources/reference/manchester\\_syntax/](http://www.co-ode.org/resources/reference/manchester_syntax/).

is inherited from Current Assets, and (hasClassification some FinancialInvestmentAsset) as well as (investmentIn some GroupCompanyOrAssociate) are inherited from the superclass Investment in Group. On the other hand, we can see that (hasFinancialInstrument some Loan) was added directly and specifies the type of financial investment asset. In this case, the added components of meaning are also apparent in the original Spanish label for the highlighted concept, which can be translated as “current assets; short-term investments in group companies and associates; loans.”

Phase 2 was applied to the asset concepts in the French and Spanish ontologies created in Phase 1. This resulted in the description, or enrichment, of 94 French asset concepts and 74 Spanish, with each concept having 7 property restrictions on average. To differentiate concepts from different ontologies, the standard short form for a concept URI is used, that is, with *namespace prefix*<sup>16</sup> followed by *local name*. Prefix *ca* : indicates French concepts, prefix *pgc07* : Spanish; concepts from the Accounting Ontology have no prefix.

### 3.3 Inference of Mappings

Phase 3 of the alignment process is automatic. First, the concept descriptions created in Phase 2, which are called (*primitive classes*), are converted into definitions, that is, (*defined classes*). In this conversion process, disjointness among siblings is also added in order to detect inconsistencies. Finally, the two ontologies are merged, and the Hermit reasoner (Motik et al. 2007) is run to infer the mappings between them.

Phase 3 was applied to the French and Spanish ontologies enriched in Phase 2. An example result is that, given the concept definitions shown in Table 2, the mapping `ca:ActifCirculantNet rdfs:subClassOf pgc07:TotalActivo`

**Table 2** Example of concept definitions resulting in an inferred mapping

<code>ca:ActifCirculantNet</code>	<code>pgc07:TotalActivo</code>
<code>owl:equivalentClass</code>	<code>owl:equivalentClass</code>
<code>((hasDepreciability some Nondepreciable)</code>	<code>((hasDepreciability some Nondepreciable)</code>
<code>or (hasGrossOrNet some Net))</code>	<code>or (hasGrossOrNet some Net))</code>
<code>and (hasClassification some Asset))</code>	<code>and (hasClassification some Asset))</code>
<code>and (hasCurrentness some (Current or CurrentByException))</code>	
<code>rdfs:label "current assets, net"</code>	<code>rdfs:label "total assets"</code>

<sup>16</sup>See <http://www.w3.org/TR/REC-xml-names/>.

can be inferred. This inference is due to the fact that the subclass has all the restrictions of the superclass, plus one more, as can be seen from inspection of the table.

## 4 Evaluation and Discussion

We used a set of *gold-standard mappings* to evaluate our method against three other systems described in Sect. 2: LogMap, YAM++, and COAL. A French accountant, aided by a Spanish accountant, manually created the gold-standard mappings between the French and Spanish balance sheet ontologies, whose creation was described in Sect. 3.1. Initially, the mappings were expressed as the standard *exactMatch*, *narrowMatch*, and *broadMatch* of SKOS,<sup>17</sup> which we call *simple mappings*. But this way of thinking did not come naturally to the accountant, and it soon became clear that a more natural way of matching the taxonomies was the use of complex mappings. An example of a complex mapping is as follows:  $F_1$  is less than the sum of  $S_1$  and  $S_2$ , where “F” stands for French and “S” for Spanish.

We restricted our investigations to mappings involving the financial asset concepts only. The evaluation is also restricted to simple mappings, because, like most other existing approaches, our approach generates only simple mappings. There are in total 46 mappings (subsumption together with equivalence) that relate exclusively to asset concepts, as shown in Table 3. The mappings created by the accountant are directed, always going from French to Spanish concepts. To enable standard reasoners to operate on the mappings, we converted the *exactMatches* into equivalence relations and the *narrowMatches* and *broadMatches* into subsumption relations. To give LogMap and YAM++, which use monolingual string matchers, a fair chance, we used English translations of concept labels in the evaluation.

As shown in Table 4, our approach based on semantic enrichment and logical reasoning gives much better results than the other systems, that is, 73.9% recall

**Table 3** Statistics of gold-standard mappings

	Simple		Complex		Total
	Asset	Others	Asset	Others	
Subsumption	<b>32</b> (13+19)	25	7	4	68
Equivalence	<b>14</b>	13	13	13	53
Total	84		37		121

The 32 simple subsumptions consist of 13 *narrowMatch* mappings and 19 *broadMatch* mappings. Bold values highlight the 46 mappings that are considered in the rest of the paper for evaluation, which are simple and relate exclusively to asset concepts. That is, when we calculate precision/recall, we consider neither complex mappings nor the simple mappings which do not relate to asset concepts

<sup>17</sup>See <http://www.w3.org/TR/skos-reference/>.



**Table 4** Comparison of logic-based alignment with LogMap, YAM++, and COAL

	<i>exactMatch</i>	<i>narrowMatch</i>	<i>broadMatch</i>	Overall recall	Overall precision
LogMap	0 %(0/14)	–	–	0 %(0/46)	0 %(0/7)
YAM++	21.4 %(3/14)	–	–	6.5 %(3/46)	50 %(3/6)
COAL	21.4 %(3/14)	38.5 %(5/13)	15.8 %(3/19)	23.9 %(11/46)	3.1 %(11/352)
Logic based	85.7 %(12/14)	69.2 %(9/13)	68.4 %(13/19)	73.9 %(34/46)	36.6 %(34/93)

Recall for each mapping type is in columns 2–4, and the overall recall in column 5; only overall precision is given in 6

and 36.6 % precision. By contrast, LogMap produced seven mappings, all incorrect; YAM++<sup>18</sup> generated six mappings, only three of which are correct. Note that although there are *exactMatch*, *broadMatch*, and *narrowMatch* in the gold-standard mappings, LogMap and YAM++ can only generate *exactMatches*. COAL, however, can be trained to generate all three kinds of mappings, which we did, using Italian and Belgian taxonomies, plus mappings between them, as the training set. The mappings for training were derived from the mapping work done by the xEBR Working Group mentioned in Sect. 1. For each concept in the French balance sheet, we first trained COAL with a set of reference *exactMatches* and then used the trained COAL to get the top *exactMatch* concept from the Spanish balance sheet. We did the same for *narrowMatch* and *broadMatch*. This results in 352 mappings. In other words, COAL generates 352 candidate mappings, whereas the logic-based approach generates 93. As can be seen in Table 4, the logic-based approach produces much better results in terms of precision and recall for all three kinds of mappings. For example, out of 14 *exactMatches* from the gold-standard mappings, the logic-based approach finds 12, whereas COAL only finds 3.

The precision achieved, 36.6 %, seems a bit low. The cause of this is redundant mappings, as opposed to incorrect mappings. Redundant mappings are not in the gold-standard mappings, but can be inferred. Associated with Table 2, the mapping, `ca:ActifCirculantNet rdfs:subClassOf pgc07:TotalActivo`, is an example of a redundant mapping. It is redundant in the sense that it can be inferred from the following two gold-standard mappings:

```
ca:BilanActifNet owl:equivalentClass pgc07:TotalActivo
ca:ActifCirculantNet rdfs:subClassOf ca:BilanActifNet
```

If redundant mappings are ignored, the precision is 100 %. Recall is not 100 %, mainly due to two sources of difficulty. One is mappings involving “Other” concepts, which are catchalls for anything that does not fall into another sibling category. The other source of difficulty is divergent categorization, for example, the Spanish taxonomy includes prepayments to suppliers in the inventory category, but the French does not.

<sup>18</sup>The YAM++ of year 2012 does not require training.

The evaluation shows that the logic-based method outperforms the other state of the art systems. It, thus, represents a successful technique—albeit limited to financial reporting—that meets a pressing need identified in Trojahn et al. (this volume), namely, the need for novel matching techniques that work in multilingual environments. Our approach has a number of further advantages: detection of incorrect mappings, reduction of effort, and better division of labor. In the course of our experiments, we observed that alignment is a painstaking, error-prone process. It took months of part-time work by the accountant to create the gold-standard and a number of incorrect mappings crept into it. A positive aspect of our method is that it detected incorrect mappings, because they caused logical inconsistencies. Another positive aspect is that it can reduce the cost of alignment. Its main cost is the effort of describing concepts, something which takes only days or weeks for one financial statement, if the required concepts are in the Accounting Ontology. Yet another advantage is that the work can be divided between two experts, one expert for each financial statement to be aligned. This bypasses the problem of requiring one person to understand both statements. It was not easy for the French accountant to interpret the Spanish balance sheet, even though it had English labels, as well as Spanish. Using our method, Spanish and French experts would independently describe their respective balance sheets, a proceeding which should result in faster work with better accuracy.

On the negative side, Protégé is too foreign for most accountants, so in our experiment, the concepts were described with the help of an accountant, but not directly by an accountant. The same applies to the Accounting Ontology. This is, however, not an insoluble problem. Solomon et al. (2000) solved it for the medical domain by means of a language tailored to domain experts. An analogous solution could work for accounting. Moreover, the enormous progress made in standardizing medical terminology should serve as an example to the accounting community, spurring the creation of a more complete and correct Accounting Ontology than the one created for this experiment.

## 5 Conclusion and Future Work

As “defined” in an XBRL taxonomy, a concept has very little semantics explicitly represented and is ready to be harnessed. Logic-based alignment, as described in this chapter, preserves the semantics given in XBRL and adds significantly more semantics by manually defining each concept in terms of widely used accounting concepts. With the semantics made explicit, financial statements from different jurisdictions can be automatically aligned. An experiment with the French and Spanish balance sheets produced a recall of 73.9% and a precision of 36.6% or 100%, if redundant mappings are ignored. The next best evaluated system gave 23.9% recall and 3.1% precision.

The results are promising, but more work is necessary before the method can be used in productive systems. The main hurdle, which, as discussed in Sect. 4, is

not insurmountable, is the mismatch between accountants and existing tools like Protégé. New tools should speak the language of accountants and also enable them to express complex mappings in arithmetic terms, like the example at the beginning of Sect. 4. The ability to automatically infer such complex mappings is an interesting research challenge, one which might be tackled by introducing a third ontology as a hub (canonical) format. Automated removal of redundant mappings is another challenge. Some automation of the concept-description process also seems possible. In addition, the method needs to be extended to uniformly deal with the small number of concepts that are textual explanations related to monetary concepts in the financial statements. Another promising line of research is the combination of the logic-based method with matcher-based methods like COAL. Finally, real acceptance of the method might require it to be embedded in existing XBRL tools, where the semantics it adds could also be leveraged by functions which search for taxonomy concepts or which extend a taxonomy with new concepts.

Tools to aid the enlargement and maintenance of the Accounting Ontology are also highly desirable. Garnsey and Fisher (2008) suggest the possibility of using natural language processing (NLP) techniques to detect and identify new financial terms. These techniques could be adapted to discover concepts in financial statements, or the accompanying documentation, that might need to be added to the Accounting Ontology. NLP could also contribute to the automation of the concept-description process. Finally, the efforts for the Accounting Ontology and the defined Accounting Taxonomies would lead to a set of reliable data shareable as multilingual Linked Data. For a discussion of multilingual Linked Data, see Vila-Suero et al. (this volume).

**Acknowledgments** The work presented in this chapter has been funded in part by the EU FP7 Activity ICT-4-2.2 under Grant Agreement No. 248458, Multilingual Ontologies for Networked Knowledge (MONNET) project, and by the DFG Research Unit FOR 1513, project B1. We would especially like to thank the xEBR Working Group<sup>19</sup> for their help.

## References

- Aleksovski, Z., ten Kate, W., & van Harmelen, F. (2006). Exploiting the structure of background knowledge used in ontology matching. In *Proceedings of International Workshop on Ontology Matching (OM'06)*.
- Allen, P. (2012). Case study: Taxonomy packages - A simple specification to solve a universal problem. *Interactive Business Reporting*, 2, 32.
- Bao, J., Rong, G., Li, X., & Ding, L. (2010). Representing financial reports on the semantic web: A faithful translation from XBRL to OWL. In *Proceedings of International Symposium on Rules (RuleML'10)* (pp. 144–152).

---

<sup>19</sup>See <http://www.xbrleurope.org/working-groups/xebr-wg>.

- Chou, C.-C., & Chi, Y.-L. (2010). Developing ontology-based epa for representing accounting principles in a reusable knowledge component. *Expert Systems with Applications*, 37(3), 2316–2323.
- Declerck, T., & Krieger, H.-U. (2006). Translating XBRL into description logic. an approach using protege, sesame & OWL. In *Proceedings of International Conference on Business Information Systems (BIS'06)* (pp. 455–467).
- Declerck, T., Krieger, H.-U., Thomas, S. M., Buitelaar, P., O'Riain, S., Wunner, T., et al. (2010). Ontology-based multilingual access to financial reports for sharing business knowledge across europe. In J. Roóz & J. Ivanyos (Eds.), *Internal Financial Control Assessment Applying Multilingual Ontology Framework*. Kiadja a Memolux Kft., Készült a HVG Press Kft. nyomdájában.
- Frankel, D. S. (2009, June). XBRL and semantic interoperability. *Model Driven Architecture Journal*, 3 (5 pp.). <http://www.bptrends.com/bpt/wp-content/publicationfiles/SIX%2006-09-COL-MDA%20Journal%202009-06%20XBRL%20v01-00-%20Frankel.pdf>.
- Gailly, F., & Poels, G. (2007). Towards ontology-driven information systems: Redesign and formalization of the REA ontology. In *Proceedings of the 10th International Conference on Business Information Systems (BIS'07)* (pp. 245–259).
- García, R., & Gil, R. (2009). Publishing XBRL as linked open data. In *Proceedings of World Wide Web Workshop: Linked Data on the Web (LDOW'09)* (Vol. 538).
- Garnsey, M. R., & Fisher, I. E. (2008). Appearance of new terms in accounting language: A preliminary examination of accounting pronouncements and financial statements. *Journal of Emerging Technologies in Accounting*, 5, 17–36.
- Gerber, M. C., & Gerber, A. J. (2011). Towards the development of consistent and unambiguous financial accounting standards using ontology technologies. In *Proceedings of the International Conference on Accounting*.
- Hoffman, C., & Watson, L. (2009). *XBRL for dummies*. Hoboken: Wiley Publishing.
- Jean-Mary, Y. R., Shironoshita, E. P., & Kabuka, M. R. (2009). Ontology matching with semantic verification. *Journal of Web Semantic*, 7(3), 235–251.
- Jiménez-Ruiz, E., Cuenca Grau, B., Zhou, Y., & Horrocks, I. (2012). Large-scale interactive ontology matching: Algorithms and implementation. In *Proceedings of European Conference on Artificial Intelligence (ECAI'12)* (pp. 444–449).
- Krahel, J. P. (2012). On the Formalization of Accounting Standards (Ph.D. thesis, State University of New Jersey).
- Li, B., & Min, L. (2009). An ontology-augmented xbrl extended model for financial information analysis. In *Proceedings of IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS'09)* (pp. 99–130).
- Motik, B., Shearer, R., & Horrocks, I. (2007). Optimized reasoning in description logics using hypertableaux. In *Proceedings of the 21st Conference on Automated Deduction (CADE'21). Lecture Notes in Artificial Intelligence* (pp. 67–83).
- Nagy, M., Vargas-vera, M., & Motta, E. (2006). Dssim-ontology mapping with uncertainty. In *Proceedings of International Workshop on Ontology Matching (OM'06)* (pp. 115–123).
- Ngo, D., & Bellahsene, Z. (2012). Yam++ : A multi-strategy based approach for ontology matching task. In *Proceedings of International Conference on Knowledge Engineering and Knowledge Management (EKAW'12)* (pp. 421–425).
- Noy, N. F. (2004). Semantic integration: A survey of ontology-based approaches. *SIGMOD Record*, 33, 65–70.
- O'Riain, S. (2012). Semantic Paths in Business Filings Analysis (Ph.D. thesis, National University of Ireland, Galway).
- O'Riain, S., Curry, E., & Harth, A. (2011). XBRL and open data for global financial ecosystems: A linked data approach. *International Journal of Accounting Information Systems*, 13(2), 141–162.
- Shvaiko, P., & Euzenat, J. (2013). Ontology matching: State of the art and future challenges. *IEEE Transactions on Knowledge and Data Engineering*, 25(1), 158–176.

- Solomon, W. D., Roberts, A., Rogers, J. E., Wroe, C. J., & Rector, A. L. (2000). Having our cake and eating it too: How the galen intermediate representation reconciles internal complexity with users' requirements for appropriateness and simplicity. In *Proceedings of the AMIA Symposium*, American Medical Informatics Association (pp. 819–823).
- Spohr, D., Hollink, L., & Cimiano, P. (2011). A machine learning approach to multilingual and cross-lingual ontology matching. In *Proceedings of International Semantic Web Conference (ISWC'11)* (pp. 665–680).
- Verdin, T., Maguet, G., & Thomas, S. (2012). Promoting XBRL for cross-border data exchange by business registers in europe. *Interactive Business Reporting*, 2, 18–21.

# Lexicalizing a Multilingual Ontology for Searching in the Assistive Technology Domain

Gregor Thurmair

**Abstract** While even large ontologies are easy to search for experts, this is not the case for end users: They need guidance to find the ontology node which fits their search intentions best. The contribution describes a multilingual and multimodal front end to a database containing products of Assistive Technology, organized as a multilingual taxonomy (EASTIN, ISO 9999). In the mapping of a user query to the “best” ontology node, the variance observed in search requests must be guided to such nodes, which requires lexicalization. The key component is a multilingual terminological database, with its entries pointing to nodes in the taxonomy. The contribution describes its development (term identification), its representation in a lexicalized model, its integration into the natural language search component (including variant treatment and normalization) and its evaluation in the search context (coverage and usability). Problems of ontology lexicalization and localization, as discussed on the side of building the ontologies, are mirrored on the side of searching them.

## 1 The Application: Assistive Technology

### 1.1 EASTIN

Access to information on Assistive Technologies (AT) is a key issue in social participation and e-Inclusion. The UN *Convention on the Rights of Persons with Disabilities* (UN 2007) declares this a fundamental right; all UN member states are obliged to comply with this convention. To support people with disabilities, many states have organized web portals which provide information about Assistive Technology products. Portals are visited by doctors, physiotherapists and other persons in the domain.

In 2005, the major European AT information providers joined in creating the European Assistive Technology Information Network (EASTIN) (Andrich 2011;

---

G. Thurmair (✉)  
Linguattec, Munich, Germany  
e-mail: [gregor.thurmair@gmx.de](mailto:gregor.thurmair@gmx.de)

Gelderblom et al. 2011; Winkelmann 2011). EASTIN provides a portal ([www.eastin.eu](http://www.eastin.eu)) where people can access all databases of its national members simultaneously on a European level.

## 1.2 The Task

In order to open the scope of this portal for additional user groups (end users) and to support persons that are not familiar with the AT domain structure and only speak their native language, a natural language front end to the EASTIN portal was built.<sup>1</sup> This front end is supposed to be *multilingual* (users should forward information requests and receive results in their native language<sup>2</sup>) and *multimodal* (offering a speech channel). It should guide end users to the node in the ontology which points best to the products of their interest. Details are given in Thurmair et al. (2012).

## 1.3 Related Work

There is significant literature on the relationship between ontologies and terminology. Like ontologies, terminology deals with hierarchies of concepts, sometimes also called, “ontologies” (Madsen and Thomsen 2009; Madsen et al. 2010). Giunchiglia et al. (2006) show how such hierarchies could be converted into lightweight ontologies.

Gangemi and Presutti (2009) discuss design patterns for ontologies. One of these patterns consists of extracting ontologies from document sets (Khan et al. 2002); it uses linguistic processing both for identifying ontology nodes by term/concept extraction (Velardi et al. 2001; Tariq et al. 2003; Gillam et al. 2005; Eynard et al. 2012) and for the identification of relations between them (Aguado de Cea et al. 2009). Extensions into the multilingual field have been proposed for ontology localization whereby either a given ontology is translated (ontology localization; Espinoza et al. 2008; Fu et al. 2009) or two existing ontologies are mapped (ontology mapping; Trojahn et al. 2008, 2010; Al-Feel et al. 2013); both use conventional methods of translation (from dictionaries to machine translation).

In the current contribution, the focus is not in the creation of the ontology; this is done by ISO (ISO 2011), and the respective national bodies take care of localizing the labels into the participating languages. However, the identification of terminology and its linking to ontology node is also required in search, as the terminology used there must also point to relevant ontology nodes.

---

<sup>1</sup>In a project called EASTIN-CL, supported by the EC under ICT-PSP-2009-5-3, n° 250432.

<sup>2</sup>Supported languages are Danish, English, Estonian, German, Italian, Latvian and Lithuanian.

Instead, the current paper contributes to the creation of terminology and its link to ontologies in multilingual search. Abusalah et al. (2009) show that ontologies perform better than simple multilingual dictionaries, which focus the translations of words, which usually are ambiguous and can lead into semantically distant areas. In addition, to support the search, it has been proposed to enrich the ontology labels by additional terms, be it by adding WordNet-related terms (Khan et al. 2002), by adding words taken from target documents (Tomassen and Strasunskas 2009) or by using machine translation (Dragoni et al. 2013). These options would not be usable in the Assistive domain, however, because its special terminology is not covered by WordNet and standard MT systems and because there are no documents as target entities but only product descriptions, with very little natural language text. So the vocabulary for query expansion must come from a different source.

The maintenance of the multilingual terminology requires special representation structures. The standard approach is to separate the ontology nodes from their linguistic descriptions and create special representations, like in lemon (Montiel-Ponsoda et al. 2011), LexInfo (Cimiano et al. 2010) and others, with close links to the Lexical Markup Framework (LMF) standard (Francopoulo et al. 2006; ISO 24613). The present contribution follows these approaches by separating the ontologies from the linguistic descriptions; however, it uses only very limited linguistic descriptions.

## 2 Ontology in the Assistive Technology Domain

### 2.1 Organization of the Assistive Technology Domain

The purpose of creating ontologies is usually to improve the search, to allow for reasoning and to help structuring a domain. In Assistive Technology, the main focus is on search support.

Information in the AT domain is structured along the lines of the ISO 9999 standard (*Assistive Products for Persons with Disability—Classification and terminology*) (ISO 9999:2011). The structure of the ontology is not motivated by concepts but by product groups, as the history of the ISO 9999 states: *With the increasing volume of international trade in assistive products, a classification was necessary to facilitate location and selection of technical aids and to provide a consistent basis for product information, prescription guidelines, legal documents, information systems, catalogues, administration of stocks and for surveys and the production of statistics* (Heerkens et al. 2012). So the ontology in Assistive Technology (AT) is product driven. It is not based on documents but on AT products and their functional differences (like *electric vs. manual wheelchairs*). The nodes represent products with similar properties, and the links represent subsets (cf. Fig. 1).

The AT taxonomy has about 860 nodes, and it is organized in three levels. At each level (but mainly on the leaf nodes), there are product descriptions attached;



Numerical Designation	Title	Description	Related AbleData Terms
15	ASSISTIVE PRODUCTS FOR HOUSEKEEPING	Included are e.g., assistive products for eating and drinking.	
15 03	Assistive products for preparing food and drink	Included are, e.g., refrigerators and freezers.	
15 03 03	Assistive products for weighing and measuring	Included are, e.g., kitchen scales, diet scales, measuring spoons and cups, cooking and meat thermometers, butter dividers, timers and liquid level indicators.	Kitchen Scale Measuring Cup Measuring Spoon Tactile Cooking Thermometer Tactile Kitchen Scale Timer Utensil Organizer Voice Output Cooking Thermometer Voice Output Liquid Measure
15 03 06	Assistive products for cutting, chopping and dividing	Included are, e.g., slicing machines, knives, cutting boards, cheese slicers, egg dividers, egg slicers, onion holders and graters.	Bagel Slicer Cheese Slicer Chopper Citrus Reamer Citrus Zester Cooking Aid Cutting Aid Cutting Board Cutting Guide Grater Kitchen Knife Kitchen Scissors Knife Sharpener Large Handle Knife Melon Baller One Hand Strainer Utensil Organizer

**Fig. 1** Example of the ISO 9999 taxonomy: codes, title, description. The “Related AbleData Terms” were taken for the search vocabulary

more than 30,000 products are currently in the databases. Maintenance is done by an ISO committee, consisting of domain experts who watch the development in AT products and adapt the taxonomy accordingly. The latest version was released in 2011.

Strictly speaking, the AT taxonomy is at best a lightweight ontology, as it organizes concepts and (subsumption) relations between them. However, it lacks a formal description and any deduction framework. Recent attempts (Andrich et al. 2012) have moved towards an ontological description of the domain formalized in RDF.

However, the topic of the current contribution—end users’ searching in a structured domain—is not affected by the formal status of such structures. Its task is to find a given node using terminology which describes its semantic content. As has already been mentioned, this terminology cannot be extracted from the document sets linked to the respective nodes, as AT is a database of products with only few lines of natural language text. As a result, the term extraction techniques (Corcho et al. 2003; Velardi et al. 2001; Lopes et al. 2009) cannot be applied, due to the lack of (multilingual) corpus data. Because search is done based on the ISO codes of the taxonomy, not by using terminology, the link between search terminology and taxonomy nodes has to be an explicit step.

## 2.2 Indexing and Search

In EASTIN, indexing would be considered as the assignment of products to taxonomy nodes; the objective is to group products with similar properties under the same node. This assignment of products to nodes is done manually, by AT domain experts, together with the manufacturers of such products. The AT product descriptions are semi-structured objects, containing formal parts (like name of manufacturer, name of product, size, price, release date) and a short (one to two sentences) free-text description of the product, mainly explaining special features of the product. Such descriptions are available in some national languages (German, Italian, etc.) and in English (often machine translated).

As for search, the usual way of searching ontologies is by navigating to the “relevant” node in the ontology and collecting the documents which are linked to this node. This is also the way the ISO 9999 taxonomy is used; the EASTIN portal offers a means to navigate in the ontology to the relevant class and then fetch the product descriptions from there. It should be noted that the system does not do free-text search; it just returns all documents available under a given ISO code.

A side effect of this approach is that the search does not need to be cross-lingual but multilingual, as its target is an ontology node identifier (in AT terms: an ISO code). So German *faltbare Gehhilfe auf Rädern mit Sitz* points to *ISO 12 06 09* just as English *folding wheeled walker with seat* or Italian *girello deambulatore con sedile* does. So the ontology can be seen as a kind of (language-independent) interlingua, accessible from many languages, and no problems of query translation need to be faced (Abusalah et al. 2009).

However, while expert users are quite familiar with the taxonomy and know which ISO code to access, end users are not familiar with navigating in ontologies. Instead they need a search paradigm which they are familiar with, namely, entering key words, like in a search engine. Therefore, the challenge is to offer an option to search *in* ontologies, i.e. where the target object of the search is an ontology node, not a document set. The system should let them use familiar keyword technology to be guided to the relevant node in the structured domain.

## 3 Terminology in the EASTIN-CL Natural Language Front End

The goal of the EASTIN-CL project is to provide easier access to the AT domain. Usability tests made in EASTIN-CL have shown that occasional and nonprofessional (end) users have difficulties in accessing it by browsing in the ontology. While searching *with* the ontology means using an ontology node to access a document (or product) set, search *in* the ontology means identifying a relevant node, other than by browsing. Easier access therefore means easier searching *in* the ontology, before

being able to search *with* it. Indeed, the extension of accessibility of the AT ontology to other end users was one of the main aims of the EASTIN-CL project.

The challenge is then to let the users ask in natural language and guide the queries somehow to the “right” node in the ontology: *Gehwagen mit klappbarem Sitz* to *ISO 12 06 09*. This is the task of a query processing component (cf. Thurmair 2004). In order to do this, the terminology (search vocabulary) used in the AT domain needs to be provided. The key element here is the link between a term and a node (ISO code). So multilingual lexicalization (assigning terminology to an ontology node) and representation of the lexical information vis-à-vis the semantic structure of the ontology are key topics not just in ontology creation but also in ontology search.

The fact that the terminology is not used to *create* or define an ontology node but to *search* for it has consequences for the terminology collection: It should cover all possible *variants* which users may use for their searches, and it will contain *ambiguous* terms, pointing to *several* nodes.

In addition, this search facility should be offered in several languages (i.e. be multilingual) and in written and spoken form. The requirement of multilinguality also implies the retranslation of the retrieved documents into the users’ native language; therefore, machine translation components need to be added.

### 3.1 *Lexicalization for Search: Master Term List*

#### 3.1.1 **Term Collection**

The first challenge is to define the terminology which users may use for searching and to link the terms to the domain nodes:

1. The most straightforward way of collecting such terminology is by observing users at their searching, creating a corpus of user data and applying learning methods to link the used terms to the nodes retrieved. However, this approach presupposes that a natural language query processing is already in place, which was not the case.
2. Next, a *corpus-based* approach (extraction of terms from the document clusters attached to a given ontology node) could be envisaged (Walter et al. 2013). This approach was tried but abandoned, for the following reasons:
  - For some languages, no texts were available, or texts were not publicly accessible (copyright).
  - The amount of words found (>100 K candidates, many in general vocabulary) would be prohibitive for building a multilingual resource (i.e. for translating all of them). Most of them are general vocabulary, and real terms (domain relevant) would have to be extracted by manual inspection.
  - The link of the term candidates to the ISO codes would still have to be created, given the massive ambiguity of the term candidates: The terms themselves result from product descriptions; they contain information items which are

rather generic (like measures, prices, colours) and point to many ISO classes and descriptions of product features which differentiate the respective products from the others, which is no good terminology for searching.

- The product descriptions contain many terms which would even lead to wrong ISO assignments. It may be described as a product feature of a *work table* that people can drive underneath with a *wheelchair*, but users looking for *wheelchair* should not be guided to this product. Careful manual indexing can avoid such wrong links.

These facts are due to the specific nature of the “documents” in the AT domain: product descriptions.

3. A third option of collecting the terminology is combining different *keyword lists* which some AT portals already provide to support their users<sup>3</sup>:
  - There is the ISO 9999 index term list, containing the key terms of the classification, and a link to the respective code(s), available in English.
  - There are key term lists of the portals REHADAT (available in English and German) and HMI (in English).
  - There are terms in the AbleData system (however, originally without ISO codes), used as examples, clarification, etc. (cf. Fig. 1).

The project decided to base (the first version of) the AT terminology on these collections, integrate and harmonize them, link them to the ISO codes and translate them into six languages.

Most of the terms found were multiword terms. In searches containing multiword terms, two indexing strategies are possible (Buder et al. 1990): Pre-coordination collects multiwords *before* searching, and post-coordination collects them *afterwards* (usually by AND-ing the single terms). Nearly all search engines use post-coordination; however, it can easily be seen that in multi- and cross-lingual contexts, multiword terms must be identified beforehand,<sup>4</sup> as they may need a specific translation: If the parts of *stuffed bag seat* are each translated in isolation, the correct German translation into *Sitzsack* will not be found, and search results will suffer from this mistake. In EASTIN-CL, as the index contains many multiword terms and there is no free-text search usable for post-coordination, pre-coordination is selected for indexing; so the majority of EASTIN index terms are multiwords.

### 3.1.2 Term Variant Treatment

Montiel-Ponsoda et al. (2011) distinguish two types of term variants: *We identify two main groups of term variants: 1) term variants that are semantically coincident but formally different, and 2) term variants that are semantically and formally different.*

---

<sup>3</sup>cf. [www.abledata.com](http://www.abledata.com), [www.rehadat.de](http://www.rehadat.de), [www.hmi.dk](http://www.hmi.dk).

<sup>4</sup>The same holds for German compounds. Both term types are analysed as sequences of single words.

The first group contains graphical (*humor* vs. *humour*), inflectional (*rollator* vs. *rollators*) and morphosyntactic variants (*nitrogen fixation* vs. *fixation of nitrogen*). The second group contains stylistic, register, diachronic and other variants.

In EASTIN-CL, a policy was followed to have just one representative for one term if possible, i.e. to reduce the variants in the lexicon to a minimum. The reason was to keep the lists small as they needed to be translated into all participating languages, which is a significant effort. Therefore, the variants of the first group were not included in the term list but treated by special normalization components during query analysis in search. Such components included:

- Orthography normalization, including mapping of US to UK spelling: Only standard spelling (in English: UK spelling) and casing are represented.
- Lemmatization of inflected forms to base forms: Only base forms are represented.
- Normalization of hyphenations (*bath tub*, *bath-tub*, *bathtub* to *bath-tub*)
- Normalization of multiwords: *wheelchair*, *manual* and *wheelchair (manual)* to *manual wheelchair*: Only the “natural” word sequence is represented. (Moreover, search of multiwords is flexible wrt position of their parts.)

If the terminology lists only use normalized forms, then of course the runtime query analysis must be able to map non-normalized forms (e.g. US spelling) to their normalized correspondents (UK spelling) in order to be successful.

More complex forms of variants (the second group of variants mentioned above) have their own entries in the terminology and are treated as synonyms in the view of translation (i.e. have the same translation); in the view of the ontology, they just point to the same ISO code.

However, even after clean-up, there is significant variance in the denominations and room for improvement.

The final term master list contains about 12,700 concepts. All terms of this list were assigned one or several ISO codes (nodes) by domain experts. Some unclear cases are marked for later refinement.

### 3.1.3 Multilinguality and Localization

EASTIN-CL is a multilingual project. Consequently, the terminology of the assistive domain must be multilingual as well. There are two ways of organizing multilingual ontologies:

- In cases where the conceptual structure of the ontology is language independent, the approach would be to assign translations into different languages to the ontology nodes. This is the case in many technical domains, like engineering, biology, etc. Espinoza et al. (2008), Trojahn et al. (2008) or Montiel-Ponsoda et al. (2009) propose ontology localization on this basis.
- In other cases, the conceptual structure of the ontology itself is sensitive to language and culture issues, like in legal domains, tax systems, military and police organization ranks. In these cases, a separate ontology for each

language must be built, and the nodes of the different language systems must be linked explicitly. A framework for this type of ontologies is presented, e.g. in EuroWordNet (Vossen 2004).

EASTIN-CL follows the majority of approaches: As the domain contains mainly product descriptions (which, conceptually, is rather language independent), the approach was to have just one ontology and multilingual terms at each of its nodes.<sup>5</sup>

In EASTIN-CL, all master terms of the domain were translated into the participating languages. As standard tools (term banks, machine translation, etc.) proved to be insufficient due to the specific nature of the domain, translations were carried out by domain experts of the EASTIN-CL partners. In unclear cases, for instance, in cases where a term was ambiguous and could point to several ISO codes, the product databases themselves were consulted (esp. the images) in order to find the best translation. Quality control (like spellchecking) was added for each language list.

The resulting term list contains translations for all 12,700 concepts into seven languages, about 90,000 terms altogether. This list was converted into the Term Base eXchange (TBX) standard<sup>6</sup> and is publicly available in METASHARE.<sup>7</sup>

### 3.1.4 Representation

In systems where ontologies are populated with lexicalizations (e.g. Tanev and Zavarella 2014, in this volume; Trapman and Monachesi 2009), as well as with multilingual correspondences (Dragoni et al. 2013; Embley et al. 2011), special attention must be paid to the representation of conceptual vs. lexicon information. An overview of design patterns of the interface between lexical and ontology information is given in McCrae and Unger (2014) (in this volume).

Models like *lemon* (Montiel-Ponsoda et al. 2011), *LexInfo* (Cimiano et al. 2010), *LIR* (Montiel-Ponsoda et al. 2009) and others (Aguado de Cea 2012) develop formal models on how to represent linguistic information in the ontology domain, in close link to the LMF standard adopted by ISO (ISO 24613) (Francopoulo et al. 2006).

The key consideration is to make a distinction between the “domain” of ontology and ontology description on one side and the “domain” of lexicons and lexicon description on the other side. The link between the two sides is established by relations like *LexicalSense* (in *lemon*) and *hasSense* (in *LexInfo*). It links terms and ontology nodes in an n:m manner.

---

<sup>5</sup>Montiel-Ponsoda et al. (2009) comment: *This model has proven to be more suitable for highly specialized domain ontologies, e.g., in engineering or technical domains.*

<sup>6</sup>[www.ttt.org/oscarstandards/tbx](http://www.ttt.org/oscarstandards/tbx)

<sup>7</sup><http://www.meta-net.eu/meta-share>.

On the linguistic side, this procedure opens all kinds of options for linguistic descriptions, similar to standard lexicon entry representations (cf. the very detailed descriptions in Cimiano et al. 2010).

In EASTIN-CL, the representation tries to identify and store only the *minimal information* needed for the querying component in order to minimize the coding effort for 90,000 terms. As such, it can be seen as a very limited subset of more elaborate models like lemon.

The term master list is multilingual. It links the terms in different languages by means of an ID (and a common link to the ontology). This way, the whole resource can be used for querying and for translation (cf. León-Araúz and Faber 2014, in this volume). This is important as the retrieved documents need to be retranslated into the query language after search, and consistent terminology for query translation and document retranslation is a prerequisite for user acceptance.

In the search processing, there is one (monolingual) lexicon for each EASTIN-CL language, consisting of all entries derived from the terminology master list and enriched by linguistic information items. It should be noted that the majority of terms which need to be represented are *multiword* terms.

An entry, as used for query processing, has the following annotations:

- An *ID* to link the terms in different languages.
- The lemma in *display form*, to be used when it should be shown to the users: (da) *hjælpemiddel til hårvask*, (en) *swivel fork with built-up handles* and (de) *elektrischer Fausthandschuh, Beutel mit Rückstoßventil*.
- The lemma in *normalized form*: normalized and lower-cased spelling, multiword parts separated by semicolon: *hjælpemiddel;til;hårvask, swivel;fork;with;built-up;handles, elektrischer;fausthandschuh* and *beutel;mit;rückstoßventil*.
- A *list of lemmata* of which the entry consists; this includes lemmatization and decomposition steps for the languages involved: *swivel;fork;with;builtup;handle, elektrisch;faust;handschuh* and *beutel;mit;rück;stoß;ventil*. This is the key field in search as query and terms are mapped using single-word lemmata.
- A *list of part-of-speech* information for each term and each of its parts (in case of multiwords).
- A *list of ISO code* nodes to which the term points: *091808;091807*.

As for the linguistic annotations, the lexicon provides the term (in normalized form and in display form) and its part of speech; for multiwords, it provides the lemmata of its parts (each with normalized lemma, display lemma and part of speech). Models like lemon provide representations of all these annotations, allowing also for the description of word components (like multiwords); the EASTIN-CL list could be converted into such annotation frameworks, as the design of its terminological resource only uses *elementary* information items and provides

them in a flat tab-delimited form, so they can be assembled easily and mapped into the more elaborate lemon or LMF categories.<sup>8</sup>

## 3.2 Search

The natural language query component in EASTIN, designed for occasional and end users, consists of three steps: query analysis, search proper and result retranslation.

### 3.2.1 Query Analysis

Query analysis must map a query input to the closest index terms. The index terms are in turn annotated with ISO 9999 codes, pointing to groups of AT products. The challenge is to narrow down the variance of search and guide the query to the best one of the terms of the index.

Beyond the monolingual lexicon described above, auxiliary language resources are available for query analysis. In EASTIN-CL, two considerations influenced the design of these resources: (1) Query processing is a runtime component, i.e. it is time and resource critical (the maximal response time for the whole system is 2 s, so the query component can take only a fraction of that). (2) The EASTIN target vocabulary is limited and basically a fixed set: Not all query input words need to be processed but only the words of the term list.

Therefore, a “static” lemmatizer and a “static” decomposer resource were implemented whereby simply inflected forms point to their lemma (lemmatizer) or word parts (decomposer; by decomposing terms like *Thorako|lumbal|orthese*, to match a query for *lumbale Orthese*). So lemmatization and decomposition consist of simple and fast list lookup.

Query analysis does tokenization, normalization, lemmatization and decomposition. A list of candidate index terms is retrieved from the (single parts of the) input words. In case of no hit, a fallback distance-based similarity search is tried. As the whole front end is multilingual, this analysis sequence is built for each supported language. Query analysis fetches a set of ISO codes for each term identified in the user query from the index.

The final step in query analysis is ranking these candidate terms. Ranking is based on the number of words in the query, the number of words of the index terms and the number of matching terms. The terms with the highest overlap of matching terms are considered to be the best. The result is mapped on a 5-point scale, and the best ranked terms are returned with their ISO codes.

---

<sup>8</sup>As EASTIN-CL also supports spoken input, an additional resource had to be provided, in the form of a pronunciation lexicon for cases where the index terms were not in the system lexicon of the speech recognizer and not recognized by it.



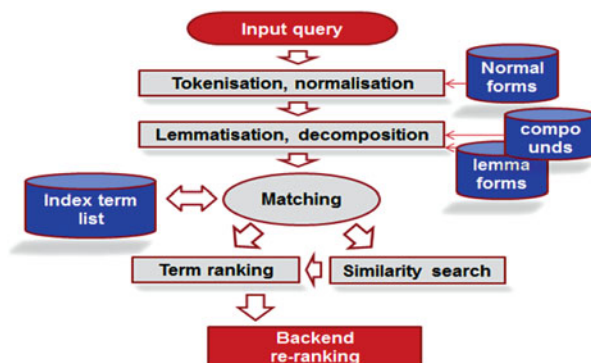


Fig. 2 Query analysis component and resources needed: seven languages

### 3.2.2 Back End Search and Retranslation

The search back end reorders the candidate list of the query processing as follows: While the query processing takes care of the best matching *index term*, the main search intention is to find the best *group of products*, i.e. the best matching ISO codes.

Therefore, the term list produced by the query analysis is re-ranked, and the highest ranked *ISO code* (not necessarily the highest ranked *term*) is offered to the users for confirmation and then used for searching the AT products. This makes the system more robust. The search interface displays which term contributed to which ISO code (cf. Fig. 3).

To avoid a situation where users find no hits, the EASTIN portal offers additional search options in addition to natural language input, like search by browsing in the ISO classification and search for products (*Tigges-Lumbal-Orthese*) or manufacturers (*All Terrain Wheelchairs Ltd*) containing the search term in their name.

Search is executed by a simultaneous access to all seven European databases linked to the EASTIN portal and a collection of the product lists returned by them, grouped under a given ISO code.

The product descriptions in the national EASTIN databases are stored in English; only some databases contain them in the national language. The multilingual front end now must retranslate the product descriptions from English into the query language. For this purpose, the EASTIN-CL front end provided machine translation web services. The MT systems were tuned for the AT domain by using the master term list and additional corpus data. Object of translation is the textual parts of the product descriptions (Figs. 2 and 3).

This way, a transparent search is enabled: Query and result presentation are in the users' native language, but the data searched for are in foreign language.

Zusammenfassung der Ergebnisse der Freitextsuche - Suche

Ihre Suche nach "Lumbalorthese" brachte das folgende Ergebnis:

Produktgruppen: 4

★★★★★ **Lumbo-sakrale Orthesen - ISO-Nummer: 06.03.06 - (253 Produkte)**  
Gefundene Schlagworte: Lumbalorthese, lumbale Orthese, Lumbalstützorthese

★★★★☆ **Lumbale Orthesen - ISO-Nummer: 06.03.04 - (8 Produkte)**  
Gefundene Schlagworte: Lumbalorthese, lumbale Orthese

★★★★☆ **Thorako-lumbale Orthesen - ISO-Nummer: 06.03.08 - (5 Produkte)**  
Gefundene Schlagworte: Thorakolumbalorthese, thorako-lumbale Orthese, Lumbalstützorthese

**Fig. 3** Search result for (German) *Lumbalorthese*. It shows three relevant ISO codes and the effect of decomposition—multiword handling in search. It also shows the re-ranking of search terms vs. the ranking of the ISO codes

## 4 Evaluation

Two types of tests were designed to evaluate (1) coverage (how well does the collected search vocabulary match the users' search intentions?) and (2) usability (how useful is the NL front end?). The test design is described in Gower et al. (2012).

### 4.1 Test Results on Coverage

In order to test terminology coverage, about 100 pictures of AT products were selected randomly and put online, asking users to enter the terms they would use to search for the type of products depicted on them. The terms which users used were analysed to find out (1) if they are in the search vocabulary and (2) if they would have pointed to the product group containing the picture.

This procedure avoids influencing users by proposing terms; it allows to verify that the terminology used by the EASTIN components is intuitive and of good coverage.

The tests of the term selection for pictures showed that the terms which users use lead to the right product group in the majority of the cases (63 %, with slight differences in the different languages); this emphasizes the good coverage of the term list (usual coverage in AT domain searches is 40–50 %, according to the domain experts).

Error analysis showed that this result can be further improved by adding synonyms and related terms to the term list. This would be an issue for future versions.

**Table 1** Results of usability tests

Query	Strongly agree (%)	Agree (%)	Neutral (%)	Disagree (%)	Strongly disagree (%)
It is easier to search products using the free text search	19	36	27	16	2
I prefer the free text search	14	37	24	25	0

Also, the test persons were AT experts. End-user tests could only start after the front end was released to the public (after the project end); so coverage tests should be repeated once the front end is in use.

## 4.2 Test Results on Usability

In order to evaluate the *usability of the approach*, users are given little tasks, and their interaction behaviour is evaluated by questionnaires: Do they succeed in their search? Which search tool do they use? Is MT of any help?, etc.

The results of the usability tests, performed with about 80 external users in six countries (languages), show a significant increase in the acceptance of the system, mainly due to the query functionality: More than 50 % of the testers consider the query component to be a very useful component (cf. Table 1).

Overall, the language technology front-end components are considered to be a significant improvement in the accessibility of Assistive Technology by the EASTIN portal.

## 5 Conclusion

In end-user front ends, the object of search is not the document base but the ontology itself. The objective of a search component is to find the “best” matching node in the ontology. Lexicalization of nodes for search, as well as localization in multilingual contexts, differs from their counterparts on the ontology production side as it needs a broader coverage to cope with all possible search term variants: Users must get hits for queries containing all kinds of terms.

Due to the specific nature of its target documents (product descriptions), the lexicalization approach in the assistive domain consists in collecting available key term lists, in linking them to the nodes of the taxonomy. Beyond the localization of the taxonomy itself, these term lists must also be made available for all participating languages, as they must support the retranslation of retrieved documents.

Localization was done by domain expert translators, due to the specificity of the terminology.

The representation of the term lists follows the approach used, e.g. in lemon to separate domain knowledge from linguistic knowledge, with only a minimum of linguistic annotations given to the terms.

At runtime, user queries must be mapped to the nodes of the taxonomy by means of the terms known to the system; this implies all kinds of normalization, from spelling to decomposition; such operations are language dependent. As most terms are multiwords, multiword support is an essential feature.

While tests have shown that (end) users appreciate the NL front end, further research would be required to adapt the original term set to the terminology really used by end users, e.g. by ontology learning tools, to improve the recall of the system.

## References

- Abusalah, M., Tait, J., & Oakes, M. (2009). Cross language information retrieval using multilingual ontology as translation and query expansion base. *Polibits. Research Journal on Computer Science and Computer Engineering with Applications*, 40, 13–16.
- Aguado de Cea, G., Álvarez de Mon, I., & Montiel-Ponsoda, E. (2009). From linguistic patterns to ontology structures. *Proceedings of International Conference on Terminology and Artificial Intelligence*.
- Aguado de Cea, G., & Montiel-Ponsoda, E. (2012). Term variants in ontologies. *Proceedings of 30th Conference of Asoc. Espan. de Linguística Aplicada (AESLA)*, Lleida.
- Al-Feel, H., Schafermeier, R., & Paschke, A. (2013). An inter-lingual reference approach for multilingual ontology matching. *IJCSI International Journal of Computer Science Issues*, 10(2), 1.
- Andrich, R. (2011). Towards a global information network: The European assistive technology information network and the world alliance of AT information providers. In G. J. Gelderblom, M. Soede, L. Adriaens, & K. Miesenberger (Eds.), *Everyday technology for independence and care*. Amsterdam: IOS Press.
- Andrich, R., Gower, V., Lyhne, T., & Petersen, M. (2012). Taxonomy of resources. ETNA Report, ICT-PSP-(270746), ETNA project. Retrieved from <http://www.etna-project.eu/resultados.html>
- Buder, M., Rehfeld, W., & Seeger, T. (Eds.). (1990). *Grundlagen der praktischen Information und Dokumentation*. Saur: München.
- Cimiano, P., Buitelaar, P., McCrae, J., & Sintek, M. (2010). LexInfo: A declarative model for the lexicon-ontology interface. *Web Semantics*, 9(1), 29–51.
- Corcho, O., Fernandez-Lopez, M., & Gomez-Perez, A. (2003). Methodologies, tools and languages for building ontologies. Where is their meeting point? *Data & Knowledge Engineering*, 46, 41–64.
- Dragoni, M., Franchescomaroni, C., Ghidini, C., Clemente, J., & Sánchez Alonso, S. (2013). Guiding the evolution of a multilingual ontology in a concrete setting. *Proceedings of 10th Extended Semantic Web Conference (ESWC)*.
- Embley, D., Liddle, S., Lonsdale, D., & Tijerino, Y. (2011). Multilingual ontologies for cross-language information extraction and semantic search. In *ER (Lecture Notes in Computer Science, Vol. 6998)*. Heidelberg: Springer.
- Espinoza, M., Gómez Pérez, A., & Mena, E. (2008). Enriching an ontology with multilingual information. *Proceedings of 5th European Semantic Web Conference*. Tenerife: Springer.

- Eynard, D., Mateucci, M., & Marfia, F. (2012). A modular framework to learn seed ontologies from text. In M. T. Paziienza & A. Stellato (Eds.), *Semi-automatic ontology development: Processes and resources* (pp. 22–47). Hershey: IGI Global.
- Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., & Pet, M. et al. (2006). Lexical Markup Framework (LMF). *Proceedings of LREC*, Genoa.
- Fu, B., Brennan, R., & O'Sullivan D. (2009). Cross-lingual ontology mapping – An investigation of the impact of machine translation. *Proceedings of 4th Asian Semantic Web Conference, LNCS 5926* (pp. 1–15).
- Gangemi, A., & Presutti, V. (2009). Ontology design patterns. In S. Staab & R. Studer (Eds.), *Handbook on ontologies* (pp. 221–243). Heidelberg: Springer.
- Gelderblom, G. J., Soede, M., Adriaens, L., & Miesenberger, K. (2011). *Everyday technology for independence and care*. Amsterdam: IOS Press.
- Gillam, L., Tariq, M., & Ahmad, K. (2005). Terminology and the construction of ontology. *Terminology 11*(1), 55–81. John Benjamins.
- Giunchiglia, F., Marchese, M., & Zaihrayeu, I. (2006). *Encoding classifications into lightweight ontologies*. University of Trento Technical Report # DIT-06-016.
- Gower, V., Andrich, R., Agnoletto, A., Winkelmann, P., Lyhne, T., & Rozis, R. et al. (2012). The European assistive technology information portal (EASTIN): Improving usability through language technologies. *Proceedings of International Conference on Computers helping People with Special Needs (ICCHP)*, Linz.
- Heerkens, Y. F., Bougie, T., & de Kleijn-de Vrankrijker, M. W. (2012). Classification and terminology of assistive products. In J. H. Stone & M. Blouin (Eds.), *International encyclopedia of rehabilitation*. <http://cirrie.buffalo.edu/encyclopedia/en/article/265/>. Visited 3/2013
- ISO 9999. (2011). ISO9999: International standard ISO 9999:2011, Assistive products for persons with disability – Classification and terminology. ISO.
- Khan, L., Luo, F., & Yen, I. (2002). Automatic ontology derivation from documents. In: *IEEE transactions on knowledge and data engineering (TKDE)*.
- León-Arauz, P., & Faber, P. (2014). Context and terminology in the multilingual semantic web. In P. Buitelaar & P. Cimiano (Eds.), *Towards the multilingual semantic web*. Berlin: Springer (in this volume).
- LMF: Lexical Markup Framework, ISO 24613. Retrieved from <http://www.lexicalmarkupframework.org/>
- Lopes, L., Vieira, D., Finatto, M. J., Martins, D., Zanette, A., & Ribeiro, L. C. (2009). Automatic extraction of composite terms for construction of ontologies: An experiment in the health care area. *RECII: Electronic Journal of Communication, Information & Innovation in Health*, 3(1), 72–84.
- Madsen, B. N., & Thomsen, H. E. (2009). CAOS – A tool for the construction of terminological ontologies. *Proceedings Nordic Conference on Computational Linguistics (NODALIDA)*.
- Madsen, B. N., Thomsen, H. E., Halskov, J., & Lassen, T. (2010). Automatic ontology construction for a national term bank. *Proceedings of the Terminology and Knowledge Engineering Conference (TKE)*, Dublin.
- McCrae, J., & Unger, C. (2014). Design patterns for engineering the ontology-lexicon interface. In P. Buitelaar & P. Cimiano (Eds.), *Towards the multilingual semantic web*. Berlin: Springer (in this volume).
- Montiel-Ponsoda, E., Aguado de Cea, G., Gómez-Pérez, A., & Peters, W. (2009). Enriching ontologies with multilingual information. *Natural Language Engineering*, 1(1), 1–27.
- Montiel-Ponsoda, E., Aguado de Cea, G., & McCrae, J. (2011). Representing term variation in lemon. *Proceedings of 9th International Conference on Terminology and Artificial Intelligence (TIA)*, Paris.
- Tanev, H., & Zavarella, V. (2014). Multilingual learning and population of event ontologies. A case study for social media. In P. Buitelaar & P. Cimiano (Eds.), *Towards the multilingual semantic web*. Berlin: Springer (in this volume).
- Tariq, M., Manumaisupat, P., Al-Sayed, R., & Ahmad, K. (2003). Experiments in ontology construction from specialist texts. *Proceedings of EuroLan*, Bucharest.

- Thurmair, G. (2004). Multilingual content processing. *Proceedings of LREC*, Lisbon.
- Thurmair, G., Agnoletto, A., Gower, V., & Rozis, R. (2012). EASTIN-CL: A multilingual front-end to a database of Assistive technology products. *Proceedings of European Association of Machine Translation (EAMT)*, Trento.
- Tomassen, S., & Strasunskas, D. (2009). Relating ontology and Web terminologies by feature vectors: Unsupervised construction and experimental validation. *Proceedings of 11th International Conference on Information Integration and Web-based Applications & Services (IIWAS)*, Kuala Lumpur.
- Trapman, J., & Monachesi, P. (2009). Ontology engineering and knowledge extraction for crosslingual retrieval. *Proceedings of International Conference Recent Advances in Natural Language Processing (RANLP)*, Borovets.
- Trojahn, C., Quaresma, P., & Vieira, R. (2008). A framework for multilingual ontology mapping. *Proceedings of LREC*, Marrakech.
- Trojahn, C., Quaresma, P., & Vieira, R. (2010). An API for multi-lingual ontology matching. *Proceedings of LREC*, La Valetta.
- United Nations. (2007). The UN convention on the rights of people with disabilities. Retrieved from [www.un.org/disabilities/](http://www.un.org/disabilities/)
- Velardi, P., Missikoff, M., & Basili, R. (2001). Identification of relevant terms to support the construction of domain ontologies. *Proceedings of Human Language Technology and Knowledge Management (HLTKM)*.
- Vossen, P. (2004). EuroWordNet: A multilingual database of autonomous and language-specific wordnets connected via an inter-lingual-index. *International Journal of Lexicography*, 17(2), 161–173.
- Walter, S., Unger, C., & Cimiano, P. (2013). A corpus-based approach for the induction of ontology lexica. *Proceedings of 18th International Conference on Applications of Natural Language to Information Systems (NLDB)*, Salford.
- Winkelmann, P. (2011). REHADAT: The German information system on assistive devices. In G. J. Gelderblom, M. Soede, L. Adriaens, & K. Miesenberger (Eds.), *Everyday technology for independence and care*. Amsterdam: IOS Press.

# Service-Oriented Architecture for Interoperability of Multilanguage Services

Yohei Murakami, Donghui Lin, and Toru Ishida

**Abstract** Since the Internet increases the opportunity to interact with foreign people in daily life, multilingual communication tools are necessary. However, the applicability of multilingual communication tools is generally limited because the quality of translation is not high enough to translate an arbitrary text correctly. To develop a multilingual environment that can handle various situations in various communities, existing language resources (dictionaries, parallel texts, part-of-speech (POS) taggers, machine translators, etc.) should be easily shared and combined beyond their complicated intellectual property problems and mismatch of their interfaces. Therefore, we introduce a service-oriented architecture to realize the Language Grid. It allows users to realize interoperability of language services and easily compose those language services to support multilingual communication. This chapter explains the system architecture of the Language Grid and its service domain model to define service interfaces and service profiles.

**Key Words** Language service • SOA • The Language Grid • Web service

## 1 Introduction

The Internet allows people to be linked together regardless of location. However, language remains the biggest barrier. Its users speak a wide variety of languages (Paolillo et al. 2005). In fact, it is not possible for anyone to learn the languages needed to access all possible information from the Internet. Though there are many successful language resources (both data and software) on the Internet, difficulties often arise when people try to use those language resources in their own intercultural activities. Complex contracts, intellectual property rights, and nonstandard application interfaces make it difficult for users to create customized language services that support their activities.

---

Y. Murakami (✉) • D. Lin • T. Ishida

Department of Social Informatics, Kyoto University, Kyoto, Japan

e-mail: [yohei@i.kyoto-u.ac.jp](mailto:yohei@i.kyoto-u.ac.jp); [lindh@i.kyoto-u.ac.jp](mailto:lindh@i.kyoto-u.ac.jp); [ishida@i.kyoto-u.ac.jp](mailto:ishida@i.kyoto-u.ac.jp)

© Springer-Verlag Berlin Heidelberg 2014

P. Buitelaar, P. Cimiano (eds.), *Towards the Multilingual Semantic Web*,

DOI 10.1007/978-3-662-43585-4\_19

To address these kinds of issues, service-oriented architectures might be a promising solution, which is a paradigm for organizing and utilizing distributed software and data by regarding them as services. Each service runs independently under the control of different ownerships, does not depend on implementation, and publishes a well-defined interface so that users can discover and invoke it remotely. These features enable users to quickly and flexibly build a system by composing the services.

In this chapter, we apply the service-oriented architecture to a multilanguage infrastructure to promote usage of language resources for multilingual communication. We have developed the Language Grid, a service-oriented platform to share language services (Ishida 2011). In this platform, end users can combine existing language services provided by researchers and users to create new language services for their own purposes. To realize the Language Grid, however, we must address the following issues:

*Service architecture:* The service platform should allow users to create services and share them. Based on various atomic services, an infrastructure for service composition should be provided. The service architecture should also allow users to develop Web applications for supporting multilingual activities on the Web based on the provided language services.

*Service domain model:* The service platform should allow users to flexibly replace a language service with another to customize language services that support their needs. To this end, the service platform should classify language services according to functionalities and define standard interfaces for each language service type. The service platform should also provide a domain model to define metadata of each language service type.

The remaining parts of this chapter are organized as follows. First, Sect. 2 explains the necessity of shifting from language resources to language services. Section 3 describes the system architecture of the Language Grid, and Sect. 4 introduces the language service domain to promote interoperability of language services.

## 2 Shift to Language Services

The service-oriented approach allows users to share and create value-adding language services. Data like multilingual dictionaries and parallel texts can be wrapped to create atomic language services to provide a translation of words or sentences. Those atomic services retrieve the translation not only by simple exact matching but also by advanced similarity matching: a parallel text service can return the translation of a sentence that is similar to the input sentence. Wrapping software like machine translators is straightforward. Even human interpreters can be wrapped as translation services, so that there is no essential difference between human translators and machine translation systems, other than their quality of service, with



human translators providing more accurate translations and machines providing translations at faster rates.

Moreover, atomic language services can be composed to create a new service according to user's needs. For instance, to translate Japanese sentences into Portuguese, we can cascade Japanese-English and English-Portuguese translators, even though there is no available direct translator handling Japanese to Portuguese. We may append further reverse translators to create back-translation, say, Japanese-Portuguese-Japanese translation. It enables users to compare original and back-translated Japanese sentences and select the translators that can produce back-translated sentences most similar to the original ones. To replace mistranslated jargon output by machine translators with the correct words in multilingual dictionaries for user domain, we need to combine part-of-speech taggers to divide the input sentences into words.

However, part-of-speech taggers are often developed in research institutes or universities and are provided only for research purposes. Their Web sites do not state that they can be used in elementary schools, hospitals, and so on. Nobody in elementary schools thinks they are useful to solve language barriers in the communities. Even if an elementary school wants to use them, the school needs to ask those providers for permission by a letter or e-mail. One of the important roles of language services is to reduce such negotiation costs related to intellectual property rights and installation costs to make language resources readily available.

### 3 The Language Grid

As illustrated in Fig. 1, the Language Grid is a service platform that allows users to share and combine language services provided by both professionals and end users in various application fields, such as disaster management field, education field, and medical care field (Ishida 2011). Major stakeholders of the Language Grid fall into three categories: *language grid operator*, *service provider*, and *service user*. Language grid operators manage the Language Grid and control language services. Service providers provide language services such as machine translations, part-of-speech taggers, dependency parsers, dictionaries, and parallel texts and register them in the Language Grid. Service users invoke registered language services for their multilingual communications. Note that a single group can act as two different stakeholders: service provider and service user.

#### 3.1 Service Layers

The Language Grid consists of the four service layers (Murakami and Ishida 2008). The bottom layer, called *P2P service grid*, aims at connecting two kinds of servers (core nodes and service nodes). Core nodes manage all requests to language

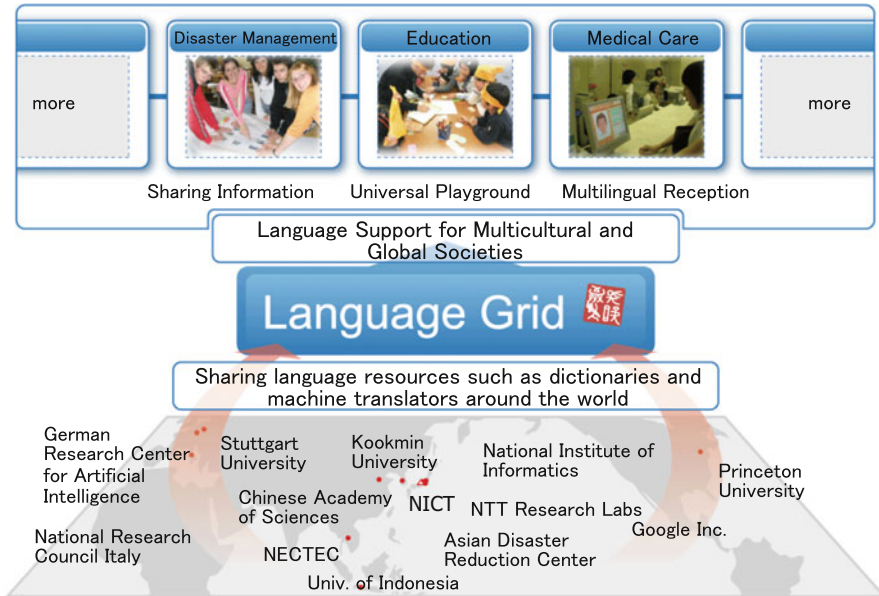


Fig. 1 Overview of the Language Grid (Ishida 2011)

services and combine multiple atomic services according to workflows, while service nodes actually invoke atomic services. The second layer is called the Atomic service. In this layer, any user can add new language resources to the Language Grid. A Web service that corresponds to a language resource is called an atomic language service. Each language resource is wrapped to develop an atomic language service. The third layer is the Composite service. Atomic language services can be composed by Web service workflows. A service described by a workflow is called a composite language service. Web Services Business Process Execution Language (WS-BPEL) and Java-based scenarios are used to describe the workflows and bind atomic language services to activities in the workflows at runtime (Khalaf et al. 2003). Different types of *application systems* including collaboration tools have been developed on the top layer. For instance, popular collaboration tools including LiquidThreads, an extension for MediaWiki that implements a threaded discussion system, and NOTA, a Web page-creating tool, have been successfully multilingualized.

### 3.2 System Architecture

This section explains service grid architecture a general-purpose architecture that supports sharing and combining services. This architecture can be customized to

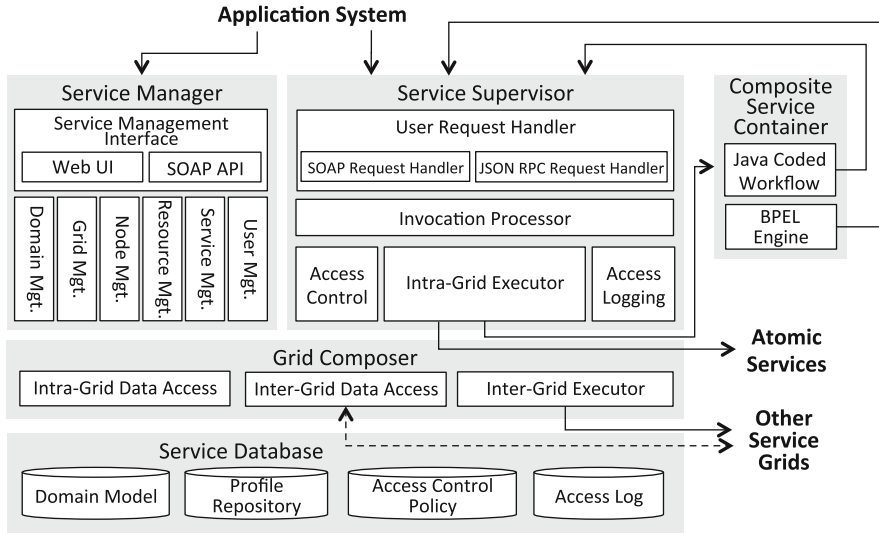


Fig. 2 Service grid architecture

any domain by defining a domain model. The Language Grid is built on this architecture by defining the language service domain. Figure 2 illustrates the service grid architecture. This architecture consists of five parts: *Service Manager*, *Service Supervisor*, *Grid Composer*, *Service Database*, and *Composite Service Container*. In the remaining parts of this section, we provide the details of the Service Manager, Service Supervisor, Grid Composer, and Composite Service Container.

### 3.2.1 Service Manager

The Service Manager consists of components managing various types of information necessary for the service grid, such as nodes, resources, services, and user information. The Domain Management handles a domain model that applies a general service grid to a specific domain. This component sets service types, standard interfaces of services, and attributes of service profiles according to domain model. The Grid Management manages federation settings of service grids. Based on the settings, the Grid Composer determines which information to be shared with which service grids. The Node Management handles node information of its service grid. This information is used by the Grid Composer to distribute registered information to other nodes within its service grid. The Resource Management and Service Management handle resource and service information registered to the service grid and the connected service grid. The information includes access control settings, service endpoints, intellectual properties associated with the resources, and access logs. Based on this information, the Service Supervisor validates service

invocation, locates service endpoints, and attaches intellectual property information to service responses. Lastly, the User Management manages user information registered to the service grid. Based on this information, the Service Supervisor authenticates users' service requests.

### **3.2.2 Service Supervisor**

The Service Supervisor controls service invocation. The control includes user authentication, access control, endpoint locating, load balancing, and access logging. The User Request Handler receives service requests through SOAP and JSON RPC and then authenticates the users. The requests are sent to the Invocation Processor. The Invocation Processor executes a sequence of preprocess, service invocation, postprocess, and logging process. The access control is implemented as a preprocess or a post-process. After passing the access control, the Intra-Grid Executor invokes the service within its service grid. To invoke the service, it locates the service endpoint. If there are multiple endpoints associated with the service, it chooses the lowest load one.

### **3.2.3 Grid Composer**

The Grid Composer not only creates a P2P grid network within its service grid but also connects to other service grids. The former is needed to improve latency if the services are physically distributed. The latter is necessary to realize federated operation of the service grids (Murakami et al. 2012). The Intra-Grid Data Access provides interfaces to read and write the Service Database in the service grid. In writing data, it broadcasts the data to other nodes using the P2P grid network so that it can share the data with other nodes in the same service grid. As a result, service users can improve latency by sending their requests to a node located near the service.

On the other hand, the Inter-Grid Data Access shares various types of information with other service grids. Based on the grid information, the Inter-Grid Data Access sends only information related to the connected service grids. The Inter-Grid Executor invokes services registered on a different service grid. To invoke a service across service grids, it replaces a requester's ID with a key exchanged between the service grids and sends the request to a core node of the other service grid.

### **3.2.4 Composite Service Container**

The Composite Service Container deploys composite services whose abstract workflows are implemented by Java or WS-BPEL. The BPEL workflows are executed by BPEL Engine like active BPEL. In invoking a component service of a composite service, Java-coded workflow or BPEL Engine can select a concrete service, based

on binding information included in a service request. Any other workflow engines like UIMA (Ferrucci and Lally 2004), Heart of Gold (Callmeier et al. 2004), and Taverna (Oinn et al. 2006) can be integrated into the Composite Service Container because the Composite Service Container is independent of workflow engines. We have bridged Heart of Gold and the Language Grid (Bramantoro et al. 2008) and apply the results to combine UIMA and the Language Grid.

### 4 Language Service Domain

To realize interoperability of language services on the service grid architecture, it is necessary to standardize service interfaces and metadata according to their functionalities. To this end, the service grid provides a service domain model for operators to classify services into several service types (Murakami et al. 2012). As illustrated in Fig. 3, the service domain model is not just a type system of data, exchanged between services, but a type system of service interfaces, service metadata, and resource metadata. This model organizes services and resources in the service grid.

Following the service domain model, we defined the language service domain consisting of 16 service types as shown in Table 1. These service types are characterized with ServiceTypeAttributes, which are classified into ones indicating which objects a given service can process and ones indicating methods the service can employ. The former is *supportedLanguages*, *supportedLanguagePairs*, *supportedLanguagePaths*, *supportedImageTypes*, *supportedAudioTypes*, and *supportedVoiceTypes*. They are used to specify languages, images, and audio files to be processed by services. The latter is *supportedMatchingMethod*. This is used to specify search functionalities implemented on language data such as bilingual dictionaries, concept dictionaries, and so on.

Moreover, we defined a service interface for each service type. To standardize the interface, we extracted common parameters of language resources belonging to the same resource type. In case of morphological analyzer, source text and source language for input parameters are common among every morphological

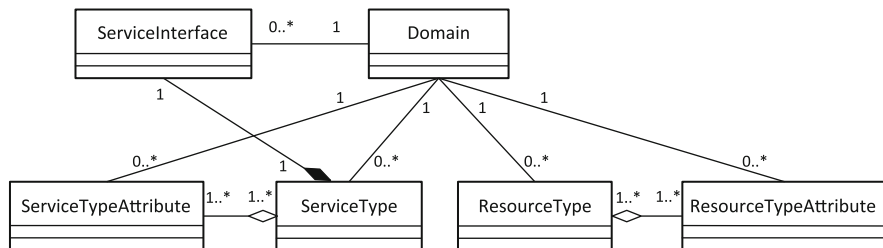


Fig. 3 Service domain model

**Table 1** Language service domain

ServiceType	ServiceTypeAttribute	ServiceInterface
BackTranslation	supportedLanguagePaths	backtranslate
BilingualDictionary	supportedLanguagePairs, supportedMatchingMethods	search
ConceptDictionary	supportedLanguages, supportedMatchingMethods	searchConcepts, getRelatedConcepts
DependencyParse	supportedLanguages	parseDependency
DialogCorpus	supportedLanguages, supportedMatchingMethods	search
LanguageIdentification	supportedEncodings, supportedLanguages	identify
MorphologicalAnalysis	supportedLanguages	analyze
MultihopTranslation	supportedLanguagePaths	multihopTranslate
ParallelText	supportedLanguagePairs, supportedMatchingMethods	search
Paraphrase	supportedLanguages	paraphrase
PictogramDictionary	supportedLanguages, supportedMatchingMethods, supportedImageTypes	search
SimilarityCalculation	supportedLanguages	calculate
SpeechRecognition	supportedLanguages, supportedAudioTypes, supportedVoiceTypes	recognize
TextToSpeech	supportedLanguages, supportedAudioTypes, supportedVoiceTypes	speak
Translation	supportedLanguagePairs	translate
TranslationWith TemporalDictionary	supportedLanguagePairs	translate

**Table 2** Output formats of morphological analyzers

Name	Language	Format
TreeTagger	English	<i>word POS lemma</i>
MeCab	Japanese	<i>word POS,subPOS1,subPOS2,subPOS3, lemma,reading, pronunciation</i>
Juman	Japanese	<i>word reading lemma POS subPOS category/domain</i>
KLT	Korean	<i>word:POS:lemma</i>
ICTCLAS	Chinese	<i>word/POS</i>

analyzer. On the other hand, we have many formats of morphemes for output parameters. Table 2 compares output formats of morphological analyzers among different languages: English, Japanese, Korean, and Chinese. Every analyzer returns

word, lemma, and part of speech tag except for Chinese analyzer. Therefore, we defined the output of morphological analysis service as an array of triples consisting of word, lemma, and POS tag. Furthermore, we enumerated POS tags available in the output of the analysis service. Since POS tags vary depending on languages, we selected a minimal set of POS tags occurring in every language: noun, proper noun, pronoun, verb, adjective, adverb, unknown, and others. Most morphological analyzers can be wrapped with this standard interface. A few morphological analyzers not complying with this interface, such as ICTCLAS, return “NULL” as unassigned parameters. This interface is designed for interoperability instead of completeness. As a result, information generated by the original morphological analyzers can be lost. When many service users need more detailed information, a new subservice type is designed by inheriting the basic morphological analysis service interface. The inherited service interface can extend the service interface while maintaining the consistency with the existing one.

This inheritance of service interfaces constructs a hierarchy of homogeneous services like an OWL-S profile hierarchy (Elenius et al. 2005), which is used to discover alternatives to the existing one. Meanwhile, to enhance interoperability among heterogeneous language services, a language service ontology has been proposed by Hayashi et al. (2008). The ontology consists of a top-level ontology and subontologies. The top-level ontology defines the relations among language service class, language processing resource class, language data resource class, and linguistic object class. A language service is provided by an instance of the language processing resource class, whose input and output are instances of linguistic object class. A language data resource consists of instances of the linguistic object class. On the other hand, each subontology organizes classes for language processing resources, for language data resources, and for linguistic objects, respectively. The interoperability of heterogeneous language services can be realized by semantics of language processing resources, language data resources, and linguistic objects defined in the subontologies.

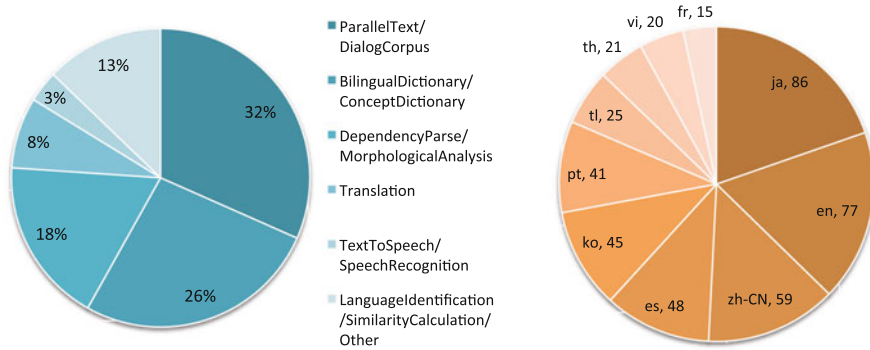
Due to limitations of space, Table 1 shows only the operation name except for input and output parameters. Refer to [http://langrid.org/service\\_manager/service-type](http://langrid.org/service_manager/service-type) for the WSDL files and more information. The attributes and interfaces help service users to compose services by searching services with the metadata and changing the services belonging to the same service type.

#### ***4.1 Atomic Language Service***

Currently, 117 atomic language services are available on the Language Grid operated by Kyoto University.<sup>1</sup> On the left side of Fig. 4, most language services are classified into language data such as parallel texts and dictionaries because many users provide various but small data created in their community. These services

---

<sup>1</sup>[http://langrid.org/service\\_manager/language-services](http://langrid.org/service_manager/language-services).



**Fig. 4** Distribution of atomic language services by service types (*left*) and supported languages (*right*)

are useful to customize general-purpose translation services to a specific domain. On the other hand, the right side of Fig. 4 shows that the number of language services supporting Japanese and English is substantially higher compared to other languages. This is due to the fact that most language services are provided by Japanese users. These are followed by Chinese, Spanish, Korean, and Portuguese. These languages represent barriers we usually encounter in Japan. To solve the bias of languages, we started federated operation of the Language Grid with Thailand National Electronics and Computer Technology Center (NECTEC) in Thailand. As a result, 22 atomic language services covering 13 Asian languages are shared through the federation. The federation accelerates sharing language services and expands the coverage of languages supported by language services.

Each atomic language service implements the corresponding service interfaces and is described using the corresponding attributes. For example, the resource CaboCha is an instance of DependencyParser type and has supportedLanguages attribute whose value is Japanese. Service CaboCha, meanwhile, can be an instance of both DependencyParse type and MorphologicalAnalysis type because results of dependency parsers generally include morphological analysis results. In addition, service CaboCha belonging to DependencyParse type has four endpoints for load balancing, two of which employ SOAP and the rest of which employ JSON RPC. Every endpoint provides the same interface, whose operation is “parseDependency.”

## 4.2 Composite Language Service

Currently, 22 composite language services are registered on the Language Grid operated by Kyoto University. Most composite language services enhance translation service by combining other language services, such as bilingual dictionary services and parallel text services. These composite language services are implemented in



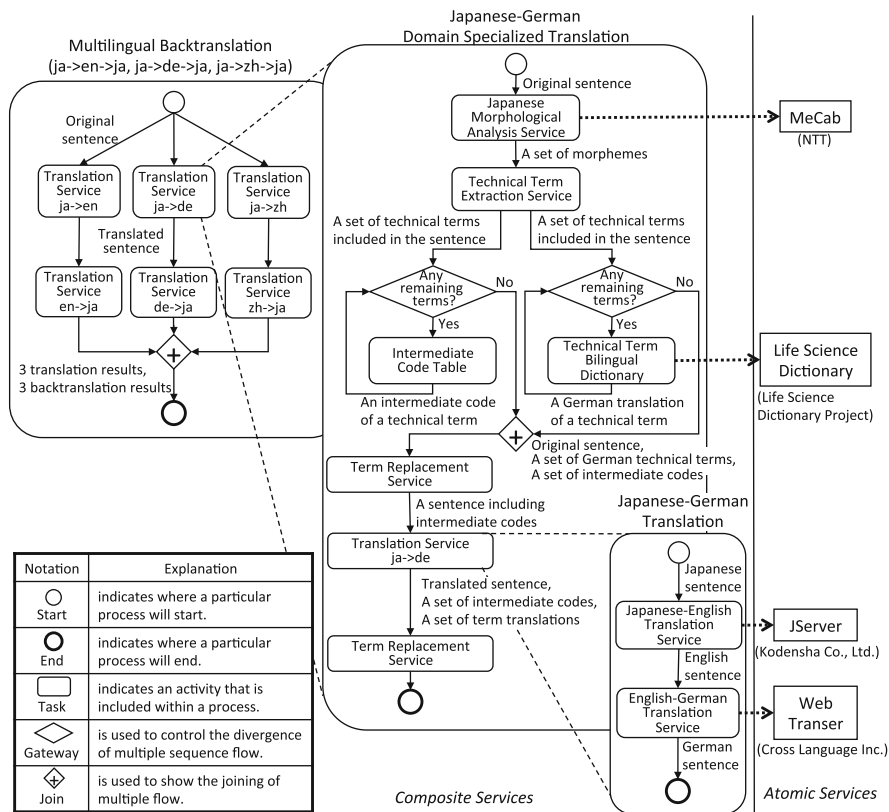


Fig. 5 An example of composite language service in BPMN

a workflow. The Language Grid uses Java-coded workflows, JavaScript, and WS-BPEL to describe the workflow.

Figure 5 shows a multilingual back-translation workflow and a domain-specialized translation workflow for improving the translation quality of technical sentences. This domain-specialized translation workflow consists of several component service types: morphological analysis service type, bilingual dictionary service type, and translation service type. To invoke the composite service, service users have to bind a concrete atomic service to each component service, such as MeCab to the morphological analysis service type, Life Science Dictionary to the bilingual dictionary service type, and a two-hop translation service consisting of J-Server (machine translator) and WEB-Transter (machine translator) to the translation service type. Service users can also invoke other combinations of concrete atomic services, as service interfaces are standardized by the language service domain model (Murakami et al. 2006). Moreover, the users can also delay binding services to choose the fastest or most popular one that provides functionality they are interested in.



Fig. 6 Application of Language Grid Toolbox: G30

## 5 Use Cases

We developed the Language Grid Toolbox (hereafter Toolbox), an intercultural collaboration support tool using the Language Grid. Toolbox is a Web application based on XOOPS, an open-source content management system (CMS). Toolbox consists of several stand-alone tools called Toolbox modules to support multilingual communities. The community administrator can construct the customized multilingual environment by activating only the Toolbox modules that are required for that community. Community members communicate with each other via the Toolbox modules. Figure 6 shows the top page of Toolbox applied to Global 30 community site, the purpose of which is to promote the collaboration among foreign and Japanese students in a campus life. Currently, 180 students join this community site. This section describes how to use the language services to support multilingual community through the use case of the Toolbox.

Toolbox provides the multilingual Bulletin board system (BBS), which lets community members communicate with each other in their native language, since its contents are translated by language services on the Language Grid. Figure 7 is a screenshot of the multilingual BBS displayed in English. The post information under participant's name indicates the language in posting. This multilingual discussion shows two Japanese students helping a Chinese student and an English student to understand a technical presentation in Japanese in a seminar. A slide of the presentation is shown on the right. Users posting a message can link it to a slide. The user can also put a pointer which clarifies the context for other participants who read the machine translation of the message. As shown in Fig. 8, the posts in various native languages are translated into a reader's native language so that the reader can easily understand the others' messages.

The messages to explain the presentation are sometimes too technical to be correctly translated. To improve the quality of technical translation, the

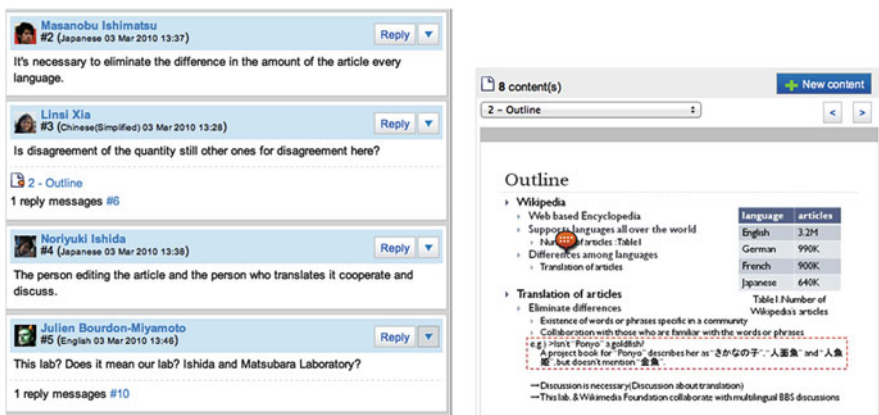


Fig. 7 Multilingual discussion on English user's display



Fig. 8 Multilingual discussion on Japanese user's display (left) and Chinese one (right)

domain-specialized translation service described in the previous section is often useful. Since most Toolbox modules rely on this composite service to support technical communication, Toolbox provides a fundamental module, called Langrid Access module, to access language services on the Language Grid and to manage the service settings, which indicate which translation services and dictionary services are used for which translation path. Figure 9 shows the interface of this module. These settings consist of six translation paths: between Chinese and English, Chinese and Korean, English and Korean, Japanese and Chinese, Japanese and English, and Japanese and Korean. Each translation path uses Toshiba English-Chinese Machine Translation, two-hop translation connecting two J-Servers, Google Translate, and J-Server, respectively. Every translation service is combined with the Agent Research Dictionary service.

Based on these settings, this module also generates a corresponding service binding information for an SOAP request. Figure 10 illustrates a sample SOAP

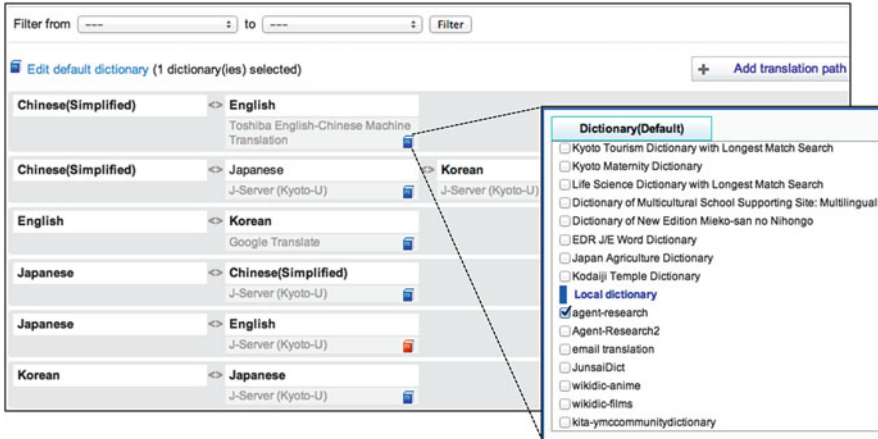


Fig. 9 Service settings for Language Grid

```

<soapenv:Envelope ...>
  <soapenv:Header>
    <ns1:binding ...>
      [{"children": [],
        "invocationName": "MorphologicalAnalysisPL",
        "serviceId": "TreeTagger"},
       {"children": [],
        "invocationName": "TranslationPL",
        "serviceId": "J-Server"},
       {"children": [],
        "invocationName": "BilingualDictionaryPL",
        "serviceId": "AgentResearchDictionary"}]
    </ns1:binding>
  </soapenv:Header>
  <soapenv:Body>
    <tran:translate ...>
      <sourceLang xsi:type="xsd:string">en</sourceLang>
      <targetLang xsi:type="xsd:string">ja</targetLang>
      <source xsi:type="xsd:string">
        This lab? Does it mean our lab? Ishida and Matsubara Laboratory?
      </source>
    </tran:translate>
  </soapenv:Body>
</soapenv:Envelope>

```

Fig. 10 SOAP request for domain-specialized translation service

request generated by the Langrid Access module. The service binding information is located between “<ns1: binding>” tags. This binds TreeTagger to morphological analysis service type, J-Server to translation service type, and Agent Research Dictionary to bilingual dictionary service type. The body of this request must adhere to the translation service interface specification, which consists of the operation “translate” and three input parameters: source language, target language, and source text. The module sends this request to the domain-specialized translation service on the Language Grid, receives the translation result, and returns it to the multilingual

BBS module. In this way, the Langrid Access module takes a role of mediator between Toolbox modules and the Language Grid.

## 6 Conclusion

The Language Grid is an infrastructure that allows end users to create new language services for their intercultural collaboration activities. This chapter proposed the service-oriented architecture to support the collection, sharing, and production of new services on the Internet and defined a language service domain to realize interoperability of language services. The main contributions of the proposed approach include the following two aspects.

*Service architecture:* We developed the service architecture for the Language Grid, including layers of P2P grid infrastructure, atomic services, composite services, and application systems. The proposed architecture applies the service-oriented approach, where language resources including data and software are wrapped as Web services so that users can easily share and combine language services for creating their own multilingual environment.

*Service domain model:* We created the service domain model to define resource types and service types, their attributes, and standard service interfaces. Using this model, we defined the language service domain to realize interoperability of language services. As a result, in a workflow, users can easily find alternate language services and change the language services belonging to the same language service type.

As the number of language services belonging to the same type increases, the horizontal service composition technique is useful to choose the best combination of language services that satisfy the user's goal under some constraints (Hassine et al. 2006). Furthermore, we need nonworkflow models for service composition to deal with a huge amount of text data because the intermediate results are too big to store in workflow engines before next invocation. The combination of a rule-based approach and streaming processing can be one of the promising solutions to this problem (Murakami et al. 2012). This technology can start invoking a next service without waiting for completing the current service invocation in a streaming fashion and enables users to insert declarative rules to change process logic runtime.

**Acknowledgements** This research was partially supported by a Grant-in-Aid for Scientific Research (S) (24220002) from Japan Society for the Promotion of Science and Service Science, Solutions and Foundation Integrated Research Program (S3FIRE) from RISTEX, Japan Science and Technology Agency (JST).

## References

- Bramantoro, A., Tanaka, M., Murakami, Y., Schäfer, U., & Ishida, T. (2008). A hybrid integrated architecture for language service composition. In *Proceedings of the Sixth International Conference on Web Services (ICWS'08)* (pp. 345–352).
- Callmeier, U., Eisele, A., Schäfer, U., & Siegel, M. (2004). The deep thought core architecture framework. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)* (pp. 1205–1208).
- Elenius, D., Denker, G., Martin, D., Gilham, F., Khouri, J., Sadaati, S., et al. (2005). The owl-editor - A development tool for semantic web services. In *ESWC* (pp. 78–92).
- Ferrucci, D., & Lally, A. (2004). UIMA: An architectural approach to unstructured information processing in the corporate research environment. *Journal of Natural Language Engineering*, 10, 327–348.
- Hassine, A., Matsubara, S., & Ishida, T. (2006). A constraint-based approach to horizontal web service composition. In *Proceedings of the Fifth International Semantic Web Conference (ISWC-06)* (pp. 130–143).
- Hayashi, Y., Declerck, T., Buitelaar, P., & Monachini, M. (2008). Ontologies for a global language infrastructure. In *Proceedings of the First International Conference on Global Interoperability for Language Resources (ICGL'08)* (pp. 105–112).
- Ishida, T. (Ed.). (2011). *The Language Grid: Service-oriented collective intelligence for language resource interoperability*. Berlin: Springer.
- Khalaf, R., Mukhi, N., & Weerawarana, S. (2003). Service-oriented composition in BPEL4WS. In *Proceedings of the Twelfth International World Wide Web Conference (WWW'03)*.
- Murakami, Y., & Ishida, T. (2008). A layered language service architecture for intercultural collaboration. In *Proceedings of the International Conference on Creating, Connecting and Collaborating Through Computing (C5-08)*.
- Murakami, Y., Ishida, T., & Nakaguchi, T. (2006). Infrastructure for language service composition. In *Proceedings of the Second International Conference on Semantics, Knowledge, Grid (SKG-06)*.
- Murakami, Y., Tanaka, M., Bramantoro, A., & Zettsu, K. (2012). Data-centered service composition for information analysis. In *Proceedings of the IEEE International Conference on Services Computing (SCC-12)* (pp. 602–608).
- Murakami, Y., Tanaka, M., Lin, D., & Ishida, T. (2012). Service grid federation architecture for heterogeneous domains. In *Proceedings of the IEEE International Conference on Services Computing (SCC-12)* (pp. 539–546).
- Oinn, T., Greenwood, M., Addis, M., Alpdemir, N., Ferris, J., Glover, K., et al. (2006). Taverna: Lessons in creating a workflow environment for the life sciences. *Concurrency and Computation: Practice and Experience*, 18(10), 1067–1100.
- Paolillo, J., Pimienta, D., & Prado, D. (2005). *Measuring linguistic diversity on the Internet*. Paris: UNESCO.

# Index

## A

Adoption of language technology, 68, 70, 71, 79, 80  
Afrikaans, 51  
Annotation property, 227, 230, 235  
Astronomy indigenous knowledge, 55  
*Atomic service*, 316

## B

Best practices and tools, 103  
Bootstrapping Linked Data (BOA), 209  
Bulgarian, 221

## C

Catalan, 221  
Chinese verbalization patterns, 90  
CIDOC-CRM, 214  
Coherent, 211, 212, 214, 222, 223  
Collaboration, 176  
Collaborative ontology engineering, 176  
*Composite service*, 316  
Concept dynamics, 32, 37, 41, 45  
Conceptualization, 33, 35, 37, 39, 41, 42  
Conceptual relations, 245, 251  
Conceptual scoping, 68–69  
Correcting-completive method, 229, 231, 237, 238, 241  
Correcting-completive patterns, 228–233  
Cross-language information retrieval, 157  
Cross-language query processing, 159, 165–166  
Cross-language query system, 156, 159  
Cross-language retrieval, 186

Cross-lingual, 104  
    comparison, 227, 228, 232, 234, 240  
    linking, 109  
Cultural diversity, 49

## D

Danish, 221  
Data categories, 236  
Data integration, 211  
Data models, 105  
Dataset publication, 114  
Data vs information, 3  
DBpedia, 138, 213–215, 221  
DeFacto, 209  
Description Logic, 214  
Disasters, 259  
Distributional semantics, 11, 261  
Domain experts, 179  
Domain-specific complement, 228, 229, 231–236, 238, 239, 241  
    alignment, 234, 235  
    extraction, 234  
    supplementation, 232, 233, 235, 238, 239, 241  
Dutch, 6, 221

## E

Ecosystem, 78, 80  
Effective discovery, 114  
Ellipsis, 228, 229, 233, 235  
    compound, 231  
    contextual, 231, 233  
    grammar, 231, 233  
    resolution, 231–233, 235, 238, 239, 241

- structural, 231–233
    - syntactic, 231
    - syntactic indicators, 229, 231–233, 238
  - Encoding issues, 113
  - End-user front ends, 308
  - English, 6, 51, 221
  - Europeana, 214
  - Evaluation, 126
    - data sets, 126
  - Event extraction, 261
  - Experiments, 131
  - Expressiveness of representation, 5
  - Extraction ontology, 159–163
    - constraints, 160
    - inference rules, 160, 162–163
    - linguistic groundings, 160–162
    - object sets, 159
    - relationship sets, 160
- F**
- Fact Checking, 209
  - Fact Validation, 209
  - Finnish, 221
  - French, 6, 221, 222
  - Fully-structured access mode, 181
  - Function, 215–217, 219, 220
- G**
- German, 6, 221, 222
  - Gothenburg City Museum, 214, 215
  - Grammar generation, 75
  - Grammatical Framework (GF), 72, 75, 78, 211–213, 215, 217, 218, 221, 223, 224
  - Grid Composer, 318
- H**
- Hebrew, 221, 223
  - High-quality multilingual linked data, 102
- I**
- Ideological analysis, 8
  - ILI. *See* Interlingual Index (ILI)
  - Indigenous knowledge systems, 49, 53
  - Inference, 213, 214
  - Information extraction, 262
  - Information vs data, 3
  - Intelligence gathering, 7
  - Interchangeability, 5
  - InterLingual Index (ILI), 244, 255
  - Interlingua mediation, 157
  - Internationalized resource identifiers (IRIs), 106
  - Interoperability, 214, 215, 223
  - ISO 9999, 297
  - Italian, 221–223
- J**
- Japanese, 6
  - Jena, 214
- K**
- Knowledge, 178
    - engineers, 179
- L**
- Labels, 109
  - Language diversity, 50
  - Language experts, 179
  - Language Grid, 315
  - Language identification, 112–113
  - Language service domain, 319
  - Language services, 314
  - Language specific DBpedia chapters, 140
  - Lemon*, 5, 16, 32–34, 38, 43, 57, 72, 80
  - Lemon Design Patterns*, DSL, 24
  - Lexical choice, 7
  - Lexical gaps, 5
  - Lexicalization, 252
    - basic, 230, 238
    - complex, 231, 238
  - Lexicalization for Search, 300–305
  - Lexical learning, 259
  - Lexicon model, 109
  - Lexico-syntactic patterns, 228, 235, 241
  - LexInfo, 5
  - Licensing, 245, 247, 250
  - Lightly-structured access mode, 181
  - Linguistic annotation, 61
  - Linguistic grounding process, 86
  - Linguistic groundings, 160
  - Linked Data, 3, 55, 209
  - Linked Data principles, 61
  - Linked Open Data (LOD), 213, 224
- M**
- Machine learning, 209, 262
  - Machine translation, 8
  - Meaning, reader-centered view, 9
  - Meaning, writer-centered view, 7



- Metadata publication, 114
  - Methodological guidelines, 102
  - ML-OntoES. *See* Multilingual ontology extraction system (ML-OntoES)
  - Modeling tools, 176
  - Modularity, 68, 70–73
  - MoKi, 180
  - Monolingual query processing, 163–165
  - Morphological analysis component, 233
  - Multidimensionality, 37, 38, 42
  - Multilingual, 244, 245, 248, 252, 253, 255
    - dynamics, 40
    - information access, 176
    - interaction, 213, 216
    - knowledge, 176
    - ontologies, 121, 178
    - resources, 114
    - terminology, 55
  - Multilingual and cross-lingual approaches, 123
    - corpus-based approach, 123
    - image-based approach, 125
    - indirect alignment composition, 125
    - instance-based approach, 124
    - linguistic enrichment-based approach, 124
    - machine learning-based approach, 125
    - manual processing-based approach, 123
    - translation-based approach, 124
  - Multilingual datasets, 104
  - Multilingualism, 51
  - Multilinguality, 175, 177
  - Multilingual labeling approach, 109
  - Multilingual lexicon, 95
  - Multilingual natural language
    - descriptions, 219, 223
    - generation, 216
    - translations, 221
  - Multilingual ontology extraction system (ML-OntoES), 158–166
    - construction cost, 168–170
    - cross-language query processing, 165–166
    - extraction ontology, 159–163
    - monolingual query processing, 163–165
    - practicalities, 166–170
    - recognition accuracy, 166–168
  - Multilingual parallel corpus, 59
  - Multilingual Question Answering over Linked Data challenge, 140
  - Multilingual search systems, 177
  - Multilingual Semantic Web, 139
  - Multilingual verbalization, conceptual models, 84
  - Multi-modal access, 180
- N**
- n-ary fact types, binary fact types, 92
  - n-ary fact types, natural verbalization, 86
  - Natural language (NL), 211, 212, 214, 221, 223, 224
  - Natural language definition, 228, 229, 233, 234, 236–239
  - Natural language front end, 296
  - Natural language interfaces, 67
  - Natural Language Processing, 209
  - Near-synonyms, 5
    - cross-lingual, 5, 11
  - NooJ, 230, 233
  - Norwegian, 221
- O**
- OAEI matching systems, 128
  - Objectification, 92
  - OMW. *See* Open Multilingual Wordnet (OMW)
  - Ontology, 176, 177, 213, 214, 216, 220, 221, 223, 224, 244
    - alignment, 236
    - class, 214, 221–223
    - engineering, 176
    - localization, 31, 33, 41, 45, 109
    - modeling, 178
    - population, 259
    - property, 214, 221–223
  - Ontology label, 227, 233, 234, 236, 238
    - labeling practices, 235
  - Ontology-lexicon, 72, 74
    - creation, 75, 77
    - design patterns, 74, 76
  - Ontology-lexicon design patterns, 18
    - Adjectives, 22
    - Event verbs, 20
    - Nouns, 18
    - Property-modifying adjectives, 22
    - Relational adjectives, 22
    - Scalar adjectives, 23
    - State verbs, 20
  - Ontology matching, 122, 177
    - cross-lingual ontology matching, 123
    - monolingual ontology matching, 122
    - multilingual ontology matching, 122
  - Ontotext, 214

- Open Multilingual Wordnet (OMW), 247, 248, 251
- Opinion extraction, 8
- Organic.Lingua, 186
- OWL
  - Horst, 214
  - 2 QL, 214
  - 2 RL, 214
- OWLIM, 213, 214
  
- P**
- POWLA, 62
- Pragmatic dynamics, 38
- Property alignment, 140
- Publishing and consuming LD, 102
  
- Q**
- QAKiS, 146
- QALD, 209
- Query, 212, 214–221, 223, 224
  - analysis, 305
  - generation, 217
- Question Answering, 209
- Question Answering over Linked Data, 139
  
- R**
- RDF, 209
- Reader-centred view of meaning, 9
- Reason-able View, 211, 213–215, 220
- Record, 215
- Relational patterns, 147
- Relation extraction, 209
- Resource Description Framework, RDF, 211–216, 219, 220, 222, 223
  - triple, 212, 215, 220
- Romanian, 221
- Russian, 221, 222
  
- S**
- Search support, 297
- Semantic, 176
- Semantic approaches, 176
- Semantic Web, 209, 212–216, 224, 259
  - imperfect vs incomplete, 10
  - research strategies, 11
- Semiautomated process for analysing source texts, 85
- Sentiment analysis, 8
- Service domain model, 319
- Service grid architecture, 316
- Service Manager, 317
- Service Supervisor, 318
- Sesame, 214
- Sharing and reuse, 72, 78, 80
- Soccer, 5
- Social media, 260
- Spanish, 221–223
- SPARQL, 209, 211, 212, 214, 216, 218, 219, 223
- Specialized knowledge, 33, 38
- Sub-term repository, 228–230, 235, 236, 239, 241
  - alignment, 235, 236, 239
  - extraction, 239
- Suggested Upper Merged Ontology (SUMO), 244, 246, 252, 254
- Swedish, 221
- Synonymy, 245
  
- T**
- Task model, 70–71, 75, 77, 80
- Teamwork activity, 178
- Term alignment, 235
- Term dynamics, 32, 35, 41, 45
- Terminological entry, 236
- Terminology, 297
- Terminology and standards, 105
- Term transparency, 234, 235
- Term variants, 301
- Term variation, 236
- Text generation, 212, 224
- Tool, 180
- Translation, 176
  - misalignments, 6
  - relations, 32, 34, 42, 43, 45
- Translations, 109, 302
- Tswana, 51
- Twitter, 260
  
- U**
- Under-resourced languages, 49
- Under-resourcedness, 52
- Uniform resource identifiers (URIs), 106
- Unstructured access mode, 181
- Use-case, 186
- Users interactions, 146
  
- V**
- Verbalization of ontology models, 84
- Verbalization pattern, 86
- Virtual assistant, 67, 79

Virtual knowledge commons, 50  
Vocabularies, 107

**W**

Weakly supervised, 259  
Web of Data, 146  
Well-defined meaning, 4  
WikiFramework, 147

Wikipedia, 138, 221, 222  
Wordnet, 244–248, 250–252, 255  
World Wide Web, 175  
World Wide Web Consortium, W3C1, 212, 213  
Writer-centred view of meaning, 7

**Z**

Zulu, 51