

Jorge Mauricio Espinoza Mejía

Ontology Localization

Departamento
Informática e Ingeniería de Sistemas

Director/es

Mena Nieto, Eduardo
Gómez Pérez, Asunción

<http://zaguan.unizar.es/collection/Tesis>



Universidad
Zaragoza

Tesis Doctoral

ONTOLOGY LOCALIZATION

Autor

Jorge Mauricio Espinoza Mejía

Director/es

Mena Nieto, Eduardo
Gómez Pérez, Asunción

UNIVERSIDAD DE ZARAGOZA

Informática e Ingeniería de Sistemas



Ontology Localization

Jorge Mauricio Espinoza Mejía

PhD Thesis

Departamento de Informática e Ingeniería de Sistemas
Universidad de Zaragoza

Advisors : Dr. Eduardo Mena
Dra. Asunción Gómez-Pérez

March, 2014

Acknowledgements

I would like to thank my advisors, Eduardo Mena and Asunción Gómez-Pérez, for having always believed in my work and for having provided insightful feedback during all stages of the research described in this thesis. Their vision was fundamental in shaping my research and I am very grateful for having had the opportunity to learn from them.

It is necessary to emphasize that this thesis is the result of a long period of research that started in 2005. In a first stage (24 months) in the Distributed Information Systems (SID) group at University of Zaragoza, we studied and designed the methods used to discover the set of candidate meanings for a given word (or words) from a pool of ontologies available on the Web. This approach was the core of our proposal to automatically discover the translations of an ontology element. The results obtained in the exploratory phase, were the base of a second stage of research more extensive and complex (29 months) within the Ontology Engineering Group (OEG) at Technical University of Madrid. In this period, we defined the methodology that supports the ontology localization activity and we developed the infrastructure that implements the methods, techniques, and tools for the management of ontology localization in distributed and collaborative environments.

The friends at the OEG and the SID group provided a precious collaborative working environment that will be hard to forget. I owe special thanks to the contribution of Jorge Gracia, Elena Montiel-Ponsoda, Raquel Trillo, Boris Billazón, Raúl Palma, M. Carmen Suárez-Figueroa, Jose Angel Ramos, Miguel Esteban, Victor Saquicela, Raúl García-Castro, Manuel Vilches, and Andrés García. Nelly Mantilla and Hernán Cuesta made our life in Zaragoza much easier.

Dra. Guadalupe Aguado de Cea was always so readily available to help me with everything I needed. Dr. Oscar Corcho shared with me some of his inspiring ideas.

Charito was the best person in the world I could have had to share all the good and bad moments of my PhD. I was very lucky to have found her. Cristina and Sofia my beautiful daughters, thanks to be with us. Last, but not least important, I would like to thank my parents, my siblings, and all my family for their great emotional support over all these years.

A special mention to Frank, who spent some days of his life in the task of verifying the correct usage of the English in this thesis. I am also indebted to Jordi Bernard for his support with regard to the formalization of this work.

Finally, I want to mention explicitly that the work contained in this doctoral thesis has been co-financed by an official scholarship from the University of Zaragoza - Banco Santander Central Hispano (citation 2005), for the European Commission in the context of the project Neon (FP6-027595), and the spanish CICYT project TIN2004-07999-C02-02.

Abstract

Our main goal in this thesis is to propose a solution to build a multilingual ontology through automatic localization of an ontology. The notion of localization comes from the Software Development area where it refers to the adaptation of computer software to non-native environments. In Ontological Engineering, the localization of ontologies could be considered as a subtype of software localization in which the product is a shared model of a particular domain, i.e., an ontology, to be used by a certain application.

In particular, our work introduces a novel approach for the multilingualism problem, describing the methods, techniques, and tools for the localization of ontological resources and how multilingualism could be represented in ontologies. It is not the goal of this work to advocate one only approach to ontology localization, but rather to show the variety of methods and techniques that can be re-adapted of other knowledge areas to reduce the cost and effort that means enriching an ontology with multilingual information. We are convinced that there is not one unique approach to ontology localization. We concentrate, however, on automatic solutions for ontology localization.

The approach presented in this dissertation provides a comprehensive coverage of ontology localization activity for the ontology practitioners. In particular, it gives a formal account of our general localization process by defining the inputs, outputs, and the main steps identified. Also, we consider various dimensions for localizing an ontology. Such dimensions allow us to establish a classification of different translation techniques based on methods taken from the discipline of machine translation. To facilitate the analysis of these translation techniques we introduce a framework that covers their main aspects. Finally, we give an intuitive view of the whole localization activity and we outline our approach to the definition of a system architecture that supports the ontology localization activity. The proposed model comprises the system components, the externally visible properties of those components, the relationships between them, and provides a base from which localization systems can be developed.

The principal contributions of this work are summarized as follows:

- *A characterization and definition of the ontology localization problems,*

based on the problems found in related areas. The characterization proposed takes into account three different problems of the localization: translation, information management, and multilinguality representation problems.

- *A prescriptive methodology for supporting ontology localization activity*, based on existing localization methodologies from the fields of Software Engineering and Knowledge Engineering, as general as possible so that the methodology can cover a broad range of scenarios.
- *A classification of the ontology localization techniques*, which can be used for comparing (analytically) different ontology localization systems as well as for designing new ones, taking advantage of state of the art solutions.
- *An integrated method for building ontology localization systems in a distributed and collaborative environment*, which takes into account the more appropriate methods and techniques depending on: i) the domain of the ontology to be localized, and ii) the amount of linguistic information required for the final ontology.
- *A modular component to support the storage of the multilingual information associated to each ontology term*. This approach follows the current trend in the integration of multilinguality in ontologies which suggests the suitability of keeping ontology and linguistic (multilingual) knowledge separated and independent.
- *A model based on collaborative workflows* for the representation of the process usually followed by different organizations to coordinate the ontology localization activity in different natural languages.
- *An integrated infrastructure* implemented within the NeOn Toolkit by means of a set of plug-ins and extensions that supports the collaborative ontology localization process.

Resumen

Nuestra meta principal en esta tesis es proponer una solución para construir una ontología multilingüe, a través de la localización automática de una ontología. La noción de localización viene del área de Desarrollo de Software que hace referencia a la adaptación de un producto de software a un ambiente no nativo. En la Ingeniería Ontológica, la localización de ontologías podría ser considerada como un subtipo de la localización de software en el cual el producto es un modelo compartido de un dominio particular, por ejemplo, una ontología, a ser usada por una cierta aplicación.

En concreto, nuestro trabajo introduce una nueva propuesta para el problema de multilingüismo, describiendo los métodos, técnicas y herramientas para la localización de recursos ontológicos y cómo el multilingüismo puede ser representado en las ontologías. No es la meta de este trabajo apoyar una única propuesta para la localización de ontologías, sino más bien mostrar la variedad de métodos y técnicas que pueden ser readaptadas de otras áreas de conocimiento para reducir el costo y esfuerzo que significa enriquecer una ontología con información multilingüe. Estamos convencidos de que no hay un único método para la localización de ontologías. Sin embargo, nos concentramos en soluciones automáticas para la localización de estos recursos.

La propuesta presentada en esta tesis provee una cobertura global de la actividad de localización para los profesionales ontológicos. En particular, este trabajo ofrece una explicación formal de nuestro proceso general de localización, definiendo las entradas, salidas, y los principales pasos identificados. Además, en la propuesta consideramos algunas dimensiones para localizar una ontología. Estas dimensiones nos permiten establecer una clasificación de técnicas de traducción basadas en métodos tomados de la disciplina de traducción por máquina. Para facilitar el análisis de estas técnicas de traducción, introducimos una estructura de evaluación que cubre sus aspectos principales. Finalmente, ofrecemos una vista intuitiva de todo el ciclo de vida de la localización de ontologías y esbozamos nuestro acercamiento para la definición de una arquitectura de sistema que soporte esta actividad. El modelo propuesto comprende los componentes del sistema, las propiedades visibles de esos componentes, las relaciones entre ellos, y provee además, una base desde la cual sistemas de localización de ontologías

pueden ser desarrollados.

Las principales contribuciones de este trabajo se resumen como sigue:

- *Una caracterización y definición de los problemas de localización de ontologías*, basado en problemas encontrados en áreas relacionadas. La caracterización propuesta tiene en cuenta tres problemas diferentes de la localización: traducción, gestión de la información, y representación de la información multilingüe.
- *Una metodología prescriptiva para soportar la actividad de localización de ontologías*, basada en las metodologías de localización usadas en Ingeniería del Software e Ingeniería del Conocimiento, tan general como es posible, tal que ésta pueda cubrir un amplio rango de escenarios.
- *Una clasificación de las técnicas de localización de ontologías*, que puede servir para comparar (analíticamente) diferentes sistemas de localización de ontologías, así como también para diseñar nuevos sistemas, tomando ventaja de las soluciones del estado del arte.
- *Un método integrado para construir sistemas de localización de ontologías en un entorno distribuido y colaborativo*, que tenga en cuenta los métodos y técnicas más apropiadas, dependiendo de: i) el dominio de la ontología a ser localizada, y ii) la cantidad de información lingüística requerida para la ontología final.
- *Un componente modular para soportar el almacenamiento de la información multilingüe asociada a cada término de la ontología*. Nuestra propuesta sigue la tendencia actual en la integración de la información multilingüe en las ontologías que sugiere que el conocimiento de la ontología y la información lingüística (multilingüe) estén separados y sean independientes.
- *Un modelo basado en flujos de trabajo colaborativos* para la representación del proceso normalmente seguido en diferentes organizaciones, para coordinar la actividad de localización en diferentes lenguajes naturales.
- *Una infraestructura integrada* implementada dentro del NeOn Toolkit por medio de un conjunto de plug-ins y extensiones que soporten el proceso colaborativo de localización de ontologías.

Contents

1	Introduction	1
1.1	Motivation	3
1.2	From Monolingual to Multilingual Ontologies	4
1.3	Context of the Thesis	6
1.4	Ontology Localization Approach	9
1.4.1	Automatic Discovery of Translations	10
1.4.2	Collaborative Localization Management	12
1.4.3	Modular Storage of the Linguistic Information	14
1.4.4	Automatic Synchronization Process	15
1.4.5	Prescriptive Methodological Guidelines	16
1.5	Structure of the Thesis	16
2	Technological Context	19
2.1	Ontologies	19
2.1.1	Ontology basics	19
2.1.2	Development of Ontologies	21
2.2	Machine Translation	22
2.2.1	MT and Localization	22
2.2.2	Classification of MT Systems	23
2.3	Methods for the Building of Multilingual Ontologies	25
2.3.1	Ab initio construction	26
2.3.2	Merging of Existing Ontologies	29
2.3.3	Translation of Monolingual Ontologies	31
2.4	Summary of the Chapter	37
3	Work Objectives	39
3.1	Goals and Open Research Problems	39
3.2	Contributions to the State of the Art	41
3.3	Work Assumptions	42
3.4	Hypotheses	43
3.5	Restrictions	44

4	Ontology Localization Problem	45
4.1	Definition of Terms	45
4.1.1	Ontology Localization Definition	46
4.1.2	Related Terms	48
4.2	Characterization of Ontology Localization	50
4.2.1	Language Equivalence Problems	51
4.2.2	Translation Problems	51
4.2.3	Management Problems	52
4.2.4	Multilinguality Representation Problems	53
4.3	Scales of the Ontology Localization Activity	54
4.4	Ontology Localization Approach	57
4.4.1	Scenarios for Localization	58
4.4.2	Automatic Localization Approach	59
4.5	Summary of the Chapter	69
5	Translation Techniques for Ontology Localization	71
5.1	Classification of Translation Techniques	71
5.1.1	Term Context Interpretation	75
5.1.2	Type of Resources Used	79
5.2	Basic Translation Techniques	84
5.3	Online MT-based Techniques	86
5.3.1	Methods Employed	86
5.3.2	Advantages	87
5.3.3	Disadvantages	87
5.4	Knowledge-based Techniques	87
5.4.1	Methods Employed	88
5.4.2	Advantages	90
5.4.3	Disadvantages	90
5.5	Corpus-based Techniques	90
5.5.1	Methods Employed	90
5.5.2	Advantages	97
5.5.3	Disadvantages	97
5.6	Semantic-based Techniques	98
5.6.1	Methods Employed	98
5.6.2	Advantages	99
5.6.3	Disadvantages	99
5.7	Analysis of Translation Techniques	99
5.8	Ontology Localization Strategies	100
5.8.1	Translation Composition	101
5.8.2	Translation Combination	102
5.9	Classification Guidelines for Ontology Localization Approaches	104
5.9.1	Type of Localization.	104
5.9.2	Localization Process.	104
5.9.3	Output.	105

5.9.4	Use case.	106
5.10	Summary of the Chapter	107
6	Lyfe-Cycle Model and Architecture	109
6.1	Ontology Localization Life-Cycle	110
6.1.1	The Automated Ontology Localization Model	110
6.1.2	Automated Localization Cycle	111
6.1.3	Data Structures.	113
6.2	Key Requirements for an Ontology Localization Infrastructure	114
6.2.1	Requirements for Collaborative and Distributed Localization Activity	114
6.2.2	Requirements to Support Automatic Ontology Translation	118
6.2.3	Requirements for an Extensible Ontology Localization Infrastructure	119
6.3	Global Description of the Architecture	120
6.4	The Ontology Management Module	122
6.5	The Localization Management Module	124
6.5.1	Synchronization Component	125
6.5.2	Localization Component	127
6.6	The Ontology Translation Module	129
6.6.1	Leverage Component	130
6.6.2	Translator Component	131
6.7	LabelTranslator System	135
6.7.1	Technical Details of the LabelTranslator System	135
6.7.2	Ontology Management	135
6.7.3	Localization Management	139
6.7.4	Ontology Translator	142
6.8	Translation Strategies Used in LabelTranslator	142
6.8.1	Translating Simple Labels	143
6.8.2	Translating Compound Labels	149
6.9	Summary of the Chapter	152
7	Methodological Guidelines	155
7.1	Neon Methodology as Framework for the Localization of Ontological Resources	155
7.2	Research Methodology	156
7.2.1	Analysis of Relevant Methodologies	157
7.2.2	Identification of Main Task	163
7.2.3	Task Completion	163
7.2.4	Analysis of Task Dependencies	163
7.3	Methodological Guidelines for Ontology Localization	164
7.3.1	Ontology Localization Actors	164
7.3.2	Ontology Localization Guidelines	166

7.4	Summary of the Chapter	172
8	Experimentation	175
8.1	Translation Algorithm Evaluation	175
8.1.1	Quality Evaluation of Ranking Method	176
8.1.2	Translation Quality Evaluation	180
8.1.3	Translation Techniques Evaluation	185
8.1.4	Translation Resources Evaluation	189
8.2	Collaborative Localization Evaluation	191
8.2.1	Overview and Objectives	191
8.2.2	Experiment Setting	192
8.2.3	Findings and Observations	194
8.2.4	Identified Strengths and Weaknesses	196
8.3	Methodological Evaluation	196
8.3.1	Usability of the Methodological Guidelines	197
8.3.2	Methodological Guidelines Evaluation Through Use Cases	199
8.4	Summary of the Chapter	210
9	Conclusions	213
9.1	Main Contributions	213
9.1.1	Identification and Implementation of the Technologi- cal Support for Ontology Localization	214
9.1.2	Development and Use of the Localization Methodology	217
9.2	Evaluation of Results	219
9.3	Future Challenges	220
	Relevant Publications Related to the Thesis	223
	Bibliography	225
A	Ontology Localization Framework Evaluation	251
A.1	Efficiency	251
A.2	Affect	252
A.3	Helpfulness	254
A.4	Control	255
A.5	Learnability	257
B	Localization User Guides	259
B.1	User Guide for Localization Managers	259
B.1.1	Installing the Environment.	259
B.1.2	Setting-up Ontology Localization Preferences.	263
B.1.3	Importing Ontology to be Localized.	265
B.1.4	Setting-up Localization Parameters.	267
B.1.5	Selecting Ontology Labels to be Translated.	270

B.2	User Guide for Translators	271
B.2.1	Setting-up Translation Preferences.	271
B.2.2	Translating Ontology Labels.	272
B.3	User Guide for Reviewers	274
B.3.1	Setting-up Revision Preferences.	274
B.3.2	Reviewing Translations.	276

List of Figures

1.1	Workflow of a typical localization project. Adapted from [Eselink, 2000]	8
1.2	Localization levels for systematic product internationalization. Adapted from [Sturm, 2002].	10
1.3	Automatic translation approach for the Ontology Localization Activity [Espinoza et al., 2009a]	11
1.4	Workflow process used to localize an ontology.	13
2.1	Different levels of analysis in an MT system.	24
2.2	The global architecture of the EWN database [Vossen, 2004].	27
2.3	Multilingual ontology deployment in the DOSE platform.	28
4.1	The Ontology Localization Activity	47
4.2	Ontology Localization levels.	55
4.3	Ontology Localization scenarios.	58
4.4	Human translator steps.	59
4.5	Automatic translation approach for the Ontology Localization Activity.	61
5.1	Classification of ontology localization techniques.	74
5.2	Ontology Example.	77
5.3	Parallel combination of translation algorithms.	102
5.4	Sequential composition of translation algorithms.	103
6.1	The Automated Ontology Localization Life-Cycle Model.	111
6.2	Typical Ontology Localization Scenario.	117
6.3	General architecture to support Ontology Localization.	121
6.4	Workflow process used to localize an ontology.	125
6.5	Synchronization of ontology and linguistic model.	127
6.6	Workflow to the Translation level.	128
6.7	Detailed Ontology Translator in a Localization System.	130
6.8	Ontology Navigator with a selected ontology element.	137
6.9	Connecting the Ontology Model with the Linguistic Model (taken from [Montiel-Ponsoda, 2011b]).	138
6.10	Linguistic Information page that support the LIR model.	139

6.11	User wizards used by the Workflow Localization Manager. . . .	140
6.12	A perspective of the Ontology Localization Activity.	142
6.13	Extract of the sample university ontology.	143
6.14	LabelTranslator strategy to localize concept, attributes and relation terms represented by simple labels.	143
6.15	Some translations of the ontology label “chair” into Spanish. . . .	146
6.16	Context of the ontology label “chair”.	147
6.17	LabelTranslator strategy to localize concept, attributes and relations represented by compound labels.	149
6.18	Algorithm to translate the compound label “AssociateProfes- sor” into Spanish.	152
7.1	NeOn Methodology scenarios for building ontology networks). . . .	156
7.2	Actors involved in the Ontology Localization Activity.	166
7.3	Ontology Localization Filling Card.	167
7.4	Tasks for Ontology Localization.	168
8.1	Level of correctness in label translations.	183
8.2	Type of errors found in the translation of ontology labels. . . .	184
8.3	LabelTranslator Configuration for Collaborative Ontology Lo- calization.	193
8.4	Results of SUMI Questionnaire for LabelTranslator.	195
8.5	Related items for the term “pest control” extracted from FAOTERM.	201
8.6	Google definitions of the ontology label “pest control”.	201
8.7	Uses and “possibly” translated documents of the ontology label “pest control”.	204
8.8	Final Ontology using an external module.	206
8.9	Extract of the sample economy activity ontology.	207
8.10	Screenshot of the Ontology Navigator view with the Translate action used by the LabelTranslator plug-in.	207
8.11	Some translations of the Ontology label “Bars” into Spanish. . . .	208
8.12	Equivalent Translations for the Term “Bars”	209
8.13	Linguistic Information associated to the Ontology Term “Bars”	210

List of Tables

4.1	Exact equivalent sample	68
4.2	Near equivalence sample	68
4.3	Partial equivalence sample	68
6.1	Some lexical templates to translate a compound label from English into Spanish.	151
7.1	Common tasks in software localization methodologies	163
7.2	Ontology Localization task and its corresponding tool.	169
8.1	Ontologies corpus statistics.	177
8.2	Five point scale for fluency and adequacy measures.	179
8.3	Results obtained in the three experiments for Spanish.	179
8.4	Results obtained in the three experiments for German.	179
8.5	Multilingual Ontologies used in the evaluation.	188
8.6	Translation Techniques Comparison.	189
8.7	Translation Resources Comparison.	191
8.8	Linguistic information related to the area of “pest control”.	202
8.9	Linguistic information related to the area of “control (of a pest)”	203
8.10	Ranked translations of the term “pest control” for French and Italian	205
8.11	Semantic fidelity evaluation results.	209

Chapter 1

Introduction

In the context of the Semantic Web [Berners-Lee et al., 2001], resources on the net can be enriched by well-defined, machine understandable metadata describing their associated conceptual meaning. Ontologies constitute the foundation upon which to build the whole new generation web, and describe human knowledge by specifying concepts related to many specific areas of interest and by modeling relationships between them.

As with the World Wide Web (WWW) [Berners-Lee et al., 1992], the success or failure of the Semantic Web will be determined to a large extent by easy access to, and availability of high-quality and diverse content [Benjamins et al., 2002]. In this respect, an important challenge that needs to be addressed is the multilingualism problem, which until now has not been properly investigated [Tjoa et al., 2005]. This problem already exists in the current Web, and should also be tackled in the Semantic Web. Studies on language distribution over WWW content show that even if English is the predominating language for documents, there exists an important amount of resources written in other languages, according to the following distribution: English 26.8%, Chinese 24.2%, Spanish 7.8%, Japanese 4.7%, Portuguese 3.9%, German 3.6%, Arabic 3.3%, French 3.0%, Russian 3.0%, Korean, 2.0%, other languages 17.8%¹. In the case of the Semantic Web the problem is similar: most of the ontologies that have been built so far have English as their basis.

Nevertheless, although English is now the *de facto* language for science and technology, other spoken languages are used and it is important to provide methods and tools both to support the definition of ontologies expressed in languages other than English and also to support interoperability across ontologies written in different languages. However, looking at the statistics of two well-known gateways of the Semantic Web as Watson² and OntoS-

¹Obtained on May 31, 2011 from <http://www.internetworldstats.com>

²<http://watson.kmi.open.ac.uk/WatsonWUI>

elect³, we can observe that the number of multilingual ontologies available on the Web is insignificant compared with the number of monolingual ontologies.

The work presented in this thesis proposes an alternative to build a multilingual ontology through automatic⁴ localization of an ontology. The notion of localization comes from Software Development where it refers to the adaptation of computer software to non-native environments. From an Ontology Engineering perspective, localization makes it possible to adapt an ontology to different languages and cultures [Suárez-Figueroa and Gómez-Pérez, 2008]. This definition has been subsequently revisited in [Cimiano et al., 2010] to refer to “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 geopolitical environment”.

We should note here that the starting point of this work in 2006 was the NeOn Methodology (see [Suarez-Figueroa, 2013]) designed in the framework of the NeOn project⁵. This methodology identifies nine flexible scenarios that covers commonly occurring situations in the ontology development process. This thesis and the research work presented in [Montiel-Ponsoda, 2011b] began at the same time to support the Scenario 9: Localization of Ontologies. A part of the problem definition was carried out together, then, the thesis of Montiel-Ponsoda derived to the definition of a model to represent multilingual information in ontologies (LIR) [Montiel-Ponsoda et al., 2011], whereas this work focused on the methods, techniques, and tools for supporting the automatic localization of ontologies.

Our main contributions are: i) the definition and characterization of the ontology localization problem, ii) the identification and implementation of the methods, techniques and tools for the automatic management of ontology localization in collaborative and distributed environments, and iii) the definition of a methodology to support the ontology localization activity.

In this chapter we first describe the motivation from which this work arises. Secondly, we explain briefly the current trends for transforming a monolingual resource to multilingual. Thirdly, we introduce some features of the software localization industry that we believe are valid in the ontology localization context. Fourthly, we introduce the main features of our proposal to perform the localization of an ontology. Finally, we present the structure of this thesis.

³<http://olp.dfki.de/ontoselect/>

⁴We will use the term *automatic* to refer to an efficient process that allows reducing the human effort of translating a domain-specific ontological resource. We are conscious that the specificity of vocabulary terms in most ontologies precludes fully-automatic translation using general-domain translation resources.

⁵www.neon-project.org

1.1 Motivation

Currently, a great effort has been done in the construction of ontologies. Although access to top-quality ontologies (e.g., Galen⁶, CYC⁷, or AKT⁸) is in many cases free and unlimited for users all around the world, most of these ontologies can be said to be essentially monolingual, i.e., documented in one natural language only, and this language is often English as an international lingua franca. However, there is a growing need for multilingual ontology resources that overcome communication barriers arising from cultural-linguistic differences, lack of excellent command of English, need for high precision in communication, etc. In fact, multilingual knowledge is even more prevalent in those countries that have more than one official language [Yang and Li, 2003]. For example, Chinese and English are official languages of Hong Kong; French and English for Canada; and Dutch, French, and German for Belgium.

Moreover, the use of ontologies has grown not only in terms of the number of application domains but also in the number of natural languages chosen to build domain specific knowledge bases. Thus, multilingual ontologies are nowadays demanded by institutions worldwide with a huge number of resources available in different languages. Basically, usage of multilingual ontologies traverses many disciplines, and has become an urgent need in certain organizations. For instance, in Agriculture, the Food and Agriculture Organization (FAO) has expressed the need for semantically structuring the information they have in different natural languages. Since all FAO official documents must be made available in Arabic, English, Chinese, French, Russian and Spanish, a large amount of research has been carried out in translating large multilingual agricultural thesauri [Chun and Wenlin, 2002], in mapping methodologies for thesauri [Liang et al., 2005, Liang and Sini, 2006], and in defining requirements to improve the interoperability of these multilingual information resources [Caracciolo et al., 2007]. In Education, the Bologna declaration has introduced an ontology-based framework for qualification recognition [Vas, 2007] across the European Union, in an effort to best match labor markets with employment opportunities. In E-Learning, educational ontologies are used to enhance learning experience [Cui et al., 2004], and to empower system platforms with high adaptivity [Sosnovsky and Gavrilova, 2006]. In the Finance domain, ontologies are used to model knowledge in the stock market domain [Alonso et al., 2005] and portfolio management [Zhang et al., 2002]. In Medicine, ontologies are employed to improve knowledge sharing and knowledge reuse. For example, a notable amount of research has focused on the creation of an ontology of traditional Chinese medicine.

⁶www.co-ode.org/galen/

⁷<http://www.opencyc.org/downloads>

⁸<http://www.aktors.org/publications/ontology/>

A further factor that has increased the need for multilingual ontologies is the development of some ontology-based systems that need to interact with information in natural languages. Some examples of these applications are: cross-lingual information retrieval [Guyot et al., 2005], multilingual question answering [Pazienza et al., 2005] or knowledge management [Segev and Gal, 2008]. These examples can serve to highlight the importance of adding multilingualism/multilingual information to ontologies, before trying to solve the numerous pending problems that still exist with the current monolingual approach.

It is worth mentioning that at the time of starting this thesis there was no well-defined and broadly accepted definitions of what the ontology localization activity entailed. In fact, from the progress made in this work, other authors introduced new adaptations of the definition of localization of ontologies, putting emphasis on adaptation of ontology to the needs of the target community (see [Cimiano et al., 2010]). In 2010, the EU Multilingual Ontologies for Networked Knowledge project (Monnet) is created to continue the research and implementation of services that allow the automatic localization of ontologies to different languages. Some of the translation services implemented in Monnet are based on the results obtained during the development of this thesis. Furthermore, it is important to emphasize that from July 2011, the Ontology Lexica Community Group⁹ at the World Wide Web Consortium (W3C) is working mainly on the development of models for the representation of lexica (and machine readable dictionaries) relative to ontologies. The development of these models can help to enhance existing language resources by linked data principles, and improve the performance of applications as varied as ontology localization, machine translation, information extraction, multilingual access and presentation and natural language generation [McCrae et al., 2012].

Finally, to our knowledge, no other study has focused on the techniques, methods and tools for this activity. For this reason, in this thesis we present our approach for localizing an ontology into different natural languages.

1.2 From Monolingual to Multilingual Ontologies

Over the last decade, research on ontologies was concentrated on methodologies and technologies for supporting the creation and management as well as the population of ontologies. There are some well recognized methodological approaches (e.g., METHONTOLOGY [Fernández-López et al., 1999], On-To-Knowledge [Staab et al., 2001], DILIGENT [Pinto et al., 2004], and NeOn Methodology [Suárez-Figueroa, 2010]) that provided guidelines to help researchers to develop ontologies. However, most of existing methodologies

⁹<http://www.w3.org/community/ontolex/>

1.2. FROM MONOLINGUAL TO MULTILINGUAL ONTOLOGIES

and technologies focus on supporting knowledge management in monolingual environments.

The broadening of ontologies from single language to multiple languages presents new challenges. Problems related to *multilingual aspects* (human communication, and communicating across cultures) and *ontology construction* (equivalence and structural problems, and human effort). The major part of these problems already has been considered in the building of other multilingual resources. This is the case with vocabularies and thesauri. A thesaurus is a list of controlled vocabulary, or keywords, organized in a hierarchical structure. Its purpose is to facilitate information retrieval and documentation. The multilingual thesauruses are nowadays commonly developed for the use on different domains. For example, the European Union (EU) has developed the multilingual Eurovoc thesaurus [Steinberger et al., 2002] to cover mainly the activities of the European Parliament; the Council European Social Science Data Archives (CESSDA) has constructed the multilingual European Language Social Science Thesaurus (ELSST) [Miller and Matthews, 2001] to facilitate the access to data resources across Europe, independent of domain, resource, language and vocabulary; and the FAO organization has carried out a large amount of research in the translations of large multilingual agricultural thesauri [Chun and Wenlin, 2002].

Currently, the great majority of the methods and techniques used for the building of multilingual thesauri have been adapted to ontologies. However, given the more complex structure and sophisticated concept relationships of ontologies on thesauri, these approaches still are open subjects of research. The two main trends are:

- *Merging of existing monolingual ontologies.* This approach uses the integration of language-specific ontologies via ontology merging techniques. However, developing a multilingual semantic framework using this approach, involves the risk of getting an unmanageable entity as an outcome, in which great care is required to define relationships between “equivalent ontologies” and to track changes and coherently update those relations [Bonino et al., 2004].
- *Reconciliation of a common set of concepts.* This approach relies on the building of a common set of concepts that could be shared by different languages. A common misunderstanding that usually affects this approach and criticizes its adoption is the great amount of material and personal resources (ontology engineering, linguist, and domain experts) used for completing the task (an example of this approach is EuroWordNet [Vossen, 1997]).

The main features and limitations of the works that use the above mentioned methods for the building of multilingual ontologies are described in Section 2.3. Considering, that the current approaches do not reduce the

cost and effort that means enriching an ontology with multilingual information, the present work introduces a novel approach for the multilingualism problem; presenting the methods, techniques and tools for the localization of ontological resources.

1.3 Context of the Thesis

The work presented in this thesis belongs to the domain of Ontological Engineering, overlapping other related fields such as Software Localization, Semantic Web, and Machine Translation (MT). In this section we do a brief review of the main aspects related to Software Localization, because, although our goal is related to the localization of ontological resources, it is necessary to analyze the more relevant features of this area to apply them to our particular purpose, indicating the differences considered. Three are the main aspects that we consider: 1) terminology related, 2) main activities, and 3) levels of localization. The first two issues have been studied in detail by different authors (see by example [Esselink, 1998, Esselink, 2000]) and the last one in [Kersten et al., 2002].

The other related fields Semantic Web and Machine Translation are part of the technological context in which we have developed our research and which will be reviewed in Chapter 2.

Terminology Related

The term localization refers to the process of adapting, translating, and customizing a product (software) for a specific market, taking into consideration the cultural conventions. In terms of software localization, it implies more than just the mere translation of the product's user interface, but also the production of interfaces that are meaningful and comprehensible, so that they can reach a larger audience. The effectiveness and efficiency of the localization process often depends on other activities, in particular globalization, internationalization, and translation.

The term GILT was coined by Cadieux and Esselink [Cadieux and Esselink, 2004] and stands for globalization, internationalization, localization, and translation. According to the Localization Industry Standards Association (LISA) ¹⁰:

- *Globalization* involves changing the way an organization does business. It is more than a technical process and involves both internationalization and localization. More specifically, globalization is the strategy of bringing an internationalized and localized product or service to the global market; thus globalization involves sales and marketing.

¹⁰<http://www.lisa.org/What-Is-Globalization.48.0.html>, October 27, 2010

1.3. CONTEXT OF THE THESIS

- *Internationalization* encompasses the planning and preparation stages for a product in which it is built by design to support global markets. In other words, internationalization makes sure that the product or service is functional in any language and content.
- *Translation* refers to the specifically linguistic operations, performed by human or machine, that actually replaces the expressions in one natural language into those of another.

In the new context, ontology localization could be referred to only by the acronym ILT - Internationalization, Localization and Translation, because considering the definition of the term *globalization*, it would take place outside the Ontological Engineering field, which defines the set of activities that concern the ontology development process, the ontology life cycle, and the methodologies, tools and languages for building ontologies. Both approaches, Software Localization and Ontology Localization, have a very pragmatical and economical orientation, since the idea is to reuse software products or ontologies already available instead of developing them from scratch. And in both approaches, the starting point is a “product” created within a certain culture and in a certain language, i.e., a monolingual product [Montiel-Ponsoda, 2011a].

In Software Localization, the original product is adapted to different cultural communities and the result will be normally used independently from the original product. In Ontology Engineering, the localized ontology may be used independently from the original ontology, or it may also happen that the ontology is expected to support an application in which several natural languages need to interoperate. In the latter case, the output will be a multilingual ontology [Montiel-Ponsoda, 2011a].

Main Activities

According to Esselink [Esselink, 2000] translation is usually only one of the activities in a software localization project where material is transferred from a language to another. Examples of activities in localization which are not necessarily part of traditional translation include: multilingual project management, software and online help engineering and testing, conversion of translated documentation to other formats, translation memory alignment and management, multilingual product support, and translation strategy consulting. Figure 1.1 shows the workflow of a typical localization project.

Every project starts with the evaluation phase in which the received material is analyzed and checked. The aim of this phase is to determine the size of the project by using for example, word counts, and identify black spots that need special knowledge or skills to be taken care of. The next phase is the project setup in which the project manager recruits team members, finds other resources and makes schedules. In this phase, the project terminology

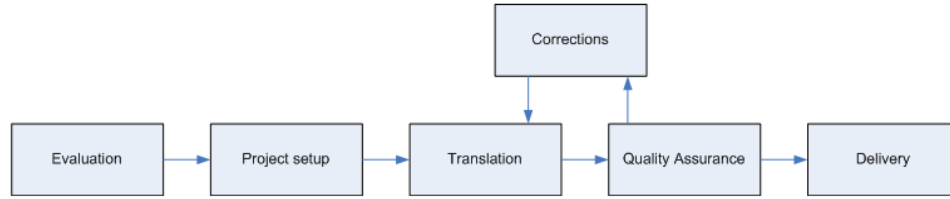


Figure 1.1: Workflow of a typical localization project. Adapted from [Esselink, 2000]

is usually created and required technical adjustments are done. After the project setup, translators start working on the material to be translated. The next phase, quality assurance, includes all methods (e.g., testing and proofreading) that are used to ensure that translations are consistent, appropriate and meet the customer's needs. Detected errors and bugs are then fixed before the delivery of the product.

Those phases have been re-adapted for the case of localizing an ontology. Furthermore, ontology localization activity has also been extended to dealing with distributed and collaborative scenarios. The current approach is based on the analysis of the process typically followed by organizations as FAO in the development and localization of ontologies. Also, to define the infrastructure requirements that support the whole life-cycle of the ontology localization activity, we took different key factors from different software localization approaches and we compared them with our own observations in the field. Concretely we analyzed three groups of requirements: collaboration and distribution of the tasks, automated translation, and extensibility.

Levels of Localization

Different levels of localization have been proposed in the Software Engineering area, depending mainly on: i) target audience, and ii) the amount of translation and customization necessary to create different language editions. The levels, which are determined by balancing complexity and need of further investigation, range from translating nothing to shipping a completely translated product with customized features [Sturm, 2002, MSDN, 2012]. Figure 1.2 shows the most accepted levels of localization:

- *Technical Level.* This level covers all technical aspects of a product. It includes the technical infrastructure and technical standards used in the foreign country the product has to be adapted to (e.g., ISO-norms for character sets, such as Unicode, ANSI, etc.). The adaptation of these issues ensures that the product works from the technical point of view and they are the basis for the next level.
- *Linguistic Level.* For most of the technical products the international

adaptations stop here, where different language versions are produced. The words and texts of the interface and manuals are translated and several aspects like punctuation, vocabulary and grammar are transferred, but often without the consideration of cultural differences.

- *Cultural Level.* The third level includes the cultural dimension of the use of products. It basically covers two areas: the context of the use and the meaning of symbols, graphics, colors, and metaphors used in the user interface. The cultural context of the product use and its position in the everyday life delivers the information concerning the required functionality. To put it in simple words, cultural localization is concerned about use of the icons, metaphors, message conventions, etc.
- *Cognitive Level.* The list of the required functions does not yet deliver the information concerning the question of how they should be presented to the user. The cognitive level therefore goes beyond the pure meaning of interface components covered by the cultural level. It encloses menu structures, priorities, interaction styles and techniques as well as basic cognitive processes used in human computer interaction. This level is undoubtedly the most underestimated, but it has a great impact on the usability of a technical product.

Taking into account that the *cognitive level* involves a conscious intellectual activity which only can be automated with difficulty, we do not consider that this level should be part of ontology localization activity. Moreover, a great part of the activities performed in this level of localization are usually considered at the time of designing an ontology. In the new approach, these levels have been adapted taking into consideration the layers of the ontology that are affected in the localization process. The new levels of localization range from the linguistic adaption of the ontology to a particular language to a cultural adaption of the ontology to a specific geo-political and cultural environment.

1.4 Ontology Localization Approach

As mentioned in the introduction, the work presented in this thesis attempts to cover both the technological and methodological aspects of the ontology localization activity. In this context, many techniques, methods, and tools developed in other areas such as: building of multilingual resources (e.g., thesauri), machine translation, software localization, distributed software development, etc., can be re-adapted for our purposes.

Of course, these methods, techniques, and tools must now consider the semantic wealth of ontologies. In the following, a brief review is given of the main aspects that have taken into account for ontology localization.

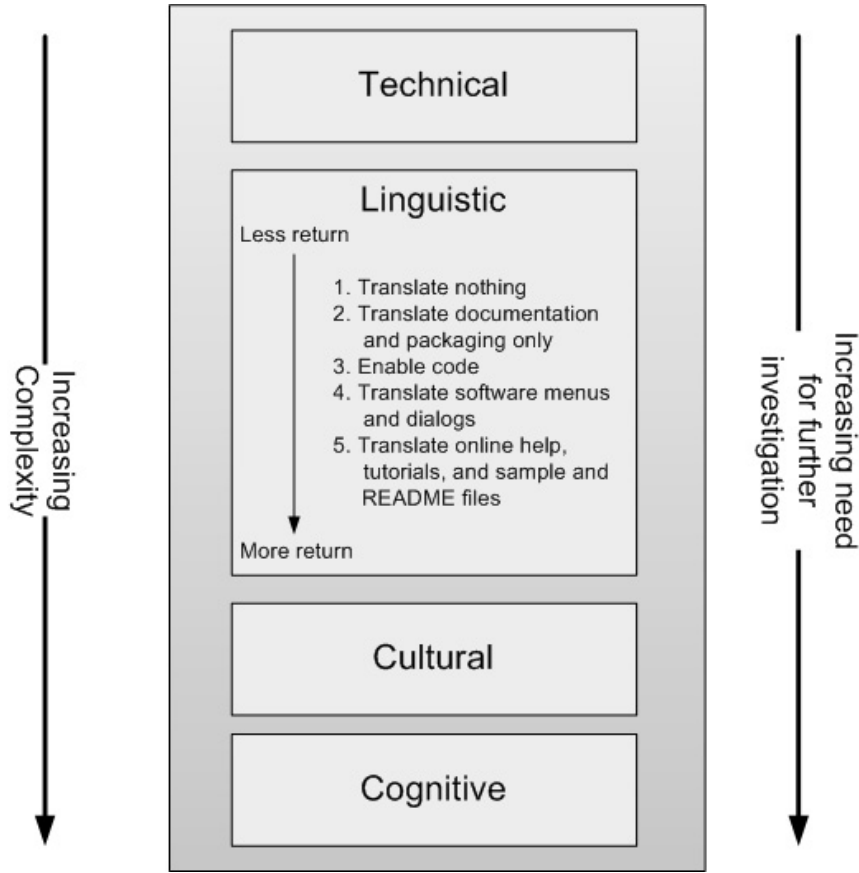


Figure 1.2: Localization levels for systematic product internationalization. Adapted from [Sturm, 2002].

Five are the major issues that we consider to support the localization of ontologies: 1) automatic discovery of translations; 2) collaborative localization management; 3) modular storage of the linguistic information; 4) automatic synchronization process; and 5) prescriptive methodological guidelines.

1.4.1 Automatic Discovery of Translations

As explained in section 1.3 a successful localization project is expensive. It involves manual work of different professionals, it requires a set of translators to discover the most appropriate translations, and it also needs reviewers to improve the quality of translations. Of these manual labors, the process of translation requires major effort [Yang, 2007]. In spite of this situation, automatic translation tools have not been used extensively in the localization industry, because the quality of these tools is still poorer [Esselink, 2000].

1.4. ONTOLOGY LOCALIZATION APPROACH

However, we believe that an automatic translation process is possible in the case of ontology elements, due that ontologies consisting of concepts and relationships that are stated clearly and succinctly [Espinoza et al., 2008a].

From our point of view this is the main contribution of Ontology Localization in this thesis. Four main steps are followed in order to discover the most appropriate translations: localization step selection, term context extraction, ontology label translation, and translation revision [Espinoza et al., 2008a, Espinoza et al., 2008b, Espinoza et al., 2009a].

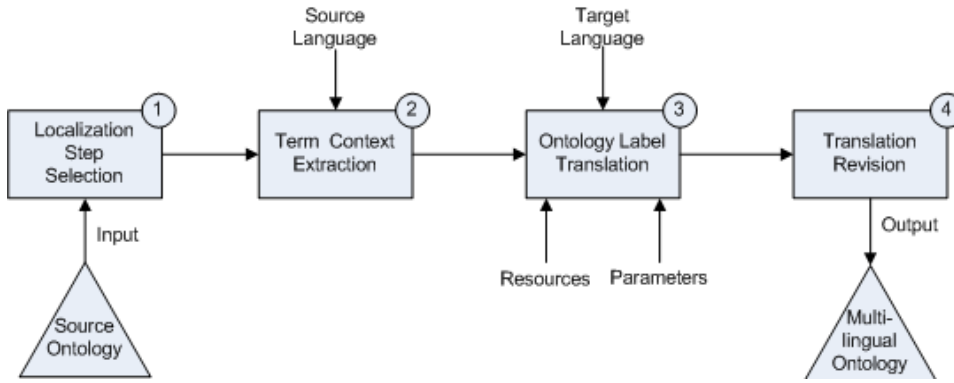


Figure 1.3: Automatic translation approach for the Ontology Localization Activity [Espinoza et al., 2009a]

- *Localization Selection.* The first step involves the selection of the ontology elements that need be localized to different natural languages. This task is especially important when only limited time or recourses are available. In such a case, it might be interesting to know which part of the ontology can best be translated.
- *Term Context Extraction.* Once the terms to be localized are identified, it is necessary to extract up to a certain depth the context of each ontology term to be localized. That is (depending on the type of term), their synonyms, textual descriptions, hypernyms, hyponyms, properties, domains, roles, associated concepts, etc. The context of an ontology term allows discerning among the different meanings that an ontology label may have.
- *Ontology Label Translation.* The goal of the label translation task is to discover the more appropriate translation of each ontology label. In

our approach, the task of translation is performed by the combination of different translation methods based on MT techniques. These techniques are combined following different translation strategies inspired from multi-engine machine translation approaches.

- *Translation revision.* A revision process is needed to evaluate the quality of the obtained translations. This task includes measuring the adequacy and fluency of all translations.

Notice that these steps cover only the translation task of the ontology localization activity. However, in this work we also describe the life-cycle model by means of representation of the major components of this activity and their interrelationships in a graphical framework that can be easily understood and communicated.

1.4.2 Collaborative Localization Management

Despite the high level of automation of our processes, we do not dream of full automation and we acknowledge that the human component is critical. We will always need highly qualified individuals to post-edit the MT output, to provide feedback, to perform a manual check of suggested translation resource entries and, most importantly, to encourage and assist one another. In fact, on most large projects today, localization is a collaborative effort, where the number of users participating in localization ranges from a handful to a couple of dozens.

Examples of such collaborative localization processes can be found in international institutions like the FAO, who have been developing and localizing the AGROVOC Thesaurus [AGROVOC, 2005], which is used widely to index agricultural information material all over the world. Thus, in the building of the AGROVOC Thesaurus, the translations were provided by the same thesaurus experts, then, a set of specialists whose main work is the translation of agricultural science literature were responsible for checking and approving the thesaurus translation work. With larger groups of users contributing to ontology localization, we believe that it is necessary to define appropriate workflows, strategies and infrastructure to support the process that coordinates the collaborative ontology localization within an organizational setting.

This process can be modeled as a collaborative workflow, describing how project participants reach consensus on ontology label translations, who can perform translations, who can comment on them, when ontology label translations become public and so on. The collaborative workflow that we propose in this thesis is designed to support all aspects of the ontology localization activity. However, the details of this process, as well as the configuration of the collaborative scenario can vary from one organization to another. Thus, in some scenarios the collaborative workflow may be

1.4. ONTOLOGY LOCALIZATION APPROACH

configured to omit the use of reviewers or may not perform the automatic localization of the ontological labels. In the following we summarize the main steps in the workflow process to localize an ontology (see Figure 1.4):

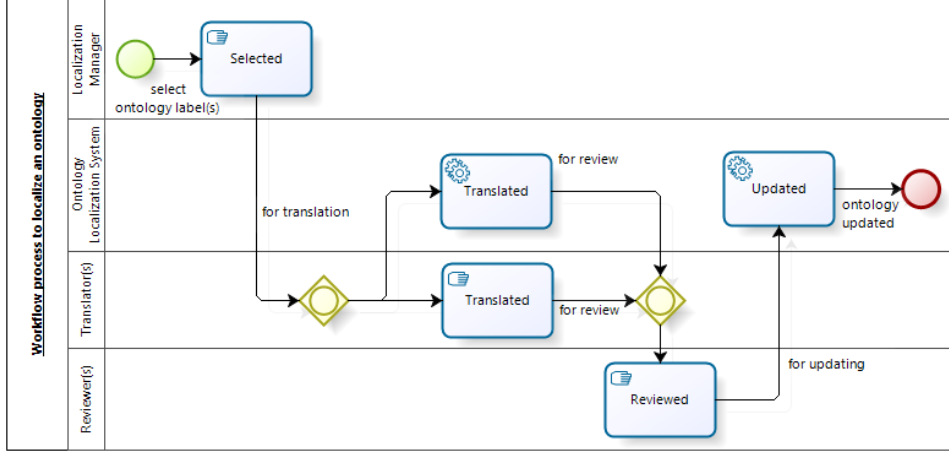


Figure 1.4: Workflow process used to localize an ontology.

- An ontology is passed to the Localization Manager for localization.
- The Localization Manager manually selects the ontology labels to be localized and sends the selected labels for translation.
- A translator downloads the selected labels to be localized and (s)he performs the translations using an automated localization tool (as proposed in this thesis) or an intensive manual process.
- Once translation activities have been accomplished, the translators upload the translated ontology labels and send them for review.
- The reviewers download the translated labels and check for possible errors.
- Finally, the Localization System updates all linguistic information of each localized label.

Our contributions about providing collaborative ontology localization can be summarized in the following points:

- Analysis of the main requirements to support a localization of ontologies in collaborative and distributed settings.
- Design of a formal model based on graphs for the representation of the activities usually followed by different organizations dedicated to localization.

- Design and implementation of a centralized repository to store the work of the different ontology stakeholders.
- Design of an architecture for managing versions of localized ontologies, controlling ontology access (through some form of check in/check out and file locking), and enabling remote or distributed access.
- Integration and implementation of an collaborative workflow to manage the sequence of translation/review/edit tasks, providing the status of tasks and processes, and notifying participants of changes in state, new work, or other information.
- Development of an architecture for supporting customizable workflows for collaborative ontology localization.
- Development of a set of interfaces that allows users collaboratively to perform the different tasks of the localization process.

All these aspects have been considered in the design of the LabelTranslator system, our approach to perform an automated localization in distributed and collaborative environments.

1.4.3 Modular Storage of the Linguistic Information

In her doctoral thesis, Montiel-Ponsoda [Montiel-Ponsoda, 2011a] analyzes the state of the art on models or formalisms to represent multilingual information in ontologies. She identifies three main ways of obtaining a multilingual ontology, depending on the layer(s) involved in the localization activity:

- *Including multilingual labels in the ontology* is the most widespread modeling option within the ontological community nowadays, because it is well supported by the most popular ontology development languages: RDF(S) [Brickley and Guha, 2000] and OWL [Bechhofer et al., 2004]. It consists of making use of the labeling functionality of RDF(S) and OWL ontology representation languages. In this case labels can be integrated in the ontology in as many languages as the user wishes. However, this approach does not permit to define any relation among the linguistic annotations themselves (e.g., saying that one is a synonym or translation of the other). This results in a bunch of unrelated data whose motivation is difficult to understand even for a human user.
- *Combining the ontology with a mapping model* assumes the existence of an original ontology and one or several ontologies localized to different natural languages, all of them represented as independent ontologies. The localized (monolingual) ontologies may have been obtained after performing the localization activity on the original ontology. This option enables independent conceptualizations in each language, what

may better capture the specificities of each culture, but the establishment of mappings or alignments among conceptualizations in different languages is by no means trivial, since mismatches arise due to each conceptualization capturing the cultural specificities of each language.

- *Associating the ontology with an external linguistic model* allows that the elements of the ontology have links to linguistic data stored outside the ontology. This type of representation allows the enrichment of domain ontologies with linguistically rich and complex models. Since these are external portable models, they can be associated to any domain ontology and published with them. Since there is just one conceptualization, this model is not as flexible as the previous one, in which cultural specificities were captured at the conceptual layer (despite the limitations imposed by interoperability and mapping discovery).

In this thesis, we follow the current trend in the integration of multilinguality in ontologies (third approach above), which suggests the suitability of keeping ontology knowledge and linguistic (multilingual) knowledge separated and independent. Three of the models that follow this trend are: LexInfo [Buitelaar et al., 2009], the Linguistic Information Repository (LIR) [Montiel-Ponsoda et al., 2011], and the Lexicon Model for Ontologies (Lemon) [McCrae et al., 2011b], which is being standardized in the W3C Ontology-Lexicon Community Group¹¹.

Our contribution here is to supply the support for a modular approach, in which the conceptualization is kept apart from the multilingual information [Espinoza et al., 2009a]. This representation form allows the inclusion of as much linguistic information as wished, as well as the possibility of establishing links among the linguistic elements within one language or across languages. In this sense, nuances or differences between languages can also be reported and even formalized in the terminological layer, in order to avoid the 100% equivalence correspondence among the different names of ontology elements. Relevant information as, e.g., the provenance of the linguistic elements, can also be included.

1.4.4 Automatic Synchronization Process

Whereas the translation process of ontology labels per se implies certain difficulties, the maintenance and updating of translated ontology labels throughout the ontology life cycle also requires special attention. The main difficulty in the management activity is to identify policies for managing changes in the ontology terms and their translated labels. In our case, this situation

¹¹<http://www.w3.org/community/ontolex/>

is even more complicated, because we provide a model where sets of ontology terms and linguistic information associated (in different languages) are separately stored (see previous aspect).

In order to keep both models synchronized we first need to find out exactly what has been changed in the ontology model, then find the equivalent places in the linguistic model, and only then start the updating. Thus, in this thesis we provide a comprehensive solution to the problems of managing the conceptual knowledge and the linguistic knowledge by means of synchronization techniques [Espinoza et al., 2009a].

1.4.5 Prescriptive Methodological Guidelines

The complexities of localization projects are very different from the complexities of software development projects. Unlike software development projects, in which well-established and precise practices and methodologies exist, the localization projects do not explain the localization process with the same style and granularity as those methodologies for developing software.

To facilitate the prompt assimilation of ontology localization by software developers and ontology practitioners, in this work we propose methodological guidelines in a manner non-oriented to researchers. We also include examples of how to use the guidelines in different cases. From a methodological point of view, our contributions can be summarized as follows:

- A characterization of the ontology localization problems.
- A study of the different strategies for representing multilingual information in ontologies.
- A prescriptive guideline to help users in the development of multilingual ontologies.

To the best of our knowledge, the study presented here is the first attempt to offer guidelines for the localization of ontologies.

1.5 Structure of the Thesis

This thesis is composed of nine chapters, including this one. Chapter 2 reviews the technological context in which this work has been developed. Particularly, we first review the aspects concerning to machine translation as the key to automatically discover the translations of the elements of an ontology. Then, we analyze the methods for the building of multilingual ontologies, followed by a description of some related works in order to do a comparison between different approaches and ours.

1.5. STRUCTURE OF THE THESIS

In Chapter 3 we provide a presentation of the objectives and contributions of the thesis. We also describe how this thesis can contribute to this field of research with the set of assumptions, hypotheses and restrictions taken into account.

In Chapter 4 we first explain the terminology related to ontology localization activity, providing the meaning that will be used for the distinct terms. Then, we propose a characterization of ontology localization based on the problems found in related areas. The chapter also describes the different scales of localization used in ontologies depending on the type of ontology elements to be localized and the level of adaptation required. Also, we analyze which elements or parts of the ontology are to undergo localization. After the foundations, we give a formal account of our general localization process by defining the input, output, and the four main steps identified.

In Chapter 5, we present a classification of different translation techniques based on the way of modeling the context used to disambiguate the candidate translations and the type of resources used to localize an ontology into different natural languages. Later in this chapter we introduce the main characteristics of different translation techniques. To facilitate the analysis of these translation techniques we introduced a framework that covers their main aspects. Then, we present at the strategic level, some natural ways to compose and combine the output of different translation techniques for obtaining ontology translations. Finally, we discuss an alternative for classifying the localization approaches.

Chapter 6 describe the life-cycle of the ontology localization activity. Then, it details the main modules to allow such an ontology localization approach in distributed and collaborative environments, but first it introduces some basic requirements for an ontology localization system. Finally, it describes general comments and different technical details related to the LabelTranslator system, our approach to performing an automated localization in distributed and collaborative environments.

Chapter 7 describes in detail the general methodology used to guide users in the development of multilingual ontologies. First, we describe the design principles contemplated when defining the methodology and the process followed to define it. Finally, it details the methodology by describing its actors, processes and tasks.

Chapter 8 is dedicated to the evaluation of our work according to the initial set of hypotheses. We describe a set of experiments that were carried out with the objective of evaluating the methodological and technological aspects of the localization activity. First, we describe the experiments used to evaluate some aspects related to the translation ranking techniques, where the task is to select the most appropriate translation of ontology labels. Then, we describe the study used to assess the usability of the LabelTranslator system for carrying out the Ontology Localization activity in distributed and collaborative environments. Finally, we describe two case

studies to measure the understanding and usability of the methodological guidelines.

In Chapter 9 we present the conclusions as well as our main contributions; we finish with the future research work.

Chapter 2

Technological Context

Along this chapter, and before presenting in detail our proposal to localize an ontology to different natural languages, we would like to describe first the technology related to our work. We start with an introduction to ontologies, describing what an ontology is from the perspective of Computer Science, and some related issues about the development of ontologies. Secondly, we present an introduction to Machine Translation (MT). We include the motivation for using MT in the localization of ontological resources and the classification of different MT systems. Finally, we describe the different methods used to build the multilingual ontologies, the main goal of the ontology localization activity. We analyze the strengths and drawbacks of each method to identify open research problems and work assumptions.

2.1 Ontologies

The Semantic Web [Berners-Lee et al., 2001] tries to achieve a semantically annotated Web, in which search engines can process the information contained in web resources from a semantic point of view, drastically increasing the quality of the information presented to the user. This approach requires a global consensus in defining the appropriate semantic structures (ontologies) for representing any possible domain of knowledge. In this sense, ontologies can be understood as the scaffolding of the Semantic Web. As ontology is one of the key terms in this thesis, this section will define its basics here.

2.1.1 Ontology basics

In [Studer et al., 1998] an ontology is defined as a formal, explicit specification of a shared conceptualization. *Conceptualization* refers to an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon. *Explicit* means that the type of concepts

used, and the constraints of their use, are explicitly defined. *Formal* refers to the fact that the ontology should be machine-readable. *Shared* reflects the notion that an ontology captures consensual knowledge, that is, it is not private, but accepted by a group. Other approaches have defined ontologies as explicit specifications of a conceptualization [Gruber, 1995] or as a shared understanding of some domain of interest [Uschold and Gruninger, 1996].

Different knowledge representation formalisms exist for the definition of ontologies. However, they share the following minimal set of components:

- *Classes*: represent concepts, which are taken in a broad sense, that is, they can represent abstract concepts (intentions, beliefs, feelings, etc) or specific concepts (people, computers, tables, etc). Classes in the ontology are usually organized in taxonomies through which inheritance mechanism can be applied.
- *Relations*: represent a type of association between concepts of the domain. Ontologies usually contain binary relations. The first argument is known as the domain of the relation, and the second argument is the range. Binary relations are sometimes used to express concept attributes. Attributes are usually distinguished from relations because their range is a data type, such as string, numeric, etc., while the range of a relation is a concept.
- *Instances*: are used to represent elements or individuals in an ontology.

There exist several categorizations of ontologies in function of a particular aspect (such as expressiveness [Lassila and McGuinness, 2001] or subject and type of structure [van Heijst et al., 1997]. An interesting classification was proposed by [Guarino, 1998], who classified types of ontologies according to their level of dependence on a particular task or point of view.

- *Top-level ontologies*: describe very general concepts like space, time, event, which are independent of a particular domain. It seems reasonable to have unified top-level ontologies for large communities of users. Some examples are Sowa's [Sowa, 1999], Cyc's [Lenat and Guha, 1989], and SUO [Pease and Niles, 2002]
- *Domain ontologies*: describe the vocabulary related to a generic domain by specializing the concepts introduced in the top-level ontology. There are several representative ontologies in the domains of e-commerce (UNSPSC¹, NAICS², SCTG³, RosettaNet⁴), medicine (GA-

¹<http://www.unspsc.org>

²<http://www.naics.com>

³<http://www.bts.gov/programs/cfs/sctg/welcome.htm>

⁴<http://www.rosettanet.org>

LEN⁵, UMLS⁶, ON9⁷), engineering (EngMath [Gruber, 1995], PhysSys [Borst, 1997]), enterprise (Enterprise Ontology [Uschold et al., 1998], TOVE [Fox, 1992]), and knowledge management(KA [Decker et al., 1999]).

- *Task ontologies*: describe the vocabulary related to a generic task or activity by specializing the top-level ontologies. For example, in the GIS domain, the task ontologies are used to enable knowledge sharing and reuse for interoperable GIS [OGC, 1996].
- *Application ontologies*: they are the most specific ones. Concepts in application ontologies often correspond to roles played by domain entities. The EBIs Experimental Factor ontology⁸, which is used to represent sample variables from gene expression experimental data, and the NIFSTD ontology⁹ which is composed of a collection of OWL modules covering distinct domains of biomedical reality, are two representative samples of these ontologies.

Our overall goal is to provide a concise methodology and implementation for localizing mainly domain ontologies. However, the general localization approach proposed in Chapter 4 can also be applied to other types of ontologies.

2.1.2 Development of Ontologies

The set of activities that concern the ontology development process, the ontology life cycle, the principles, methods and methodologies for building ontologies, and the tool suites and languages that support them, is called *Ontological engineering* [Gómez-Pérez et al., 2004]. Of the classification described above, the *Task* and *Domain ontologies* are the most complex to develop: on one hand, they are general enough as is required for achieving consensus between a wide community of users and, on the other hand, they are specific enough to present an enormous diversity with many different and dynamic domains of knowledge and millions of possible concepts to model.

Basically, the construction of domain ontologies relies on domain modelers and knowledge engineers that are typically overwhelmed by the potential size, complexity and dynamicity of a specific domain [Sánchez and Moreno, 2008]. In computing literature, various methodologies have been reported for developing ontologies (more details in [Gómez-Pérez et al., 2004]). Several tools such as Ontolingua [Farquhar et al., 1996], OilEd [Bechhofer

⁵<http://opengalen.org>

⁶<http://nih.gov/research/umls>

⁷<http://saussure.irmkant.rm.cnr.it/ON9/index.html>

⁸<http://www.ebi.ac.uk/efo/>

⁹<https://confluence.crbs.ucsd.edu/display/NIF/Download+NIF+Ontologies>

et al., 2001], Protégé [Noy et al., 2001], OntoEdit [Sure et al., 2002], Top-Braid [Allemang and Polikoff, 2004], and NeOn Toolkit [Hasse et al., 2008] are developed for the construction and management of ontologies. The most prominent of these are Protege and NeOn Toolkit. However, at the moment of writing this thesis only the NeOn Toolkit provides the technological and methodological support for localizing an ontology.

2.2 Machine Translation

In this section we first discuss the use of Machine Translation (MT) in the ontology localization activity. Then, we present a brief description of MT systems and how these systems can be distinguished according to different criteria. The decision to exclude lengthy discussions of MT in this thesis was made because of two reasons. First, MT has already been the subject of much discussion and is particularly well documented in literature. Second, we do not provide a contribution to the state of the art in MT, our intention only is to study the features of MT approaches and identifying the best-suited for the localization of ontologies.

Readers interested in learning more about MT can refer to [Wilks, 2009, Llitjs, 2009] for enlightening discussions of both linguistic and computational issues in MT research and development, as well as detailed descriptions of different approaches to MT and of specific MT systems.

2.2.1 MT and Localization

In general, the main goal of machine translation systems is to translate text from one language into another, with or without human assistance. A distinction should be made between systems where MT is used to help people understand foreign text, and where it helps to produce translations. Whereas in the former case the quality of the translation does not matter too much, as long as the meaning is preserved; in the latter setting of so-called computer-assisted translation systems, quality is the main concern. It is this latter kind of MT that we are dealing with in this thesis.

Although MT has not been used extensively in the localization industry, this situation has changed recently [Ruopp, 2010, Hudik and Ruopp, 2011]. New approaches and technology providers are emerging, and both clients and suppliers are giving a serious look into MT technology for localization. In fact, different pilot projects to assess the viability of MT technology for localization projects, showing positive results from the perspective of both cost and productivity have been successfully implemented [Muntés et al., 2012].

Below, we summarize the main issues that are changing the perception of using MT in the process of localization [Esselink, 2000, Global Vision, 2007, Vashee, 2007]:

- Some products contain massive amounts of information; therefore it is impossible for humans to translate them fast enough to meet their rapid expansions.
- Some technical terms are never or rarely used by users, however these terms can be obtained by means of automatic mechanisms of knowledge acquisition using available repositories on the Web.
- Commercial machine translation tools can produce very satisfactory results with the proper use of terminology and an adequate source text preparation.
- MT can produce results that do not require a lot of human interaction when the text is limited in vocabulary range and it uses a general sentence structure.

Most of the issues described above are applicable to ontological resources. In addition, we argue that MT can be a valid tool for automatic ontology localization since ontologies consist of concepts and relationships that are clearly and succinctly stated [Espinoza et al., 2008a]. This is the reason why in the following section a brief review of the main aspects related to MT systems is given.

2.2.2 Classification of MT Systems

According to [Och, 2002], MT systems can be distinguished under the following criteria: the type of the input text, the application, the level of analysis and the technology used.

Type of Input

For our purposes, the input text (*words or sentences*) can typically be expected to be grammatical and well-formed. The task is more complicated in case of a *speech translation system*. Then, the system has to deal with speech recognition errors and spontaneous speech phenomena such as ungrammatical utterances, or hesitations. Therefore, a speech MT system has to be able to deal with ‘wrong’ input. In this thesis, we deal principally with *word and simple sentence translation systems*.

Applications

There are various types of applications for MT technology. In *gisting*, the aim is to produce an understandable raw translation. A possible goal is that a human is able to decide whether a foreign language text contains relevant information. To extract information from the document, typically a human translation would be performed. In *post-editing* applications, the aim is to

produce a translation that is then corrected by a human translator. In *fully automatic MT*, the computer is used to produce a high quality translation. For our purposes, we are interested in *fully automatic applications* and only in some cases in *post-editing applications*.

Analysis

Typically, three different types of MT systems are distinguished according to the level of analysis that is performed. Figure 2.1 [Och, 2002] gives the standard visualization of the three approaches: *direct translation*, *transfer approach* and *interlingua approach*.

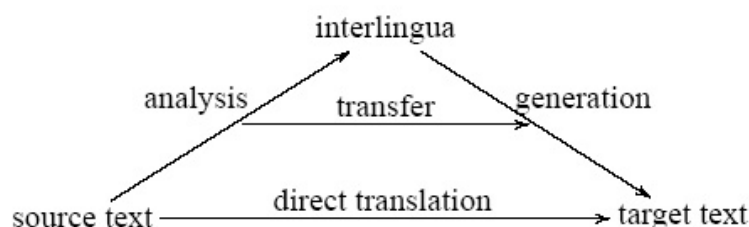


Figure 2.1: Different levels of analysis in an MT system.

The simplest approach is the *direct translation* approach in which a word-by-word translation from the source language to the target language is performed. In the *transfer approach*, the translation process is decomposed into three steps: analysis, transfer, and generation. In the analysis step, the input sentence is syntactically and semantically analyzed producing an abstract representation of the source sentence. In the transfer step, this representation is transferred into a corresponding representation in the target language. The target language sequence is produced in the generation step. In the *interlingua approach*, a very fine-grained analysis produces a completely language independent representation of the input sentence. This representation is used to produce the target language sentence.

An often claimed advantage of the *interlingua approach* is that developing translation systems between all pairs of a set of $n > 1$ languages is more efficient. There are only n components which need to be translated into the interlingua and n components are needed to translate from it. In a transfer approach or a direct translation approach, the development of $n * (n - 1)$ components for each pair of languages is needed [Och, 2002].

Despite apparent advantages, the interlingua based approach has been used far less widely than other approaches. Perhaps the reason for this lies in the heinous difficulty of defining a universal, or even widely applicable, interlingua. In addition, the construction of the necessary resources to an interlingual machine translation system continues to be a labor intensive

2.3. METHODS FOR THE BUILDING OF MULTILINGUAL ONTOLOGIES

process often resulting in knowledge-based systems that suffer from a lack of robustness. Such systems may work well on certain types of phenomena, but their complex knowledge-based foundation makes them difficult to extend to new phenomena or languages [Dorr et al., 2002]. For these reasons we decided to exclude the interlingual approach as a suitable solution for ontology localization.

Technology

MT systems can also be classified according to their core technology [Och, 2002, Costa-Juss, 2008]. Two kinds of systems can be distinguished: *rule-based* and *empirical-based* MT systems.

In the *rule-based machine translation* (RBMT) approaches, human experts specify a set of rules, which are aimed at describing the translation process. This is typically a very expensive work for which linguistic experts are needed. Using a *empirical machine translation* (EMT) approach, the knowledge is automatically extracted by analyzing translation examples from a parallel corpus. Within the empirical-based approaches we can further distinguish between *example-based* (EBMT) and *statistical MT* (SMT). In EBMT, a translation of a new sentence is performed by analyzing similar translation examples. In SMT, parallel examples are used to train a statistical translation model. The decision rule used to decide for the actual translation is derived from statistical decisions.

A major advantage of EMT approaches to MT is that these systems can be developed very quickly for new language pairs and domains [Och, 2002]. If consistency of terminology is the main factor to consider, then RBMT allows precise control of the terms used. To put this in simple words, if the aim is to use the technology to just to give a ‘gist’ of the meaning, EMT may be a more attractive option, but if the aim is to use an automated translation coupled with the skills of human translators, RBMT might be more useful.

Both technologies have their strengths and weaknesses; therefore, we believe that an appropriate combination of these and other technologies can be of great help for the localization of ontological elements.

2.3 Methods for the Building of Multilingual Ontologies

The multilingual processing is one of the most important issues that Computational Linguistics and Semantic Web faces; this is the reason why in the following section we give a detailed overview of the different methods that can be used to support the building of multilingual ontologies - the main goal of the ontology localization activity.

Basically, all methods used for the building of multilingual thesauri could be adapted in the building of multilingual ontologies. The ISO 5964 [ISO, 1985] (Guidelines for the Establishment and Development of Multilingual Thesauri) recognizes three approaches for the construction of these resources:

1. *Ab initio*¹⁰ *construction*, i.e. the establishment of a new thesaurus without direct reference to the terms or structure of any existing thesaurus. This method needs to be adopted when a new multilingual system is being established and an existing thesaurus does not already exist.
2. *Reconciliation and merging of existing thesauri*, i.e. two pre-existing monolingual thesauri of similar domain which are used in order to create a new multilingual thesaurus. This situation may occur if a new system is being formed on the basis of two or more pre-existing monolingual systems.
3. *Translation of an existing monolingual thesaurus*, i.e. a monolingual thesaurus covering the subject field of the proposed multilingual thesaurus, and serving as the source language.

These three approaches have been adopted in different ways to build multilingual ontologies. We would like to point out that the first two methods are the current trends in the ontology engineering field. The last approach is an emergent research field and it is the main focus of this thesis. After a short explanation of each approach, we will describe some relevant works.

2.3.1 Ab initio construction

In case of ontological resources, this localization procedure is generally adopted when: i) the ontology is being developed from the start and multilinguality is included at the same time or ii) the decision of building a multilingual ontology is taken during the first stages of the ontology development.

The common feature of the majority of works that adopt this localization procedure is the effort towards the construction of an upper-level conceptual core. This conceptual core makes on one hand the ontology under consideration accessible in many languages, and on the other allows ontology to represent many different cultures. Below we provide the most relevant works in this area.

Two well known efforts that adopt this procedure are EuroWordNet (EWN) [Vossen, 1997] and MultiWordNet (MWN) [Bentivogli et al., 2000]. EuroWordNet explicitly was based on and had the same structure of the

¹⁰Latin phrase meaning “from the start”; literal meaning being something done “from scratch”.

2.3. METHODS FOR THE BUILDING OF MULTILINGUAL ONTOLOGIES

Princeton WordNet¹¹ [Miller, 1995], developed as a monolingual lexical database for American English. The work initiated in the EWN project is now being continued by the Global Wordnet Association (GWA)¹².

The aim of EuroWordNet was to develop a multilingual lexicon with wordnets for several European languages (English, Dutch, Spanish and Italian), which could be used “to improve recall of queries via semantically linked variants in any of these languages”. The general approach for EWN was to build the multilingual database taking advantage of existing resources in each language. Participants from each country were responsible for a language specific wordnet using their already available tools and resources built up in previous national and international projects. As in WordNet, information about nouns, verbs, adjectives and adverbs was organized in synsets. A synset is “a set of words with the same part-of-speech that can be interchanged in a certain context” [Vossen, 2004]. Synsets are related to each other by semantic relations, such as hyponymy or meronymy, for example. The wordnets in EuroWordNet are considered “autonomous language specific ontologies”. Then, multilingual wordnets are interconnected through an Inter-Lingual-Index (ILI), a list of unstructured meanings mainly from Princeton WordNet, specifically WordNet1.5, that provide the mappings across the wordnets as illustrated in Figure 2.2.

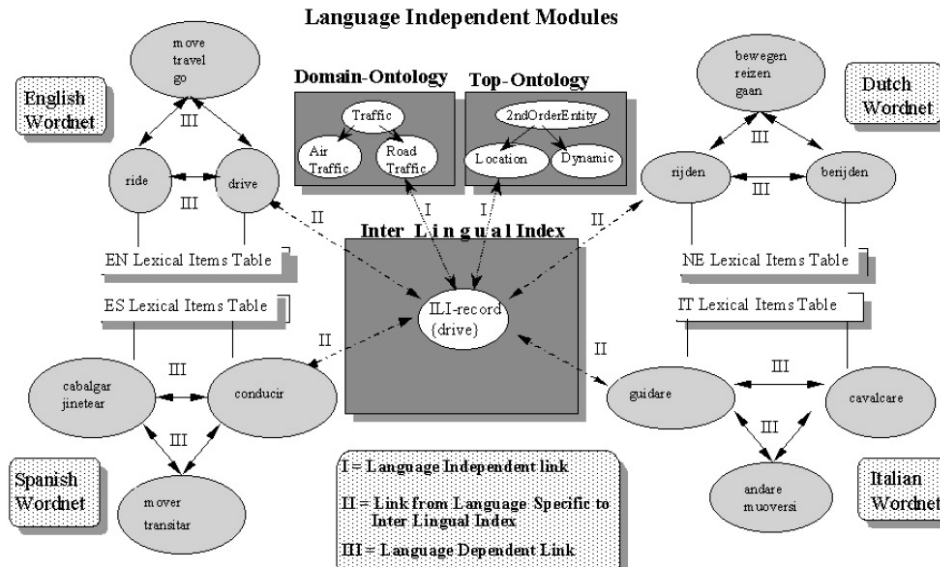


Figure 2.2: The global architecture of the EWN database [Vossen, 2004].

MultiWordNet (MWN) is a multilingual lexical database including information about English and Italian words. The model adopted within MWN,

¹¹<http://wordnet.princeton.edu/>

¹²http://www.globalwordnet.org/gwa/gwa_grid.htm

consists of building language specific wordnets keeping as much as possible of the semantic relations available in the WordNet. This was done by building the new synsets in correspondence with the WordNet synsets, whenever possible, and importing semantic relations from the corresponding English synsets; i.e., if there are two synsets in WordNet and a relation holding between them, the same relation holds between the corresponding synsets in the new language. The MWN model minimizes the discrepancies that can appear when two wordnets are built independently for two different languages, by strictly adhering to the WordNet building criteria and subjective choices. However, MultiWordNet explicitly recognizes the presence of “lexical gaps” in the correspondence between different languages, due to missing direct translations of some words.

Another approach is given by [Bonino et al., 2004], in which the authors introduce a simple approach to multilingual semantic elaboration using the Distributed Open Semantic Elaboration platform (DOSE). This approach uses a language independent ontology in which concepts are defined as high-level entities for which language dependent definitions are specified. Such entities are linked to a set of different definitions, one for each supported language, and a set of words that the authors call synset. Figure 2.3 shows the multilingual ontology deployment used in the DOSE platform.

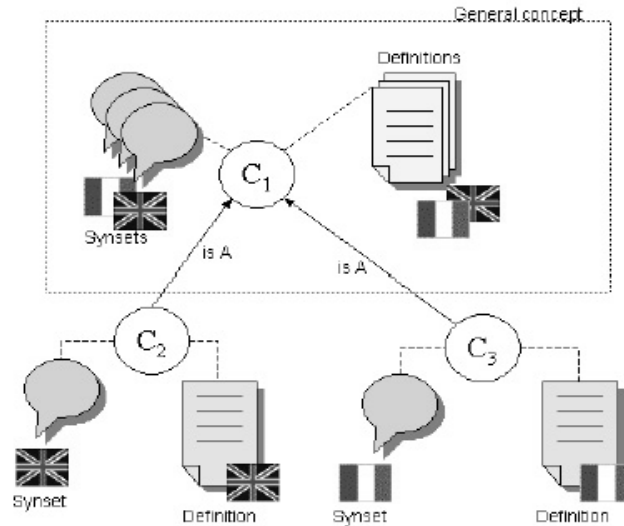


Figure 2.3: Multilingual ontology deployment in the DOSE platform.

The ontology is physically distinct from definitions and synsets, allowing separate management of concepts and language-specific information, isolating the semantic and the textual layers. This assumption guarantees sufficient expressive power to model conceptual entities typical of each language and, at the same time, reduces redundancy by collapsing all common

2.3. METHODS FOR THE BUILDING OF MULTILINGUAL ONTOLOGIES

concepts into a single multilingual entity. Synsets and textual definitions are created by human experts through an interactive refinement process. A multilingual team works on concept definitions by comparing ideas and intentions, aided by domain experts with linguistic skills for at least two different languages, and formalizes topics in a mutual learning cycle.

Finally, the work presented in [Segev and Gal, 2008] proposes an ontology-based model for building multilingual applications. Their model was based on a global ontology manually designed for a specific domain. Additionally, this model uses local context to specify the ontology. The combination of ontologies and contexts lends itself well to multilingual applications in which a single ontology fails to capture all nuances that stem from language and cultural differences. The procedure used for adapting an existing ontology to the needs of a multilingual environment includes the following four steps, *selection*, *collection*, *extraction*, and *adaptation*. In the selection step an existing ontology is chosen. In the collection step, sample documents that represent ontology concepts are collected. Contexts are extracted from sample documents in the extraction step. Finally, extracted contexts are associated with ontology concepts. The ontological system works simultaneously in multiple languages, and it is easily expandable and adaptable to other languages. As a final comment, we can say that some of the steps in this approach need intensive human labor.

Main advantages and shortcomings

An advantage of the works that adopt this approach is that it is easier to ensure language neutrality (i.e. lack of bias towards any one language). However, the costs of producing such ontology, as well as the definition and multilingual equivalence of its terms, have to be established a priori.

There are still two critical issues that need to be solved before the building of multilingual from scratch or other similar efforts can be used as a shared conceptual framework for all languages: the scarcity of lexical semantic information (especially from endangered languages), and the lack of a linguistically-motivated shared conceptual core as the basis of multilingual conceptual representation.

2.3.2 Merging of Existing Ontologies

This approach for localizing an ontology may be adopted when monolingual ontologies already exist in similar domains and therefore a multilingual ontology could be quickly and robustly constructed from monolingual resources. The aligning and merging of ontologies is actually one of the most active domains of investigation in the Semantic Web community [Euzenat and Shvaiko, 2007, Ehrig, 2007]. However, the issue of mapping ontologies written in different natural languages is still relatively unexplored.

The works that adopt this localization procedure try to identify the similarities between heterogeneous ontologies and then try to automatically create suitable mappings for transformation. Depending on the way that alignments between ontologies are discovered, these works can be grouped into two categories: cross-lingual ontology alignment approaches and generic approaches that involve machine translation tools and monolingual ontology matching techniques. Note that the ontology localization activity proposed in this thesis can contribute as a plausible solution for the approaches that use MT tools. More details can be consulted in the Section 5.9.4

Cross-lingual Alignment Approaches

Dorr *et al.* [Dorr et al., 2000] and Palmer & Wu [Palmer and Wu, 1995] took a structural approach to this problem. They focused on HowNet verbs and used thematic-role information, which denotes the contexts in which a particular verb may occur. The HowNet thematic-role specifications are mapped to word classes in an existing classification of English verbs called EVCA [Levin., 1993], whose structure is similar to that of the verb classes in HowNet. These mappings are then used to align English EVCA verbs to Chinese HowNet verbs.

In [Chen and Fung, 2004] the authors proposed an automatic technique to associate the English FrameNet lexical entries to the appropriate Chinese word senses. Each FrameNet lexical entry is linked to Chinese word senses of a Chinese ontology database called HowNet. First, each FrameNet lexical entry is associated with Chinese word senses whose part-of-speech is the same and Chinese word/phrase is one of the translations. In the second stage of the algorithm, some links are pruned out by analyzing contextual lexical entries from the same semantic frame. In the last stage, some pruned links are recovered if its score is greater than the calculated threshold value.

Carpuat et al. [Carpuat et al., 2002] merged thesauri that were written in English and Chinese into one bilingual thesaurus in order to minimize repetitive work while building ontologies containing multilingual resources. A language-independent, corpus based approach was employed to merge WordNet and HowNet by aligning synsets from the former and definitions of the latter. Similar research was conducted in [Malaisé et al., 2007] to match Dutch thesauri to WordNet by using a bilingual dictionary, and concluded a methodology for vocabulary alignment of thesauri written in different natural languages.

Monolingual Alignment Approaches with MT Support

Asanoma [Asanoma, 2001] aligned the Japanese Goi-Taikei ontology with WordNet by first translating a significant subset of the WordNet synonym sets (synsets) into Japanese, automatically matching these based on (mono-

2.3. METHODS FOR THE BUILDING OF MULTILINGUAL ONTOLOGIES

lingual Japanese) lexical overlap, and filling in the gaps for the remaining classes based on their hierarchical positioning relative to the aligned classes.

Trojahn *et al.* propose a multilingual ontology mapping framework in [Trojahn et al., 2008], which consists of smart agents that are responsible for ontology translation and capable of negotiating mapping results. In [Fu et al., 2009a] Fu *et al.* present the SOCOM Framework which is designed specifically to achieve cross-lingual ontology mapping. The SOCOM framework divides the multilingual mapping task into three phases: an ontology rendering phase, an ontology matching phase and a matching audit phase. The first phase of the SOCOM framework is concerned with the rendition of an ontology labeled in the target natural language, particularly, appropriate translations of its labels. The second phase concerns the generation of matching results in a monolingual environment. The third phase of the framework aids ontology engineers in the process of establishing accurate and confident mapping results.

Main Advantages and Shortcomings

The principal advantage of this approach is the existence of a great quantity of ontologies that can be used to build a multilingual ontology. However, some problems arise, i.e. the difference in the hierarchical structure of the relevant ontologies, as well as differences in the semantics of the terms. While equivalency is sought between terms, this does not imply that the hierarchical structures themselves must also be equivalent. Therefore, developing a multilingual ontology using this approach involves the risk of getting an unmanageable entity as an outcome, in which great care is required to define relationships between “equivalent ontologies” and to track changes and to coherently update those relations [Bonino et al., 2004].

2.3.3 Translation of Monolingual Ontologies

This localization approach is adopted when an ontology has to be built in a certain target language (e.g., English or Spanish) and there is a monolingual ontology covering the domain of the proposed multilingual ontology. Note that, this approach is the main focus of this thesis.

In this section we describe the main features of related works that we consider more relevant. But we first classify them according to the level of automatization and collaboration used to localize an ontology into different natural languages. Basically, the methods for translating an ontology can be grouped in four different categories: by hand, using a community, automatically and semi-automatically.

Translating an Ontology by Hand

One of the benefits of translating an ontology by hand lies in the total control of how the translation is being done, and for small ontologies the amount of work is acceptable. There are however some disadvantages to this method: the translation will be a somewhat subjective one, since the translation is a product of one or few people's skills, which is affected by their education, experience and temper. Furthermore, the amount of time and effort, and therefore expenses, to complete this task will be very high when translating larger ontologies. On the other hand, in cases of very specialized knowledge, this approach may not be very successful.

Using a Community to Translate an Ontology

A different approach to localize an ontology into different natural languages is to use a community to translate the ontology. This approach combines the advantages of human translations with speed. If the community is big enough, it would be possible to let them translate a complete ontology. Of course, spammers and people with other bad means as well as inconsistency must be ignored. Therefore a certain threshold could be introduced: a word must have been translated at least a certain number of times, after which the translation is approved automatically. Some works that adopt this approach are:

AGROVOC: Caracciolo [Caracciolo et al., 2007], examines some of the issues associated with the development of AGROVOC, a multilingual thesaurus designed to cover the terminology of all subjects of interest to FAO (agriculture, forestry, fisheries, food and related domains such as environment). According to Caracciolo [Caracciolo et al., 2007], translations are provided by native speakers of the target language. Translations are typically made of the English version and sent to FAO for validation and inclusion in the master version. Apparently, terms are assigned a unique number e.g., "Abalone" is assigned to the number five in English. The translation of the word, e.g., in French "ormeau" is also given the number five in the French version. As a result, the result is not alphabetically ordered in each language, but multiple names are attached to a single concept across the languages.

The number of terms in the vocabulary varies substantially by language. The differences can be the result of discrepancies in language, but also control over additions to the vocabularies in those languages. Information about the stability of such vocabularies is limited. Information about the impact of that stability on the use or expansion of these vocabularies is also limited. Furthermore, based on these different sizes of the vocabularies, either some vocabularies are under-specified or some are over-specified, or characteris-

2.3. METHODS FOR THE BUILDING OF MULTILINGUAL ONTOLOGIES

tics of languages differentiate themselves from other language vocabularies for the same set of concepts.

UK Data Archive Thesaurus: In [Balkan et al., 2002] the authors describe an approach for the building of a multilingual thesaurus for the social sciences. The thesaurus was produced by the UK Data Archive (UKDA) as part of the EU-funded LIMBER (Language Independent Metadata Browsing of European Resources) project and was derived from their in-house English monolingual thesaurus, HASSET (Humanities and Social Science Electronic Thesaurus). The multilingual thesaurus is available in four languages English, French, German and Spanish and in various formats, including RDF (Resource Description Framework).

Basically, the construction of the thesaurus proceeded in two stages: first the monolingual thesaurus was reduced, and then the translation of the reduced thesaurus was carried out. In the first step reduction of monolingual thesaurus, different policies were adopted for the task of restructuring the UKDA HASSET for use across Europe. The reduction was made with the understanding that the rationale behind ELSST was to produce a common ontology which could be extended via local extensions to cater to the cultural and institutional needs of the individual archives and also allow for inclusion, via mappings, to specialized thesauri in certain subject areas. The translation process was carried out by a team of translators at the UKDA, who met on a regular basis to discuss problems as they arose. They provided feedback to those working on the monolingual thesaurus, so that changes to the monolingual thesaurus, such as the addition of scope notes, could be implemented where necessary. Verification of the translations was carried out by bilingual information experts at the appropriate CESSDA sites.

Automatically Translation of an Ontology

This approach uses different resources and automatic tools to ensure that information represented in an ontology using one particular natural language achieves the same level of knowledge expressivity if translated into another natural language. The main advantage of this approach is that it does not need human labor to discover the translations of an ontology. However, in the process of achieving this goal, some important challenges have to be addressed. The details on how these challenges are going to be reached and their implementation and evaluation will follow in the next chapters of this thesis.

In the following paragraphs we briefly describe the main features of the most relevant works, ordered by similarity with our approach. We wish to point out that to the best of our knowledge works that support the localization of an ontology do not exist. The different approaches that can

be found in the literature to discover automatical translations of ontology terms offer only a partial solution to the problem:

LabelTranslation: LabelTranslation is a strategy and a platform created for supporting the multilingual extension of ontologies existing in just one natural language. This tool was developed in order to support “the supervised translation of ontology labels” [Declerck et al., 2006] and, at the same time, to allow for the semantic annotation of multilingual web documents using the resulting multilingual labels of ontologies. By “supervised translation” is meant that this approach foresees the intervention of the domain expert or translator in case of a lack of results or for validation. Therefore, LabelTranslation offers a semi-automatic strategy. LabelTranslation can be integrated into any ontology engineering platform to enable its users to translate their ontologies inside the application. For the development of LabelTranslation already available multilingual semantic resources and basic natural language processing tools were reused for providing a semi-automatic translation of labels in ontologies. In the current version of the LabelTranslation platform three types of multilingual resources are included: i) EuroWordNet (EWN), a semantic lexical resource, ii) Wikipedia¹³, the multilingual free encyclopedia on the Web, based on knowledge of the word, and iii) BabelFish¹⁴, an on-line translation service used as “fallback position” [Declerck et al., 2006].

The steps for the translation approach are summarized:

1. Upload of an ontology in the LabelTranslator platform
2. Selection of the ontology labels to be translated in one of the target languages (en, es, de)
3. The system accesses the EWN database to find the selected term (or part of a term), and also checks in the WordNet database, only if the source language is English
4. Result(s) (synset and gloss) are displayed, if the matching is successful. Users can then validate the suggestions, modify the translation and save it in the database. A disambiguation problem can as well occur (see Disambiguation problem below)
5. If the matching in EWN is not successful, the system checks in Wikipedia, which also uses a mechanism for relating entries in the various available languages
6. If steps three and five do not provide any results, the system turns to BabelFish

¹³<http://es.wikipedia.org/wiki/Wikipedia>

¹⁴<http://babelfish.altavista.com/>

2.3. METHODS FOR THE BUILDING OF MULTILINGUAL ONTOLOGIES

7. If the translation is still not satisfactory, the user can enter a translation, together with partof- speech information and a definition

If the same translation session is repeated in the future, the system will return the translation already saved in its memory. Developers of Label-Translation give priority to the EWN resource because a “high quality in the translation is expected since EWN has been built following semantic considerations and validated by language and/or domain experts” [Declerck et al., 2006]. In the translation step using EWN (step three), sometimes more than just one result (or synset) is returned, which could be the appropriate equivalent translation for the label in the ontology. Then, glosses offered by EWN can be of great help, since the system can use them for disambiguating. Two approaches -or a combination of both- can be used, and these are the following (Note that LabelTranslator developers suggest the implementation of a hybrid approach combining both strategies):

- *Rule-based strategy*: the terms in the gloss of the target language are also present in the ontology; source and target languages share the same or similar glosses.
- *Static strategy*: based on two gloss-based similarity measure algorithms used in the Perl package WordNet::Similarity.

In order to solve the disambiguation problem in Wikipedia (step 5.), the user can go to the Wikipedia encyclopedic articles and manually check that the content, context, etc. of a term match with the ontology content.

Ontoling: OntoLing [Pazienza and Stellato, 2006] is a framework for a semi-automatic linguistic enrichment of ontologies. This framework was developed for “supporting manual annotation of ontological data with information from different, heterogeneous linguistic resources” [Pazienza and Stellato, 2006]. The latest version of OntoLing even helps the user with automatic suggestions through the exploitation of different linguistic resources. By exploiting existing bilingual resources, OntoLing helps in the development of multilingual ontologies, “in which different multilingual expressions coexist and share the same ontological knowledge” [Pazienza and Stellato, 2006]. In this sense, if ontologies are already available in one natural language, this tool helps in the process of ontology localization or, as has been defined by its developers, in the “multilingual enrichment process” [Pazienza and Stellato, 2006].

In the current version of OntoLing, two language resources are available for the linguistic or multilingual enrichment, WordNet¹⁵, for the linguistic enrichment of ontologies with English labels, and DICT dictionaries¹⁶, for

¹⁵<http://wordnet.princeton.edu/perl/webwn>

¹⁶<http://www.dict.org/links.html>

the linguistic and multilingual enrichment of ontologies. This last resource accesses a compendium of multiple on-line monolingual and bilingual dictionaries, as for example, all bilingual Freedict Dictionaries: English-German, English-Arabic, English-Croatian, English-Hungarian, etc.

Since OntoLing has been developed as a plug-in for Protégé, the user has to upload an ontology in the Protégé ontology editor in order to use it. Any Protégé plug-in, exploiting linguistic resources, includes a linguistic watermark package, i.e., a package that contains abstract classes and interfaces for accessing linguistic resources. As already mentioned, the current package contains two implemented linguistic interfaces related to freely available resources, namely: WordNet and DICT dictionaries. Steps and techniques of this localizing tool are summarized in the following:

- Open an ontology in the Ontology Panel of the Protégé editor
- Select from the OntoLing menu of available linguistic resources those that will be visualized during the translation task
- OntoLing accesses the selected linguistic resources by means of a wrapper called Linguistic Interface. With this Linguistic Interface the user visualizes the linguistic information in the Linguistic Browser Panel embedded in the Protégé framework.
- The ontology can be enriched with:
 - Additional labels for the selected class, i.e., synonyms
 - Glosses as descriptions for the selected class
 - IDs of the selected senses as additional labels for the selected class. This is useful if pointers from ontology concepts to senses from a given linguistic resource are needed.
- The user checks the suggestions offered by the linguistic enrichment module and selects the appropriate ones.
- Selections are added to the ontology.

Regarding the automatic linguistic enrichment of ontologies, this is currently under development. Moreover, this functionality only will be available if the ontology is in OWL (Web Ontology Language)¹⁷, and the loaded linguistic resource is a taxonomical lexical resource and/or a linguistic resource with glosses. The enrichment component will exploit the taxonomical structure of the glosses of the linguistic resource to judge which linguistic information can be used to enrich the ontology.

¹⁷<http://www.w3.org/TR/owl-features/>

Semi-automatical Translation of an Ontology

A combination of the above approaches results in a semi-automatical translation. In this approach, translations are made automatically after the result is verified by a person or community. The main advantage of this approach is that the system does the time intensive lookup of words, and the human intervention could be limited to verifying the uncertain results. To distinguish the certain from the uncertain results, multiple methods can be used. One suggestion is to use multiple translation sources. For every word, these sources are referenced for their translation(s). If the sources agree on one translation, this one is most likely correct and does not need a validation by a human. Another idea is to consider a translation successful if one source only returns one translation, thus excluding the fact of an ambiguous word. Of course, in this case, the source should be totally complete and all-knowing.

2.4 Summary of the Chapter

In this chapter we have presented the technological context related to the thesis. We have first presented an introduction to ontologies, describing what an ontology is from the perspective of Computer Science. Also we have explained some issues about the development of ontologies, focusing on the methodologies and tools that can be used to build a multilingual ontology.

Later, we have introduced the key role of MT and other related technologies as tools that allow for an automatic localization process. With regard to MT technologies, we have explained how these systems can be distinguished according to different criteria.

A classification of the different methods used for the building of multilingual ontologies has been defined. Finally, we have described the main features of the most relevant works with respect to our goal - to localize an ontology to different natural languages.

All works surveyed in this chapter show that, although the great majority of methods and techniques used to build multilingual thesauri have been adapted to ontologies, the current approaches do not reduce the costs and efforts from enriching an ontology with multilingual information. Therefore, the present work introduces a novel approach for the multilingualism problem.

Chapter 3

Work Objectives

This chapter presents the goals of our work, together with the open research problems that we aim to solve. Besides, we detail the contributions to the state of the art, the work hypothesis, the assumptions considered as a starting point for this work and the restrictions of the presented results.

3.1 Goals and Open Research Problems

The goal of our work is to establish ontology localization as a new research problem, which has not been explored yet and which in our opinion needs to be investigated. To accomplish this overall goal, we have decomposed it into the following conceptual and technological objectives:

- Goal 1. The definition of a methodology that supports the ontology localization activity. We propose a method that guides users through the adaptation of an ontology to different languages and cultures.
- Goal 2. The development of an infrastructure that implements the methods, techniques, and tools for the management of ontology localization in distributed and collaborative environments.

In order to achieve the first objective, the following (non-exhaustive) list of open problems must be solved:

- There is no consensus about the characterization and classification of ontology localization problems. Some characterizations related to localization have been proposed under different perspectives: machine translation [Hutchins, 2007], natural language processing [Briscoe, 1991], multilingual thesauri building [ISO, 1985], etc.
- The lack of prescriptive and detailed methodological guidelines for the ontology localization activity. Although methodologies exist in the

Software Localization Industry [Collins, 2001, Müllner, 2009, Jevsikova, 2009], these are general methodologies that do not define each activity or task in an exact manner; they do not clearly state its purpose, its inputs and outputs, the actors involved, when its execution is more convenient, and the set of methods, techniques and tools to be used to execute them. Therefore, it is difficult to use either of them in the activity of localization of ontologies, if we want to facilitate a prompt assimilation of ontology practitioners.

- The lack of studies about the identification and classification of translation techniques that may help to reduce the effort of localizing an ontology manually. Different approaches address the translation problem, but from different perspectives. For example thesauri localization [Balkan et al., 2002, Caracciolo et al., 2007], query translation in Cross-Language Information Retrieval (CLIR) [Nie, 2010], or meta-data records translation for MultiLingual Information Access (MLIA) [Chen et al., 2012], however, none of these approaches considers the most salient properties of the ontology localization activity as the way of disambiguating the candidate translations or the type of resources used.

With regard to the second objective, the following (non exhaustive) list of open research problems must be solved:

- From a methodological perspective, no study exists relating to the desirable requirements for an ontology localization infrastructure. Some key factors such as: i) collaboration and distribution of the task; ii) automatic translation; and iii) extensibility, present in the definition of infrastructure requirements of related works, could be used to convert them into positive influences for ontology localization.
- From a technology perspective, there exist two open problems:
 - The majority of the existing models to store the multilingual information associated to each ontology term uses a non-modular approach. In this approach the multilingual information is embedded in the ontology by means of the RDF(S)/OWL predicates¹. This way of representation has important limitations related to the restricted amount of linguistic information that can be attached to ontology terms.
 - There is no definition of a system architecture that performs for an automatic, distributed and collaborative ontology localization. None of the related works introduced in section 2.3.3, provide

¹An example of these predicate labels are: `<rdfs:label>` and `<rdfs:comment>`

a description of the system components, its properties, and the relationships between these ones, which supply a base from which localization systems can be developed.

For tackling these issues we make several contributions related to the two main objectives.

3.2 Contributions to the State of the Art

In this thesis, we aim to giving solutions to the previous open research problems. Chapter 4 and chapter 7 will describe the contributions for the first objective and chapter 5 and chapter 6 will describe the contributions related to the second one.

With regard to the first objective (the creation of a methodology to support the ontology localization activity), the thesis presents contributions to the state of the art in the following aspects:

1. *A characterization and definition of the ontology localization problems* based on the problems found in related areas. The characterization proposed takes into account three different aspects of localization problems: translation, information management, and multilinguality representation.
2. *A prescriptive methodology for supporting ontology localization activity.* This methodology is based on existing localization methodologies from Software Engineering and Knowledge Engineering, as general as possible so the methodology can cover a broad range of scenarios. This methodology describes the localization activity with its sequential task, actors, inputs and outputs.
3. *A classification of the ontology localization techniques*, which can be used for comparing (analytically) different ontology localization systems. The classifications of localization methods also provide some guidelines which help in identifying families of localization approaches.

The second objective deals with the identification and implementation of the methods, techniques and tools for the management of ontology localization in distributed and collaborative environments. This work proposes contributions to the state of the art in the following aspects:

- *An integrated method for building ontology localization systems in a distributed and collaborative environment*, which takes into account the more appropriate methods and techniques depending on: i) domain of the ontology to be localized, or ii) the amount of linguistic information required in the final ontology.

- *A modular component to support the storage of the multilingual information associated to each ontology term.* This approach follows the current trend in the integration of multilinguality in ontologies which suggests the suitability of keeping ontology knowledge and linguistic (multilingual) knowledge separated and independent.
- *A customizable model based on collaborative workflows* for the representation of the process usually followed by different organizations to coordinate the process of ontology localization to different natural languages. We propose a customizable workflow which allows to define the different actors used to support the ontology localization activity.
- *An integrated infrastructure* implemented within the NeOn Toolkit by means of a set of plug-ins and extensions (i.e. Localization Support Feature and Workflow Support) that supports the collaborative ontology localization process. Additionally, appropriate user interfaces have been implemented as part of the Workflow Support Feature, to support the possible actions/operations that users can perform according to the collaborative process. To our knowledge, other approaches that coordinates the localization of ontologies in a distributed way do not exist.

All these contributions are backed up by a large number of experiments. These experiments show how the proposed methodological and technological solutions have been applied to real-world cases, tackled in the context of the following R&D projects: the EU funded NeOn Project², the National Project “GeoBuddies”³, and the EU Monnet Project⁴.

3.3 Work Assumptions

The work described in this thesis is based on the set of assumption listed below. These assumptions help explain the decisions taken for the development of the methodological and technological solutions and the relevance of the contributions presented. Assumptions A1-A3 are related to our first objective, whilst the assumption A4 and A5 are related to the second one.

A1 According to the Neon Methodology, the ontology localization activity has to be performed once the conceptual model of the ontology is stable.

A2 We are working with ontologies of different domains.

²<http://www.neon-project.org>

³<http://mayor2.dia.fi.upm.es/oeg/index.php/es/completedprojects/97-geobuddies>

⁴<http://www.monnet-project.eu/Monnet/Monnet/English?init=true>

- A3 We assume correctly spelled ontology labels, considering the syntactic rules of the source language.
- A4 The localization of ontologies can be performed by one ontology engineer or by a team of ontology engineers, translators, and linguists who may be geographically distributed.
- A5 The collaborative ontology localization within an organization usually follows a well defined process for the coordination of the translation activities.

3.4 Hypotheses

Once the assumptions have been identified and presented, the set of hypothesis of our work are described. This set of hypothesis covers the main features of the proposed solutions and they will be validated through this thesis:

- H1 The characteristics of the ontology labels such as i) the similar lexical formats used to name terms (e.g., concepts as nouns) [McCrae et al., 2011a], ii) the low percentage of spelling errors [Espinoza et al., 2008a], iii) the significantly smaller size than a sentence [McCrae et al., 2011a], make these labels amenable to automatic translation.
- H2 The use of specific translation methods for localizing simple and compound ontology labels instead of a unified method, could improve the values of precision and recall.
- H3 The use of more than one resource into the translation process gives a wider range of translation candidates to choose from, and the correct translation is more likely to appear in multiple translation resources than in a single translation resource.
- H4 An appropriate combination of translation methods leads to better localization results than only using one at a time.
- H5 Localization methodologies in other areas are general enough to be taken as a starting point to develop an easy to use and understand ontology localization methodology.
- H6 It is possible to define a unified method to independently localize an ontology to different natural languages of i) the domain of the ontology, and ii) the process used to discover the translations of each ontology element.

- H7 The collaborative process usually followed by organizations for the localization of ontological resources can be modeled by means of collaborative workflows.
- H8 The implemented infrastructure is usable with regard to the efficiency, effectiveness and users satisfaction.

3.5 Restrictions

Finally, the following set of restrictions defines the limits of our contributions and allows the determination of future research objectives. Most of these restrictions are related to the technological aspects of the contribution (R1-R3), while R4-R7 are related to the experimentation.

- R1 The proposed method and technology do not consider the optimization of the localization process of the generated system, neither in terms of the space required during the localization nor in terms of the time needed to complete the localization.
- R2 An ontology localization systems does not necessarily have to find a translation for each ontology label. The localization of all the labels of an ontology is normally a desirable feature, however in some cases the localization depends on the degree of the shareability of the conceptualizations.
- R3 The process of localization only considers translations of ontology labels and instances. Translation of annotations labels as `rdfs:comments` are not supported yet.
- R4 We do not include support for the argumentation of the selected translations.
- R5 We are only considering ontologies expressed in OWL as input of the ontology localization activity.
- R6 The LabelTranslator system proposed for the translation of ontology labels works only with the natural languages: English, Spanish and German.
- R7 The method for localizing ontologies covers the translation to one target language per time, but does not consider the translation of labels into different natural languages simultaneously.

Chapter 4

Ontology Localization Problem

Open and dynamic systems, such as the Web and its extension, the Semantic Web, are by nature distributed and heterogeneous. Such characteristics implicate that the ontologies used to describe content and services can be represented using different formats and, more specifically, different natural languages. In this scenario, multilingual ontologies are required. As we described in section 2.3 there are two current trends for the building of multilingual ontologies, however these approaches do not reduce the cost and effort that comes with enriching an ontology with multilingual information.

In this chapter we first explain the terminology related to ontology localization activity, providing the meaning that will be used for the distinct terms. Secondly, we describe the problems that characterize the ontology localization activity. Thirdly, we briefly analyze the different scales of localization used in ontologies depending on the type of ontology elements to be localized and the level of adaptation required to make the ontology accessible to speakers of different natural languages. Also, we analyze which elements or parts of the ontology are to undergo localization. Fourthly, we provide the foundations for the thesis, giving a formal account of our general localization process by defining the input, output, and the four main steps identified.

4.1 Definition of Terms

In her doctoral work about Multilingualism in Ontologies, Montiel-Ponsoda [Montiel-Ponsoda, 2011a] identifies different definitions about ontology localization (see [Suárez-Figueroa and Gómez-Pérez, 2008, Cimiano et al., 2010], in the sense of “*the adaptation of the ontology and its natural language documentation to the needs of the target users*”). She also shows that both Software Localization and Ontology Localization have a very pragmatical

and economical orientation, since the idea is to reuse software products or ontologies already available instead of developing them from scratch. Based on this premise, she arrived at the conclusion that in “*Ontological Engineering, the localization of ontologies could be considered as a subtype of software localization in which the product is a shared model of a particular domain, i.e., an ontology, to be used by a certain application*”.

Despite the quantity and quality of definitions identified in the research works introduced above, most of the authors do not provide a guide on what really involves the ontology localization activity from a technical point of view. The definition that we propose in this thesis is intended to help understand the process for localizing automatically an ontology. We believe, that ontology localization cannot be fully or correctly understood without being contextualized in reference to a number of interdependent processes. From an Ontology Engineering perspective, these processes can be referred to as a group with the acronym ILT - Internationalization, Localization and Translation. In the following section we define in detail these three processes from the perspective of ontological engineering.

4.1.1 Ontology Localization Definition

In this section we introduce the definition of the ontology localization activity exactly as we perceived it in this work. It does not pretend to solve each particular problem nor to strictly cover the complete field. It aims at serving as a guide for this thesis. Thus, rather than attempting to cover the entire spectrum of research in ontology localization, we will concentrate on process, methods and techniques to automatically discover the translations of the elements of an ontology. The automatic process of translating monolingual ontologies into other natural languages is the core of the localization activity and will be explained in detail in the remainder of this chapter. A classification of localization approaches focusing on the translation of the different elements of an ontology will be explained in Chapter 5. Additionally, in Chapter 6 we will give an intuitive view of the whole life cycle of the localization activity.

Given one ontology O , ontology localization means that for each entity (concept, attribute, relation, or instance) expressed in a source natural language, we try to find a translation term, which has the same intended meaning, but in a different target natural language(s) L . There are some other parameters that can extend the definition of the localization activity, namely: (i) the localization parameters to accept a translation as suitable, e.g., weights, thresholds; and (ii) external linguistic and semantic resources used by the localization process to obtain the translations, e.g., text corpus, ready-to-use MT systems, or machine readable dictionaries.

This can be schematically represented as illustrated in Figure 4.1. The definition is inspired in the work presented by [Euzenat and Shvaiko, 2007]

4.1. DEFINITION OF TERMS

for ontology matching. The formalization of this process will be introduced in section 4.4.2.

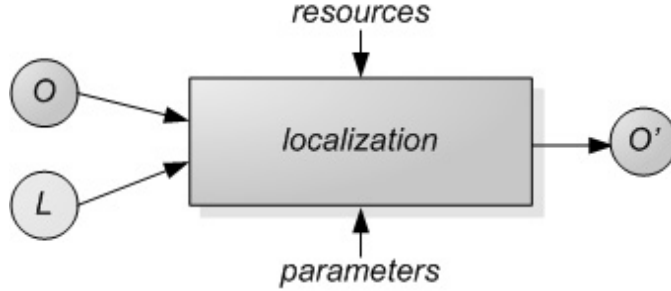


Figure 4.1: The Ontology Localization Activity

For clarification we provide a short definition of ontology as used in our scenario. So far we have considered ontologies without being precise about their meaning. An ontology can be viewed as a set of assertions that are meant to model some particular domain. Usually, the ontology defines a vocabulary used by a particular application.

Definition 1 (Core Ontology) *A core ontology is a structure*

$$B := (C, \preceq_C, R, \sigma, \preceq_R, I),$$

consisting of,

- *three disjoint sets C , R and I whose elements are called concept identifiers, relation identifiers and instances identifiers (or concepts, relations and instances, for short).*
- *a partial order \preceq_C on C called concept hierarchy or taxonomy.*
- *a function $\sigma: R \rightarrow C \times C$ called signature, where $\sigma(r) = \langle \text{dom}(r), \text{ran}(r) \rangle$, where $\text{dom}(r)$ and $\text{ran}(r)$ are the domain and range of a relation $r \in R$.*
- *a partial order \preceq_R on R , called relation hierarchy .*

We will denote by OE , the union set, $OE = C \cup R \cup I$. An element $oe \in OE$ is called an ontology element.

Relationships between concepts and/or relations as well as constraints can be expressed within a logical language such as first-order logic or Horn-logic. A formal definition of logical language has not been included at this stage of the research, because it is not relevant to the rest of the definitions.

Definition 2 (Lexicon) *Let $L = \{l_1, \dots, l_n\}$ be a set of natural languages, and Nat_{l_i} , $1 \leq i \leq n$, be a set of strings in the language l_i . A lexicon Lex for a core ontology B in a set of natural languages L is a set of functions $Lex = \{Lex_{l_1}, \dots, Lex_{l_n}\}$*

$$Lex_{l_i}: OE \rightarrow Nat_{l_i}$$

where OE are the ontology elements of the core ontology B .

In this work, an ontology consists of a core ontology, as well as a corresponding lexicon.

Definition 3 (*Ontology*) An Ontology O is therefore defined by the following tuple:

$$O := (B, Lex),$$

consisting of,

- the core ontology B .
- the lexicon Lex .

We will say that $O = (B, Lex)$ is a multilingual ontology if the set L of natural languages associated to the lexicon Lex has more than one natural language.

Once we have defined the concept of Ontology as used in our scenario, our aim is to define other processes related with the ontology localization activity.

4.1.2 Related Terms

As we explained before ontology localization cannot be fully understood without being contextualized in reference to two interdependent processes: *internationalization* and *translation*.

Internationalization.

When an ontology is developed, its design is inevitably influenced by the culture and native language of their developers. To adapt an ontology successfully to international regions or markets, the culturally and linguistically-dependent parts of the ontology must be carefully designed, a process referred to as *ontology internationalization*. This process includes, for example naming conventions of ontology terms and/or hyphenation and morphological rules of the ontology elements.

Thus, ontology internationalization can be defined as the process of generalizing an ontology so that it can handle multiple languages and cultural conventions without the need of re-designing it. Internationalization takes place at the design level. There are two key reasons for ontology internationalization:

4.1. DEFINITION OF TERMS

1. To ensure that an ontology is properly designed and therefore can be accepted in international markets, and
2. To ensure that an ontology is localizable.

In the first case, the labels and descriptions used in the ontology are concise, clear, and they do not contain any jargon or slang. The second reason above mentioned will help to reduce the localization costs by developing the ontology in a way that ensures a smooth localization process. One way to do this is by following a standard for the naming of labels. Some works [Fried et al., 2007, Schober et al., 2007] have proposed naming conventions for ontology terms. These guides are used in specific applications such as ontology verbalization¹. We claim that the definition and use of style guidelines should also be extended to ontology engineering.

Translation.

Translation can be generally defined as the process of “transferring a text from a source language and culture into a target language and culture with a certain purpose” (adapted from [Nord, 1997]). Considering this definition, we agree with [Montiel-Ponsoda, 2011a] in that the translation process may be considered the mother activity that encompasses Ontology Localization.

Depending on localization purpose, the process of translation can be categorized in: *instrumental* and *documental* [Montiel-Ponsoda, 2011a]. In the first case, the goal of the target ontology can be to have the same function in the target community as the original ontology in the source ontology. The purpose of the translation can also be to “document” the ontology in another language to make it accessible to a community which speaks another language. In both cases and just as it occurs in software localization, in ontology localization the emphasis should be placed on automatic translation tools that allow users to avoid the manual effort of building a multilingual ontology.

Some authors believe that this solution is not viable, since machine translation (MT) today suffers from several critical limitations to support the translation of ontology labels [Segev and Gal, 2008]. The general awareness is that automatic translation tools have yet to achieve a level of proficiency comparable to human translation. However, since ontologies consist of concepts, attributes and relations that are stated clearly and succinctly, we hypothesize that ontology components are more readily translatable than full-length text [Espinoza et al., 2008a].

When translating ordinary text, one has to deal with textual phenomena such as anaphora or metaphors, and much care must go into assuring that

¹The verbalization makes ontologies accessible to people with no training in formal methods. The goal of the ontology verbalization is to produce natural language from the definition of class or properties.

one obtains clear and natural-sounding sentences. This is not such a big issue in ontology labels, which tend to have text with single words, compound words, named entities, short phrases, or short sentence fragments. Also, the ontology labels have characteristics, which make these amenable to MT:

- *Consistency.* The lexical formats used for naming ontology terms are very similar [Espinoza et al., 2008a]. Also, the labels used for describing ontology elements commonly use a upper/lower case distinction. It poses some advantages to MT because it allows performing word segmentation. Some works have shown that having a basic word segmenter helps MT performance [Koehn and Knight, 2003, Habash and Sadat, 2006, Chang et al., 2008]. Additionally, we can rely on the initial uppercase letter to identify a phrase initial word.
- *Accuracy.* The spelling accuracy of the labels of an ontology is reported to be approximately 97.0%-99.5% [Espinoza et al., 2008a]. These values are very important because the typographical errors can affect the translation quality. Furthermore, sentence boundaries (used in ontology term comments), which are absolutely crucial for parsing in MT, are usually clear in the ontologies through the use of accurate methods of punctuation.

We define the ontology label translation task as finding, for an individual label l in the source language S , the correct translation, either a word or phrase, in the target language T . Clearly, there are cases where l is part of a multi-word term that needs to be translated as a unit. For this case, this approach can be extended by preprocessing the data in S to find short-phrases, and then executing the entire algorithm treating short-phrases as atomic units. In this thesis, we do not explore the extension of this approach to the translation of sentences (e.g., comments of ontology term). Nevertheless, we focus on the translation of simple and compound labels. The technical details of this process will be explained in section 4.4.2.

4.2 Characterization of Ontology Localization

Once the main concepts of the ontology localization activity are defined, we describe in this section the localization problems that must be taken into account when deciding on the method of localizing an ontology.

The characterization of the localization problem in ontologies has been analyzed previously in [Montiel-Ponsoda, 2011a]. In this approach, the author provides a classification of different categorization relations (language equivalence problems in our work) that are shared among different cultures, no matter how different the linguistic structures are that express them in

each language. However, she does not envision problems as: the identification of translation mechanisms that preserve the semantics of the original ontology term, the management of changes in ontology terms and their translated labels, and the representation of the multilingual information.

4.2.1 Language Equivalence Problems

Due to the fact that cultures classify the world in a different way, when translating ontologies we may encounter different types of situations:

- *Existence of an exact equivalence.* This is typical of highly specialized technical and engineering fields such as Mechanics, in which there is a direct/complete equivalence among the terms in different languages referring to a certain object or process. In this case there is little place for synonyms or variants. E.g., ‘Wärmekraftmotor’ in German is translated as ‘heat engine’ in English.
- *Existence of several context-dependant equivalents.* When one term in a language can be translated by several equivalents in another language, and the user has to choose the most suitable depending on the context of the ontology and the word connotations. For example, the English term ‘girl’ can be translated into Spanish as ‘niña’, ‘chica’, ‘joven’, o ‘hija’. Each translation reflects different nuances of the concept and it will be necessary to find out which equivalent is needed in the ontology to translate the English term ‘girl’ with the most approximate or suitable equivalent.
- *Existence of a lexical gap.* This is mainly due to mismatches at the conceptualization level, i.e., when a certain culture categorizes reality with a degree of granularity that does not correspond to the granularity degree of the other culture, resulting in a lexical gap in the target language. For example, in French there is a difference between big rivers that flow into the sea, which are called ‘fleuves’, and rivers that flow into other rivers, ‘rivières’. In most topography ontologies in English this distinction is not made.

4.2.2 Translation Problems

In order for a translation algorithm to be useful in ontology localization, it has to produce reliable translations that exactly correspond to what a human translator would produce. Having a computer assistant that requires the translator to do significant corrections to its suggestions is nearly as good as having no assistant at all. For our purposes, the aim of this process is to suggest terms that are translations of the original concepts in the ontology. This particular task may be performed in at least two ways:

- The first option involves term translation from source language into target languages(s) followed by a *cross-lingual retrieval* of the senses of each translated term in the source language. Note that in this case the senses of the translations is in the same language of the term to be localized. To compute the similarity between the senses of the translations and the sense of the term under consideration a similarity measure is necessary. We identified these measures as *language dependent* because the compared terms need to be defined using the same natural language.
- The second option involves term translation from source language into target languages(s) followed by a *monolingual retrieval* of the senses of each translated term in the target language. In this case, to be able to compare the terms using their contexts, we need to use a similarity measure that will allow cross-lingual comparison of words and their contexts. We identified these measures as *language independent*.

In both cases, this process requires that the full meaning of each term be accurately rendered from a source language into target natural language, with special attention paid to cultural nuances. In the MT literature, some authors have investigated the improvement in the quality of MT approaches, incorporating context-rich approaches from word sense disambiguation (WSD) methods [Carpuat and Wu, 2007, Apidianaki, 2009]. However, to the best of our knowledge, the inclusion of a disambiguation method of ontological terms for improving the ontology localization activity not been tested yet.

The formulation of the translation task as a word-sense disambiguation task has multiple advantages. First, if we knew the correct semantic meaning of each word in the source language, we could more accurately determine the appropriate words in the target language. Secondly, the availability of large amounts of resources from which we can infer the senses for disambiguating the words.

4.2.3 Management Problems

Whereas the translation process of ontology labels *per se* implies certain difficulties, the maintenance and updating of translated ontology labels throughout the ontology life cycle also requires special attention. The main difficulty is to identify policies for managing changes in ontology terms and their translated labels. Up to now, none of the works on managing ontology changes [Palma et al., 2008, Tudorache et al., 2008] dealt with changes of ontology elements with multilingual information. Several situations could happen:

4.2. CHARACTERIZATION OF ONTOLOGY LOCALIZATION

- *An ontology term is added* then the ontology label should be translated to all supported languages.
- *An ontology term disappears* then all their translations should be removed.
- *An ontology term is renamed* then all multilingual labels should be re-translated.

In all cases this process can be performed using two operation modes: instant mode or batch mode. The first is executed when changes are applied, while the batch mode can, for instance, be executed at the end of the user's session.

4.2.4 Multilinguality Representation Problems

As a result of the Localization Activity, we obtain an ontology in which ontology labels are represented in different natural languages. According to the state-of-the-art, we can identify three main models of representing multilinguality in an ontology:

- *Inclusion of multilingual information in the ontology by means of human readable labels.*² This has been the most used approach by the Ontology Engineering Community until now, which allows multilingual labels to be associated to ontology terms.
- *Creation of one conceptualization per culture and language involved, and mappings established between the different conceptualizations.* Each conceptualization will reliably reflect the categorization of the reality that each language makes. However, the effort required in the development of the various conceptualizations and the linkage among conceptualizations is by no means trivial. A representative example of this approach is the well-known EuroWordNet³ lexicon.
- *Association of external multilingual information to the ontology.* Different models have been proposed to associate linguistic data to ontologies: a) the Linguistic Information Repository (LIR) [Montiel-Ponsoda et al., 2011], specially designed to account for cultural and linguistic differences among languages; b) LexInfo [Buitelaar et al., 2009] combines the linguistic elements represented in LingInfo [Buitelaar et al., 2006] and LexOnto [Cimiano et al., 2007], and c) the Lexicon Model for Ontologies (Lemon) whose focus is on the linguistic enrichment of ontologies from a morphosyntactic viewpoint. The main advantage of

²We refer to *rdfs:label* and *rdfs:comment* properties used to describe a resource with human readable text in addition to “pure” RDF properties.

³www.illc.uva.nl/EuroWordNet/

this third approach is that it does not require the effort of creating additional conceptualizations, and it can take advantage of the ontologies already available on the Web to create multilingual ontologies.

The choice among the three models will be mainly determined by the shareability of the conceptualization and the amount of linguistic information required for the final ontology [Espinoza et al., 2009b].

In the next section, we describe the different scales of localization used to enrich an ontology to different natural languages.

4.3 Scales of the Ontology Localization Activity

In our approach, the “scales of localization”, or degrees of localization refer to the relationship among the amount of translation required and the customization necessary to create different ontology language editions. These scales are inspired on product localization scales proposed in the Software Engineering area (see section 1.3).

An ontology is a fairly complex structure and it is often more practical to focus on the localization of different levels of the ontology separately rather than trying to directly localize the ontology as a whole. This is particularly true if we want a predominantly automated localization rather than entirely carried out by human users/experts. Another reason for the scale-based approach is that the degree of complexity of the translation techniques involved in the localization depends on the type of ontology elements that can be localized and the level of adaptation required.

The selection of the elements or parts of the ontology that can be undergo to the localization requires a detailed analysis of the layers into which an ontology can be divided. According to Morris (1938) the science of semiotics⁴, namely, the study of formal languages and the theory of signs considers the existence of several levels in the definition of a language: *syntax*, *semantics*, and *pragmatics*. Syntax deals with the structure of symbols, semantics with their meanings, and pragmatics with their contexts of usage.

These three levels, have been adapted for different reasons to the problem of interoperability and translation. For example, in the context of semantic interoperability, some authors have proposed classifications of the problems to be faced when managing different ontologies in, possibly, different formats [Chalupsky, 2000, Euzenat, 2001, Klein, 2001]. Other authors have refined these levels with the purpose of building and maintaining ontology translation systems in: *lexical*, *syntax*, *semantic*, and *pragmatic* layers [Corcho, 2011] or for supporting the association of explicit semantics based on

⁴The subject of semiotics was originally spelled “semeiotics” to honour the English philosopher, John Locke (1632-1704) who in “An Essay Concerning Human Understanding” (1690) first coined the term “semiotike” from the Greek word “semeion” meaning “mark”, “token” or “sign”.

4.3. SCALES OF THE ONTOLOGY LOCALIZATION ACTIVITY

ontologies with relational data sources in: *lexical*, *syntactic*, *representation paradigm*, *terminological*, *conceptual* and *pragmatic* layers [Barrasa, 2007].

In Ontology Localization, we find that Montiel-Ponsoda [Montiel-Ponsoda, 2011b] takes the Barrasa's classification (mentioned above) for identifying the layers that may be involved in this activity. She states that the *lexical*, *syntactic* and *representation paradigm* layers not to be affected by Ontology Localization, because their design options do not depend on the inclusion of multilingual information. The *terminological layer* is the one most clearly affected by the Localization Activity, since the labels that name ontology terms will have to be expressed in different natural languages. Regarding the *conceptual* layer, she identifies this layer as a potential candidate to be modified during the Localization Activity because, certain languages and cultures categorize certain knowledge spheres in a way that may not be shared by other cultures. Finally, the *pragmatic* layer may also need to undergo localization to meet target language needs, however, she discards this layer in order to generalize as much as possible the ontology localization activity.

The research conducted in this work focuses on three ontology layers to define the scales of localization. On the one hand the *terminological* layer leads to different scales of linguistic adaption depending on the ontology elements to be translated, and on the other hand, the *conceptual* and *pragmatic* layers lead to different degrees of transformation of the conceptual model depending on the required cultural adaptation.

Figure 4.2 shows the different scales for ontology localization. These degrees range from the linguistic adaption of the ontology to a particular language (*linguistic level*) to a cultural adaption of the ontology to a specific geo-political and cultural environment (*cultural level*)

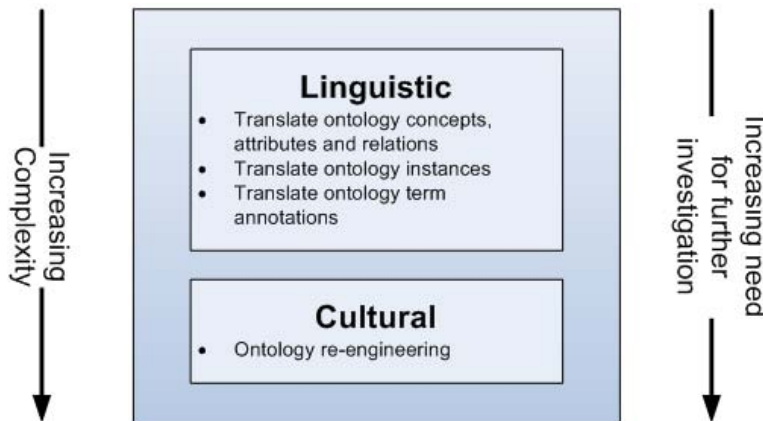


Figure 4.2: Ontology Localization levels.

Linguistic Level

This level involves the adaptation (translation) of the ontology to a particular language. This adaptation will affect the lexical layer of the ontology. Obviously, the lexical layer is language-specific and is thus clearly affected by any ontology localization process, even when the adaptation is done within the same linguistic system. This means that the changes motivated by the cultural environment in which the ontology is to be used -be it within the same linguistic system or not- will be reflected at the lexical layer [Cimiano et al., 2010]. This task of adapting the lexical layer to another language is however crucial to make the ontology accessible to speakers of another language. A straightforward way to localize the lexical layer is to provide a 1:1 translation for each label, definition and the accompanying documentation [Cimiano et al., 2010].

We consider different levels of difficulty to translate ontology elements:

- *Level 1 - Translation of ontology concepts, attributes and relations.* At this level, the localization process involves translating the labels⁵ or identifiers⁶ of concepts, attributes and relations. These labels or identifiers can be simple or compound words. The main difficulty in this level is to get the label specifications and its context correctly. Consider for example the word ‘plant’, which depending on the context can be translated into Spanish as ‘planta’ in the sense of “living organism” or ‘fábrica’ in the sense of “industrial plant”.
- *Level 2 - Translation of ontology instances.* The main complication at this level is to decide which instances should be translated and which ones no. A big part of the instances are represented by a proper name, and therefore should not be translated (e.g., a label containing “Michael Schumacher” should not be translated). However, other instances as by example “South America” should be translated to other natural languages.
- *Level 3 - Translation of ontology term annotations.* The main difficulty at this level includes translating a huge amount of phrases which are part of the annotation of concepts, attributes or instances in an ontology. To provide a human-readable description of a term, the Ontology Web Language (OWL) uses for example the `rdfs:description` statement, where a textual comment can be added. Thus, this level involves the difficulty to translate long pieces of text correctly.

⁵Name of an ontology term.

⁶The identifier is the name used in the URI reference of a term. The identifiers make the ontology much more readable.

Cultural Level

The second level includes the ontology adaptation to a particular culture. From a software product perspective this level covers two areas: the context of the use and the meaning of symbols, graphics, colors and metaphors used in the user interface. However, from an ontology perspective the cultural level needs to adapt an ontology to a specific geo-political and cultural environment.

We perceive that the cultural adaptation is a special form of *ontology re-engineering* which means transforming the conceptual model of an existing and implemented ontology into a new, more correct and more complete conceptual model which is re-implemented. In this case, the adaptation to a different geo-political and cultural reality may require more than a 1:1 translation, i.e. a change as well in the underlying conceptualization.

In all cases, the localization process needs to take care of cultural discrepancies of source and target communities due to differences at the cultural or geopolitical background - will have an impact on the lexical layer, since language is the means we have to understand and experience reality. For instance, the most appropriate translation for the English term ‘computer’ in the sense of “a machine for performing calculations automatically”, is ‘ordenador’ in Spain, but ‘computadora’ in South America.

In this thesis we only describe the factors related to the automatic process for localizing ontologies when the lexical layer undergoes modifications without affecting the conceptualization. The details of how to localize an ontology to the cultural level is out of the scope of this work.

Once we analyzed the different scales of localization depending on the type of ontology elements to be localized and the level of required adaptation, in the next section we give a formal account of our general approach to localize an ontology to different natural languages.

4.4 Ontology Localization Approach

After the foundations, the next step in the research methodology is the actual creative and innovative one. This section describes our approach for ontology localization based on the previous findings. In fact, this will be achieved through several elements. First, we describe the different scenarios of localization used to enrich an ontology to different natural languages. Secondly, we shape a general underlying approach for localization. Specific methods for each task are then going to lead to a concise basic approach [Espinoza et al., 2008a, Espinoza et al., 2008b, Espinoza et al., 2009a]. Then, we explain the individual steps in detail.

4.4.1 Scenarios for Localization

At this stage, two scenarios can be identified to perform the localization activity:

- **Scenario 1: Bottom-up approach.** This approach starts from an ontology already conceptualized and the goal is to get the translations of the ontology labels in different natural languages. Usually, these ontologies are designed without taking into account the multilingual and localization aspects.
- **Scenario 2: Top-down approach.** Starting at the ontology design level, the ontology is built in such a way that later on it can be easily adapted to new languages and cultural conventions. In section 4.1, we introduced the notion of internationalized ontologies, which we see as ontologies built with the aim of supporting multilingual descriptions of the conceptualizations they provide, since they are to be used in a multilingual scenario.

Figure 4.3 shows the relationships between the scenarios above described. In general, ontology localization should be performed after the internationalization task, because internationalization takes place at the level of the ontology design. In any case, localization does not ‘per se’ come after internationalization, nor is it a subordinated task. Translation is a more independent task since one can translate without internationalizing, although translations are greatly affected by the other two concepts.

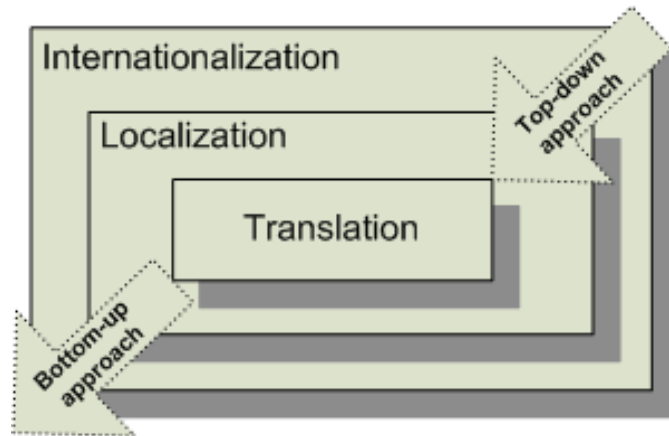


Figure 4.3: Ontology Localization scenarios.

The generic approach for localizing ontologies that we explain in the following section can be used in both scenarios.

4.4.2 Automatic Localization Approach

The localization activity is a complex task that involves different manual tasks (e.g., management, translation, revision, etc.). From these tasks we consider that the translation task definitively requires the most effort. For this reason, one of the main objectives of this work is to integrate an automatic translation process in the localization of ontologies.

Identifying the Phases of the Translation Process

We propose that the starting point in the design of a general translation process for the ontology localization activity should be guided by the observation of how the translation process is performed by a human expert. Thus, we first consider the nature of the “translation process” itself. Malmkjær [Malmkjær, 2000] points out that the translation process may be used to designate a variety of phenomena, from the cognitive processes activated during translating, both conscious and unconscious, to the more “physical” process which begins when a client contacts a translation bureau and ends when that person declares satisfaction with the product produced as the final result of the initial inquiry. In translation practice, of course, the cognitive aspects are expressed within the physical aspects.

Few studies deal specifically with the identification and characterization of the phases or stages for modeling the human translation process (see [Starren and Thelen, 1988, Nord, 2005, Englund, 2005]). For our purposes, we adopt the approach presented in [Starren and Thelen, 1988] which is organized in four steps: 1) meaning discovering, 2) finding receptor(target) language equivalents, 3) checking the meaning of the receptor(target) language item, and 4) formulation of the final translation. Figure 4.4 illustrates the steps above described.

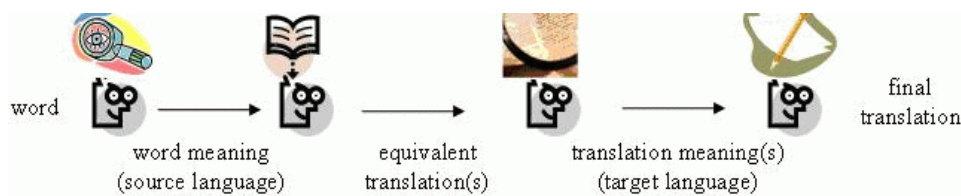


Figure 4.4: Human translator steps.

In order for a human translator to be able to discern among the different meanings a word may have (homonymys or polysemic words), (s)he needs to analyse the context. Depending on the context in which the word is used, a certain meaning will be selected, whereas the rest of potential meanings of the word will be discarded. This process is performed almost unconsciously in the translator’s mind if (s)he has a good command of the subject and the terminology used in it. For example, if the word to be translated is “bank”,

and the text is about finances, the translator will undoubtedly assign the meaning “financial institution” to the word “bank”.

The next step in the translator’s mind is to look for possible equivalents of the word “bank” with the meaning of “financial institution” in the target language. Assuming the translator’s proficiency on the subject in both, the source and target culture, the translator will look for an equivalent concept in the target culture. If “bank” is going to be translated into Spanish, the translator has to find out if the English word is referring to a “savings bank” or to an “investment bank”, for example, since in the first case, “bank” would be translated into “caja” and in the second into “banco”. Here again the context is essential for the translator to make the right choice. At this stage, it is difficult to separate this action into two steps, finding language equivalents and checking their meaning) because concepts are represented by lexicalizations, and they come together as indivisible items in the translator’s mind. In order to take the final decision on which the most appropriate translation for a certain word is, the translator will have to take into account two additional aspects: 1) which is the concept in the target language that better matches the concept in the source language?, and 2) which is the purpose of the translation? Once the purpose and context have been checked again, the translator is able to select the most appropriate translation for the source word.

In the following, we give a formal account of our general translation process by defining the input, output, and the four main steps identified.⁷

Defining the Automatic Translation Process

Figure 4.5 illustrates input, output, and the four main steps of the general automatic localization approach. This detailed stepwise approach of ontology localization is novel and one core contribution. Notice that these steps cover only the translation task in the Ontology Localization Activity. The description of the life-cycle model for localization which involves tasks that extend far beyond the translation process itself will be introduced in Chapter 6.

In the following sections, the individual steps will be explained in more detail.

Input

The input of the process is one or more ontologies, which need to be localized to different natural languages.

If more than one ontology is taken, then each ontology is processed individually. Pre-known lexical or multilingual information of the ontology

⁷It should be noted here that all phases were re-labeled to describe their functionality in the localization activity.

4.4. ONTOLOGY LOCALIZATION APPROACH

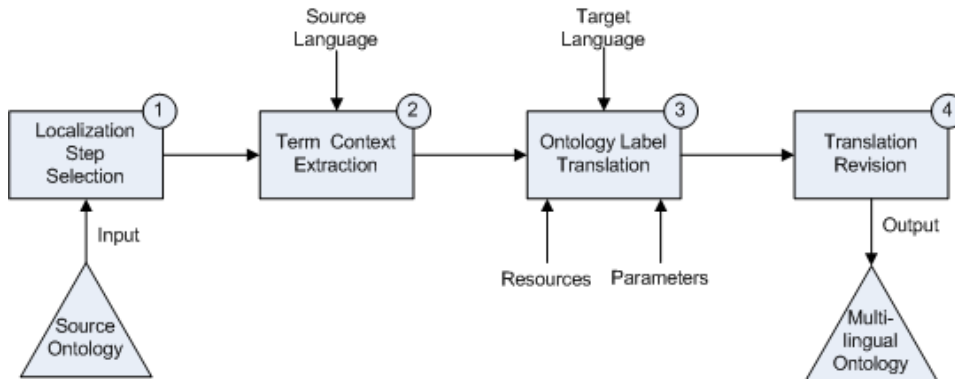


Figure 4.5: Automatic translation approach for the Ontology Localization Activity.

terms to be localized may be very useful, giving to the localization algorithm good starting points for discovering the translations of other ontology elements. For example, it may be useful to know that the word “river” is more specific (e.g., as provided by the `skos:narrower`⁸ property) than the English concept “watercourse” even if the former is not a literal translation of the latter. This lexical term may help to disambiguate the possible candidate translations. Also, if the same concept contains multilingual information indicating that a translation in French of “watercourse” is “cours d’eau” (e.g., as provided by the `rdfs:label`⁹ property), then, it is possible to use this information as intermediate language of an indirect translation¹⁰ in other language.

Localization Step Selection

Before the localization of ontology elements can be initiated, it is necessary to choose which element actually to consider from the ontology. This step may choose to discover the translations of certain candidate ontology elements and ignore others (e.g., only localize ontology concepts and not ontology relations)

Definition 4 (*Localization Selection*) *Given an ontology O , we define*

⁸The Simple Knowledge Organization System (SKOS) is a common data model for sharing and linking knowledge organization systems via the Semantic Web.

⁹This property allows to include the tag “@lang” to create multilingual labels in RDF.

¹⁰Indirect translation is translation into language C based on a translation into language B of a source text in language A [Landers, 2001].

a localization selection. Sel_O , as a subset of the ontology elements OE of O

$$Sel_O \subseteq OE$$

To the best of our knowledge, there are no specific methods for selecting the space of candidates to be localized. We consider that the implementation of a specific selection method may depend on various factors, for example: the time and recourses available for performing the localization or the use of the ontology after localization.

In [Prins and van den Broek, 2004] the authors propose a method to make a semi-automatic “intelligent” translation of only a part of the ontology. They use as strategy of selection to find out which concepts of the ontology contain the most instances and then translating the concepts one by one. The intuition of this approach is based on the fact that instances are the ontological terms which are used more in their particular cases (multilingual semantic search). To decide which branches are the most important, they use a script that visualizes the amount of concepts and instances in the ontology. This visualization allows the user to identify the parts that have an enormous amount of subclasses and none or only few instances, and then starting with the translation of the selected terms. In our work, we use a similar localization selection strategy, in which the user may choose to localize the complete ontology or only certain ontological elements.

Context Extraction

In order to translate an ontology element $oe \in OE$ from the ontology O , one must consider its context¹¹. The context of an ontology term allows discerning among the different meanings that an ontology label (defined in the lexicon Lex of the ontology) may have. Notice that, the inclusion of this phase in our generic translation approach goes in the line of incorporating a word sense disambiguation method to improve the quality of the obtained translations.

The clues for discovering the context of an ontology term can be found not only in the surrounding terms, but also in other terms semantically related to the terms under consideration. In other circumstances the clues to extract the context of a term are found in the textual descriptions and also in the practical interpretation of the term.

In the rest of this thesis, we will not further distinguish between labels and words, which will be interchangeable.

Definition 5 (*Ontology Term Context*) Let U_{oe} be a set of ontology elements such that $oe \in U_{oe}$. The context of the ontology element oe , ctx_{oe}

¹¹Context is the environment in which a word is used, and context, viz. word usage, provides the only information we have for figuring out the meaning of a new or a polysemous word.

4.4. ONTOLOGY LOCALIZATION APPROACH

is the set

$$ctx_{oe} = U_{oe} \cup Lex(U_{oe})$$

where $Lex(U_{oe}) = \{Lex(u) \mid u \in U_{oe}\}$

The selected ontology elements in an ontology term context may vary according to the requirements of the used algorithm of translation. In any case, the context of the ontological term (concepts C , relations R , and instances I) needs to be extracted from intensional and extensional ontology definitions. The main goal of the context of an ontological term is to reduce the “noise” of the word translation, since a single source word can be translated into many words in any target language. For example, the concept term *chair* of the sample ontology can be translated from English to Spanish as the nouns: *silla* (seat), and *cátedra* (professorship), or as the verb: *presidir* (take the chair). However, if we use as context the term *professor*, we can limit the number of obtained translations.

We consider two different dimensions for modeling the context ctx of an ontology element oe :

- *Context interpretation.* This dimension is concerned with the way to encode the context used by a particular translation technique.
- *Context size.* This dimension makes reference to the number of elements used to define the context of a term. It is difficult to establish the minimum size that the context should have for determining the meaning of a term.

The combination of these two dimensions provides a broad range of possibilities for the translation algorithms. In section 5.1.1 we categorize these dimensions and we show how these can be used to classify different translation techniques.

An orthogonal dimension that needs to be considered is the *context feature selection*. This dimension aims to select the most relevant context features, removing the features least useful and thus improving efficiency or accuracy. There are several techniques to determine which words make up the context of a word: distance-based window, syntactic based-window, relatedness computation between words, etc. [Gamallo, 2007]. Some of these techniques have been applied to a word sense disambiguation domain; see for example [Mihalcea, 2002, Decadt et al., 2004, Gamallo, 2007, Gracia and Mena, 2009]. In our work we use a mechanism for the context selection, which is based on the relatedness computation between words [Espinoza et al., 2008a].

Term Translation

For a given ontology element oe from ontology O , this step tries to discover the more appropriate translation. The translation computation of an ontology element oe is done by using a wide range of translation functions. Each translation function is composed of the context of the ontology term, the target language(s) in which the ontology should be expressed, and the linguistic and semantic resources to obtain the translations.

Definition 6 (*Translation similarity function*) Let oe be an ontology term, ctx_{oe} be an ontology term context for oe , and l be a natural language. A function

$$ts_{ctx_{oe}}^l: Nat_l \rightarrow [0, 1]$$

is called translation similarity function.

The translation similarity function will use a variety of linguistic and semantic resources to obtain the translations.

Definition 7 (*Localization ontology*) Let $O = (B, Lex)$ be an ontology, Sel_O be a selection of O , $\lambda \in [0, 1]$ a real number, and l be a natural language. The localization ontology O for the selection Sel_O to the natural language l with threshold λ is a ontology $O' = (B', Lex')$

$$O' = ts_O^\lambda$$

that it holds:

- $B = B'$;
- $Lex' = Lex \cup \{Lex_l\}$;
- for all $oe \in Sel_O$, $ts_{ctx_{oe}}^l(Lex_l(oe)) > \lambda$

where $ts_{ctx_{oe}}^l$ is a given translation similarity function for any $oe \in Sel_O$ and given context ctx_{oe} .

It is denoted by tl_{oe} , the translation of oe , i.e. $tl_{oe} = Lex_l(oe)$.

Different techniques can be used to perform the ontology localization task. A classification of these techniques will be extensively defined and explained in the next chapter.

Evaluation

In our approach we consider the translation task in a very specific setting of computer-assisted software localization. This setting imposes that the expected translations have a high quality. However, we recognize the need for an adequate procedure to evaluate and to guarantee the quality of the translation. Thus, from the obtained translations, we need to identify their quality.

4.4. ONTOLOGY LOCALIZATION APPROACH

Definition 8 (*translation evaluation*) For each obtained translation of an ontology element oe , evaluation is defined as

$$eval: tl_{oe} \rightarrow [0, 1]$$

where, tl_{oe} is the result of the translation of a ontology element oe into a target natural language.

Ideally, without any time or money constraints, translation output could be judged by humans to provide an idea of the system's performance. Obviously this is not the case when we need a fast way of evaluating translations. Goutte [Goutte, 2006] reviews a few automatic MT evaluation metrics from two different approaches¹²: string matching based and information retrieval (IR) based.

All metrics presented below rely on a number of reference translations to which the translation output is compared. This does not mean that all words to be translated must have reference translations, only benchmark words. This does however mean that the performance measured automatically on that benchmark may not carry over to a different body of labels, especially in a different domain.

String Matching Techniques. These metrics are based on the computation of the minimum edit (Levenshtein) distance . This identifies the minimum number of insertions, deletions and substitution necessary to transform one string into the other. Some metrics that use this approach are:

- *Word Error Rate (WER)* is computed as the sum of insertions, substitutions and deletions, normalised by the length of the reference word. A WER of 0 means the translation is identical to the reference. One problem with WER is that this measure does not guaranteed a value between 0 and 1 and in some settings a wrong translation may yield a WER higher than 1.
- *WERg* [Blatz et al., 2004], normalises the sum of insertions, substitutions and deletions by the length of the Levenshtein alignment path, i.e. insertions, substitutions, deletions and matches. The advantage of this metric is that it is guaranteed to lie between 0 and 1, where 1 is the worst case (no matches).
- *Position-independent Error Rate* does not take into account the ordering of words in the matching operation. In fact it considers the translations and the reference as bag-of-words and computes the differences between them, normalized by the reference length.

¹²This section is a summary taken from [Goutte, 2006]

In fact, any string comparison technique may be used to derive similar translation evaluation metrics. One such example relies on the “string kernel”, and allows to take into account various levels of matching depending e.g., on the part-of-speech of the words, or to take into account synonymy relations [Cancedda and Yamada, 2005].

IR-style Techniques These metrics use measures inspired by Information Retrieval. In particular the n-gram precision is the proportion of n-grams from the translation that are also present in the reference. These may be calculated for several values of n and combined in various ways.

- *BLEU*: This metric proposed by [Papineni et al., 2002] is the geometric mean of the n-gram precisions for $4 \leq n \leq 1$, multiplied by an exponentially decaying length penalty. This penalty compensates for short, high precision translations such as “the”.
- *NIST*: This metric was used in the MT evaluation rounds organised by NIST [Doddington, 2002]. NIST computes the arithmetic mean of the n-gram precisions, also with a length penalty. Another significant difference with BLEU is that n-gram precisions are weighted by the n-gram frequencies, to put more emphasis on the less frequent (and more informative) n-grams.
- *F-measure*: The F-measure [Melamed et al., 2003] is the harmonic mean of the precision and recall. It relies on first finding a maximum matching between the translation output and the reference, which favors long consecutive (n-gram) matches. The precision and recall are then computed as the ratio of the total number of matching words in the maximum match over the length of the translation and reference, respectively.
- *Meteor*: The Meteor evaluation system improves upon the F-measure in at least two ways. It uses some linguistic processing to match stemmed words in addition to exact matches, and it puts a lot more weight on the recall in the harmonic mean [Lavie et al., 2004].

BLEU and NIST are the metrics that are currently most widely used, and the ones all other MT evaluation metrics have to be compared with. The F-measure claims to provide higher correlation with human judgements [Melamed et al., 2003], but this is apparently not always the case, especially for smaller segments [Blatz et al., 2004]. Empirical evidence [Lavie et al., 2004] suggests that putting more emphasis on recall further improves the correlation. In fact it shows that recall alone often correlates best with human judgement, at odds with the exclusive use of precision in BLEU and NIST.

Output

As we explained before, we consider a multilingual ontology as the output of the ontology localization process. A multilingual ontology express the correspondences between entities belonging to the ontology to be localized and the multilingual terms pertaining to a natural language. A correspondence must consider the two corresponding entities (ontologies entities and multilingual terms) and the relation that is supposed to be held between them.

In the following part we first provide the definition of multilingual ontology like it is used in our work. Then, we describe the other important component of the output of the localization, the relation that holds between the source entities with their translations.

Definition 9 (*multilingual ontology*) *A multilingual ontology $O' = (B', Lex')$ is an association of ontology elements OE with a set of translation terms T pertaining to a different natural languages L .*

$$O' : OE \rightarrow T_L.$$

where, each ontology element $oe \in OE$ is labeled by a set of translation terms $t_1, t_2, \dots, t_n \in T$ in the language l of the lexicon Lex_l . We denote $O'(oe) = \{t_1, t_2, \dots, t_n\}$.

The multilingual ontology defines also the reciprocal relation

$$S_L : T_L \rightarrow OE$$

by $S_L(t) = \{oe \in OE | t \in O'(oe)\}.$

The next important component of a multilingual ontology is the relation that holds between the ontology elements oe and its translations T (see *translation problems* in the section 4.2).

We consider that ontology localization algorithms should primarily use the equivalence relation ($=$) for expressing synonymy or equivalence relationship. However, according to guidelines for the establishment and development of multilingual thesauri [ISO, 1985] equivalence is divided into: exact equivalence, inexact equivalence, partial equivalence, single-to-multiple equivalence and non-equivalence. In the following part we briefly explain each case. In all examples, the letter X represents the source ontology element that needs to be localized and the letter Y its translation(s):

- *Exact equivalence (inter-language synonymy)*: the terms in X and Y are semantic and culturally equivalent. Table 4.1 shows a sample of equivalents terms in different languages.
- *Inexact or near equivalence (inter-language quasi-synonymy, with a difference in viewpoint)*: the terms in X and Y express the same general

Table 4.1: Exact equivalent sample

German	English	French	Dutch
Schiennennetz	Rail network	Résau ferroviaire	Spoorwegnet

concept but the meanings of the terms in X and Y are not exactly identical. Often the differences are more cultural than semantic, i.e. there is a difference in connotation or appreciation. In the case of inexact equivalence the terms can be treated as if they were exact equivalents. Table 4.2 shows a sample of inexact equivalence terms in different languages. The terms in Spanish and English are equivalent, however the term in French is only a near equivalence.

Table 4.2: Near equivalence sample

English	Spanish	French
Historic settlements	= Asentamientos históricos	≈ Site de peuplement

- *Partial equivalence (inter-language quasi-synonymy, with a difference in specificity)*: the term X in one of the languages has a slightly broader or narrower meaning than the preferred term Y in the other language. Table 4.3 shows a sample of partial equivalence terms in different languages. In this case, there are three possible solutions: i) treat the terms as exact equivalents., ii) adopt the terms from each language as loan terms in the other languages, and iii) treat the situation as single-to-many equivalence (see next case).

Table 4.3: Partial equivalence sample

German	English
Wissenschaft	Science

- *One-to-many equivalence (too many or not enough terms)*: to express the meaning of the term X in one of the languages, two or more terms Y are needed in the other language. The issue of one-to-many equivalence can be solved by using “coined terms”. A coined term represents a concept new to the target language, which accepts the concept and constructs a new term in its language to express it.
- *Non-equivalence*: no existing term Y with an equivalent meaning is available in the target language for a term X in the source language. Just like the previous case the solution is the “coined terms”.

We believe that the equivalence relationships above described can be represented using relations from the ontology language. For instance, using

OWL, it is possible to take advantage of `owl:equivalentClass` for describing an exact or near equivalence; `rdfs:subClassOf` for representing a partial equivalence; and `owl:disjointWith` for representing a non-equivalence. We consider that one-to-many equivalence is a special case of exact equivalence. These relations correspond to set-theoretic relations between classes: equivalence ($=$); disjointness (\perp); more general (\supseteq). They can be used without reference to any ontology language.

4.5 Summary of the Chapter

During this chapter we have first explained the terminology related to ontology localization, providing a standardized vocabulary for each one of the related terms. A description of the problems that have to be taken into consideration to localize an ontology has been presented later.

Then, we described the different scales of localization used to enrich an ontology to different natural languages. Also, we analyzed which elements or parts of the ontology are to undergo localization. Basically, we identified that the terminological layer is most clearly affected by the Localization Activity, since the labels that name ontology terms will have to be expressed in different natural languages.

The core of the chapter has been presented in Section 4.4 where we have detailed the different steps that involves the localization process. The different steps have been explained by means of a formal definition. We have showed that a variety of techniques can be used to develop multiple approaches for localizing ontologies. These techniques will be classified in the next chapter and further detailed in latter ones.

Chapter 5

Translation Techniques for Ontology Localization

Having defined what the ontology localization problem is, we attempt at classifying the methods and techniques that can be used for localizing an ontology to the linguistic level. All techniques proposed in this work aim to reduce the effort of localizing an ontology manually. Therefore, the major contributions to discovering appropriate ontology element translations arise from the machine translation discipline. The different analyzed approaches address the translation problem but from different perspectives, such as query translation in Cross-Language Information Retrieval (CLIR) or meta-data records translation for Multilingual Information Access (MLIA). In this work we have attempted to consider these approaches, focusing on translating ontology elements, and aiming to provide a common conceptual basis for their analysis.

In this chapter, we present a classification of different translation techniques based on the way of modeling the context used to disambiguate the candidate translations and the type of resources used to localize an ontology into different natural languages. To facilitate the analysis of these translation techniques we introduced a framework that covers their main aspects. Then, we present, at the strategic level, some natural ways to compose and combine the output of different translation techniques. Finally, we discuss an alternative for classifying the localization approaches.

5.1 Classification of Translation Techniques

Following the complexity of ontology localization, a variety of techniques exists that can be used to solve the different stages of this activity. However, in this thesis we only focus on those techniques that are directly related with the *translation phase* described in the previous chapter. A complete description of all stages of the ontology localization activity will be given in

Chapter 6.

As we already mentioned in the introduction of this chapter, for our classification we have analyzed state of the art approaches that address the translation problem from similar perspectives, because to the best of our knowledge, our contribution is the first attempt to classify translation techniques for ontology localization. In this respect, we have analyzed approaches used for both *query translation* in CLIR and *metadata records translation* (MRT). We used these approaches because the queries and metadata records have similar properties to the ontologies labels. In [McCrae et al., 2011a] the authors demonstrate that the terms used to designate concepts are frequently just noun phrases and are significantly shorter than a usual sentence. In the case of the relations between concepts (object properties) and attributes of concepts (data type properties), these are occasionally labeled by means of verbal phrases.

CLIR¹ refers to the retrieval of documents that are in a language different from the one in which the query is expressed. Since there are two different languages, a certain translation process is required to find a common representation through either query translation or document translation. According to the literature, query translation is a popular approach to CLIR, because although large-scale document translation approaches have been reported [Oard and Hackett, 1997], they have their limitations: computational expensiveness and restricted document representation by incomplete MT systems. In this thesis, we will ignore document translation approaches, because these do not present any problems that are specific to ontology localization. Compared with document translation, query translation is more flexible, and light-weight [Ballesteros and Croft, 1998, Jang et al., 1999, Gao et al., 2002]. However, this approach raises two particular problems: the selection of the appropriate translation terms/words, and the proper weighting of them [Grefenstette, 1998]. Notice that these problems are similar to those found in the translation of the lexical information of an ontology (see section 4.2 for more details).

MRT is the process of converting metadata records describing objects in a digital collection from one language into another [Chen et al., 2012]. As in CLIR, the main effort of this approach is to select an adequate translation strategy to improve the quality of the obtained translations. Nevertheless, in the last few years, researchers have worked on CLIR and MRT problems intensively, so, we will use this knowledge for our purposes.

For classifying translation techniques, we propose two categories based on the most salient properties of the ontology localization dimensions (see Figure 5.1). These categories are:

¹This task has also been termed multilingual, translingual, or cross-lingual IR by some groups. [Oard, 1997] contains a brief note on the different connotations of these terms and how they came about.

5.1. CLASSIFICATION OF TRANSLATION TECHNIQUES

- *Term Context Interpretation* classification is based on the way of modeling the context and the size or depth of the context used to disambiguate the candidate translations.
- *Type of Resources Used* classification is based on the type of resources used to localize an ontology into different natural languages.

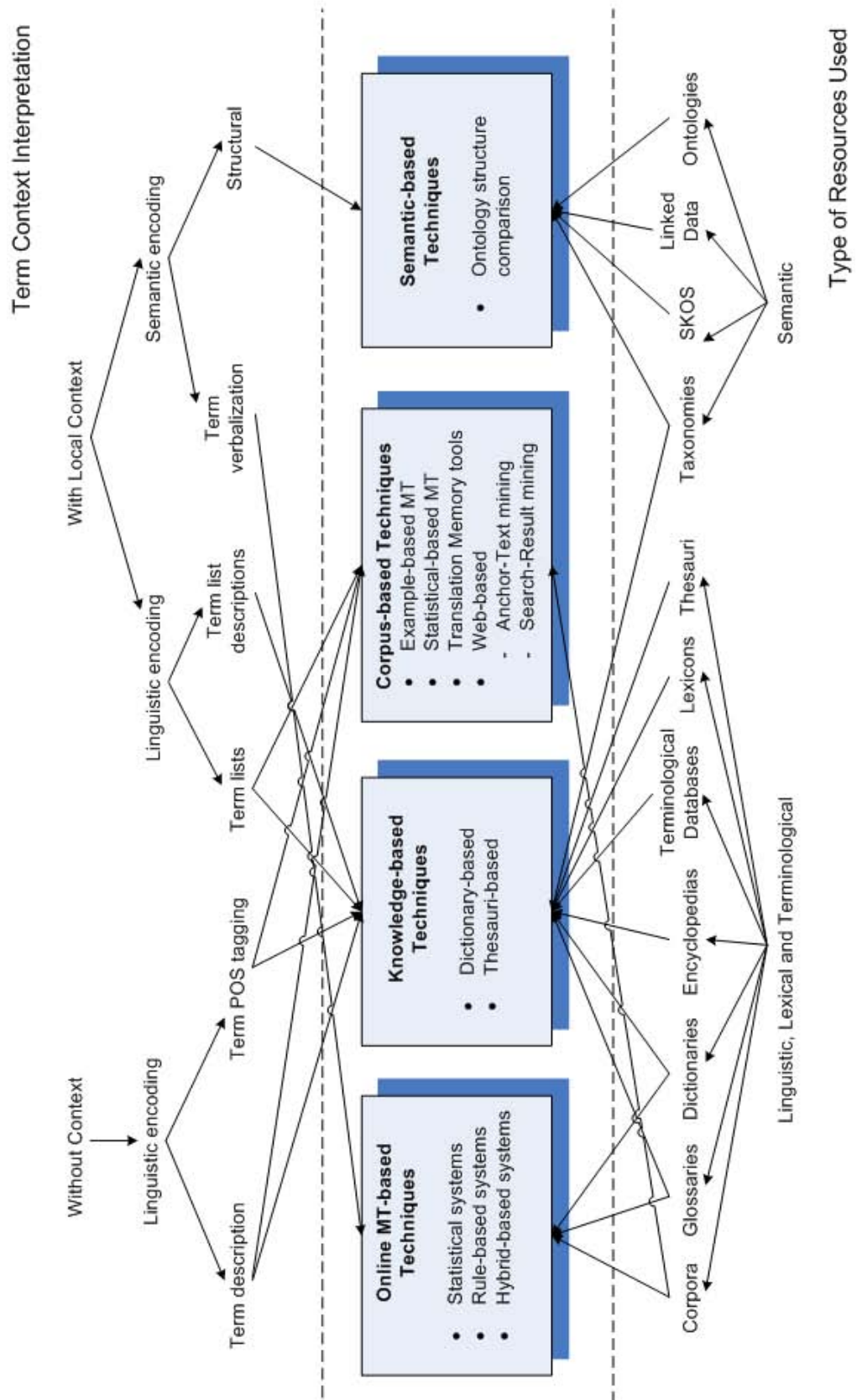


Figure 5.1: Classification of ontology localization techniques.

The overall classification of Figure 5.1 can be read both in descending (focusing on how the translation techniques interpret the context of each ontology term for disambiguating candidate translations) and ascending (focusing on the type of resources used for discovering candidate translations) manner in order to reach the *Translation Techniques*.

5.1.1 Term Context Interpretation

The *Term Context Interpretation* classification is concerned with the way of modeling the term context used for disambiguate the candidate translations:

- The first level is categorized depending on size or depth of the context: *without context* and *with local context*. These categories are assumed to be independent of the modalities used for encoding the context of each ontology term to be translated.
 - a. The *without context* approach uses only the information related to the term itself as context. This option is sometimes disregarded, but it contains important information about the internal structure of the ontology term, e.g., term annotation (see `rdfs:comment`), or type of term (concept, relation, or instance).
 - b. In the *with local context* approach the context involves a narrow group of terms centered on the ontology term itself, which fairly well approximates contexts starting from the immediately surrounding of direct relationship terms to the whole ontology.

We wish to point out that the division of context into different sizes allows for the showing of their relative influence on translation techniques. One can argue by example that there are no distinct boundaries between local context that uses a small set of terms and a local context that uses many related terms. There are only more or less influential context features, whose general tendency is that their influence diminishes with increasing distance from the ontology term itself.

- The second level of this classification decomposes these categories, taking into consideration the way of encoding the context of each ontology term to be translated. There are two different points of view for context pre-processing: *linguistic* and *semantic*.
 - a. The *linguistic encoding* processes the context of an ontology term as linguistic objects. Basically, the linguistic encoding approach uses the information obtained from the lexicon of the ontology in order to generate the term context.

- b. The *semantic encoding* processes the context as the entities that appear hierarchically organized in an ontological structure. In this approach of encoding, the context is obtained from the entities that are part of both lexicon and core ontology. In other words the semantic encoding makes use of all information of the ontology.
- The third level of this classification particularize the categories above mentioned in seven groups of syntactic and semantic context knowledge: *term description*, *term POS tagging*, *term list association*, *term description association*, *term verbalization*, and *structural context*. The first two groups have as context the information of term itself. The rest of the categories use a local context, but with the difference that both *term verbalization* and *structural* groups use a semantic encoding; the other groups use a linguistic encoding approach.

To illustrate the different ways of modeling the context of an ontological term, this section contains an ontology example of the university domain (see Figure 5.2). Concepts are depicted as rectangular boxes, relations as ellipses, annotation values as hexagons, and instances as rounded boxes. Ontology relations are drawn as solid arrows, whilst the instantiations of concepts and relations are depicted as dotted, arrowed lines. The example contains five concepts *person*, *professor*, *full professor*, *associate professor*, and *faculty*; one object relationship *belongsTo*; two attribute relationships *hasFullName* and *hasName*; and three instances *Computer*, *Edu*, and *Asun*. The example is fictitious and any concurrences with the real world are purely by chance.

- a. *Term description*. This category is represented by the use of a short description in the natural language of the ontology term under consideration. Usually these descriptions help clarify the meaning of the ontology terms. The `rdfs:comment` property can be used to define an ontology term description in the natural language (see RDF(S)² for more details). The term description context of the concept *professor* of our sample ontology can be:

$$ctx_{professor} := (\text{ a professor is a member of the faculty ...})$$

- b. *Term POS tagging*. In this case the context is represented by the use of the grammatical category of the term. In order to obtain the grammatical information of a term, the Part-of-Speech (POS) [Church, 1988, DeRose, 1988, Garside, 1987] tagging is a natural option. POS tagging is the process of assigning a part-of-speech like noun, verb, pronoun, preposition, adverb, adjective

²www.w3.org/TR/rdf-schema/

5.1. CLASSIFICATION OF TRANSLATION TECHNIQUES

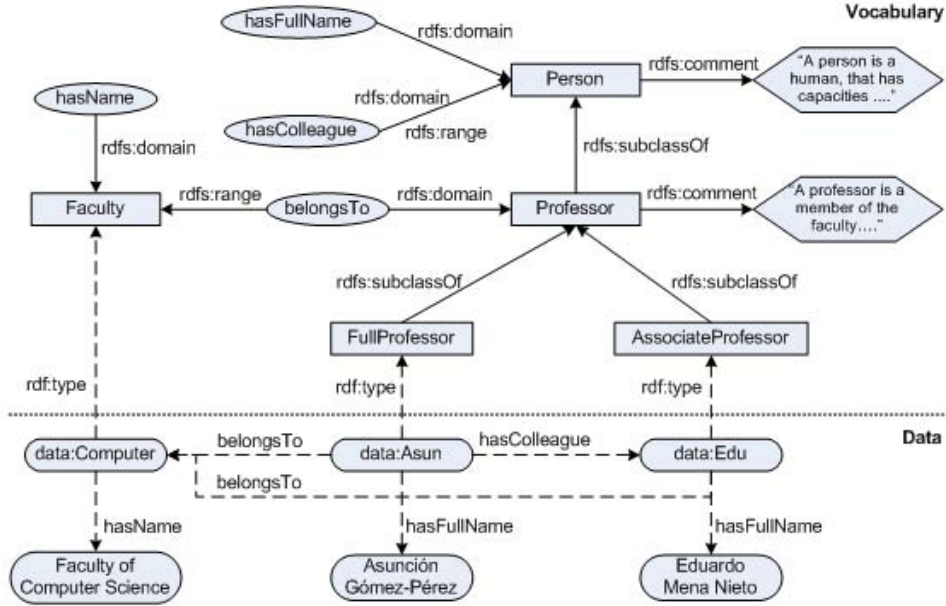


Figure 5.2: Ontology Example.

or other lexical class marker a word of a text. Most POS taggers³ need at least one short phrase from which it is possible to derive the lexical categories, or parts of speech of each word.

We have identified that for the majority of ontology compound labels (e.g., *AssociateProfessor*) it is not necessary to have an additional processing to determine the part of speech of each token. However, for obtaining the POS of a single term (e.g., *Professor*) additional information is required, i.e., the relationship with adjacent and related words in a phrase, sentence, or paragraph. One way to solve this problem is to use empirical rules to annotate a simple term. Based on our experience, we propose the following rules:

- The concepts, instances and attribute relations are considered nouns.
- All the rest of the terms (e.g., object relations) are considered verbs.

Another option is to try to generate a natural language sentence from the ontology term. In literature this process is known as

³A Part-Of-Speech Tagger (POS Tagger) is a piece of software that reads text in some language and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc. Available POS tagger tools can be consulted in the web page of the Stanford Natural Language Processing Group (<http://www-nlp.stanford.edu/links/statnlp.html#Taggers>).

ontology verbalization. Some authors have studied this problem extensively (see [Hewlett et al., 2005, Flied et al., 2007, Schober et al., 2007] for example). This approach can be used for verbalizing the ontology term and then use a part-of-speech tagger to discover the POS of the term.

The context of the concept term *professor* will be:

$$ctx_{professor} := (NN)$$

Where, NN represents a singular noun.

- c. *Term lists*. This category is represented by the use of a bag-of-words consisting of n words adjacent to the target ontology term. The list of terms obtained is independent of the semantic relationship between adjacent terms. Thus, for instance the context to depth two of the ontology term *professor* can be:

$$ctx_{professor} := (person, fullProfessor, associateProfessor, ...)$$

- d. *Term list descriptions*. In this category, the descriptions (in the natural language) of surrounding terms in the context are expanded to include descriptions of the terms related to subsumption relations in ontology. The natural language descriptions of each term can be extracted from the `rdfo:comment` property. For the ontological term *professor* the context can be:

$$ctx_{professor} := (\text{a professor is a member of the faculty ; a person is a human, that has capacities or attributes})$$

In the example, the second description belongs to the broader term *Person*.

- e. *Term verbalization*.⁴ To model the context using this approach, it is necessary to transform an ontology term into a natural language sentence. As we commented previously recent works already have studied the way of generating natural language sentences from ontology elements.

Intuitively, we can see that this option is an alternative to the approach previously described. Also, the ontology term verbalization has some advantages. In contrast to the *term list description* approach, where the descriptions not always define the exact meaning of the term, term verbalization reflects exactly the meaning of the ontological term. An example of the *term verbalization context* for the sample term *professor* is shown in the following:

⁴According to the Merriam Webster dictionary, one of the definitions of verbalization is to use words to express or communicate meaning. In this thesis we use this term in the same sense.

5.1. CLASSIFICATION OF TRANSLATION TECHNIQUES

$ctx_{professor} := (\text{ a Professor is a Person; a Professor belongs To Faculty;})$

- g. *Structural term context.* The structural context is encoded exactly as the entities appear together in an ontological structure. Also, the structural context uses all logical relations represented in an ontology, such as equivalence, subsumption, disjoint, etc.

5.1.2 Type of Resources Used

Following with the explanation of the categories used to classify the different translation techniques introduced in Figure 5.1, in this section we describe the classification of the type of resources used to perform a particular translation technique.

The classification is categorized depending on the richness of the internal structure of the resource: *linguistic, lexical and terminological* and *semantic* resources. The first type of resources groups together similar words without much distinction in the kind of similarity relation (e.g., linguistic databases, dictionaries, thesauri, Web, etc). The semantic resources on the other hand group together objects denoted by words (or more complex lexical items) according to a principled set of paradigmatic (meta-)relations like synonymy, hyponymy, meronymy, antonymy and syntagmatic (meta-)relations according to the dependency structure (e.g., lexical databases and ontologies). Notice that both types of resources can be used during the search and disambiguation of translation candidates.

In the following part we will give a brief overview of these resources (for more details, cf. [Ide and Veronis, 1998, Litkowski, 2005, Agirre and Stevenson, 2006]). Our purpose is not to analyze and compare the existing definitions of these resources, but to justify the convenience of their reuse in the ontology localization activity. Whenever possible, we will be referring to multilingual and online resources.⁵

Linguistic, lexical and terminological resources

The main linguistic, lexical and semantic resources involved in the classification of the translation techniques are the following:

- a. *Corpora.* According to [McEnery, 2003] corpora are defined as large collections of general or subject specific documents. Nowadays, corpus primarily means a collection of texts held in electronic form, capable of being analyzed automatically or semi-automatically rather than manually and for different purposes. Elaborate typologies of corpora have been proposed in literature ([Baker, 1995, Laviosa, 1997]) taking into

⁵The definitions and justifications of the resources here described, are a short summary of the deliverable found in [Espinoza et al., 2010].

consideration aspects such as: the relationship of translations between the different language sections of the corpus, and the number of languages represented in the corpus. According to this, the two main types of corpus are:

- *Parallel corpora* can be defined as corpora that contain source texts and their translations. Parallel corpora can be bilingual or multilingual. They can be uni-directional (e.g., from English into Chinese or from Chinese into English alone), bi-directional (e.g., containing both English source texts with their Chinese translations as well as Chinese source texts with their English translations), or multi-directional (e.g., the same piece of writing with English, French and German versions). These resources can be better exploited by Translation Memory tools, which align translation equivalents. Translation memory is a technology that enables the user to store translated phrases or sentences in a special database for local reuse or shared use over a network [Esselink, 2000].
 - *Comparable corpora*, in contrast, can be defined as corpora containing sets of texts that are collected using the same sampling frame and similar balance and representativeness [McEnery, 2003], e.g., similar features such as the same proportions of the texts of the same genres, in the same domains, in a range of different languages, in the same sampling period. However, the texts of a comparable corpus are not translations of each other. Rather, their comparability lies in their same sampling frame and similar balance. The Web as corpus offers a valuable resource for building and contrasting comparable corpora on the same domain. With the enormous growth of the Information Society, the Web has turned into a reliable test bed of data for natural language processing, not only in terms of data size but also in terms of data type (e.g., multilingual data, link data).
- b. *Glossaries*. These resources can be defined as alphabetical lists of terms or words found in or related to a specific topic or text. It may or may not include explanations, and its vocabulary may be monolingual, bilingual or multilingual [Wright and Budin, 1997]. These resources are of interest in the ontology localization activity because they usually contain the specific terminology of a domain. They can be monolingual or multilingual. In the case of monolingual glossaries, the most useful information they provide are definitions of terms, which can be used as contextual information for disambiguation purposes. If they are bilingual, they normally contain lists of translation pairs, which can be used in the translation process and need to be further disambiguated.

5.1. CLASSIFICATION OF TRANSLATION TECHNIQUES

- c. *Dictionaries*. The dictionaries are, according to [Varó and Linares, 1997], “books in which lexemes of a language are gathered and explained in the form of headwords or lemmas following an alphabetical order”. A machine-readable dictionary (MRD) is a dictionary in an electronic form that can be loaded in a database and can be queried via application software. It may be a single language explanatory dictionary or a multi-language dictionary to support translations between two or more languages, or a combination of both. MRDs are considered a valuable source of information for use in Natural Language Processing (NLP) because they contain an enormous amount of lexical knowledge. Some examples of MRDs that could be used in ontology localization are WordReference⁶, Wiktionary⁷, the Merriam-Webster’s Online Dictionary⁸ or Leo⁹.
- d. *Encyclopedias*. These resources are defined as documents that contain information on all branches of knowledge or treat comprehensively a particular branch of knowledge usually in articles arranged alphabetically often by subject (Glossary of Library Terms). Nowadays, one of the best-known online encyclopedias is Wikipedia. Wikipedia¹⁰ defines itself as a free, web-based, collaborative, multilingual encyclopedia project supported by the non-profit Wikimedia Foundation. Others resources of this type are DBpedia¹¹, which is a community effort to extract structured information from Wikipedia and make this information available on the Web or Freebase¹², which is a large collaborative knowledge base consisting of metadata composed mainly by its community members. Apart from the detailed information that can be found about a certain article, these resources are interesting for ontology localization because of two major reasons:
- They offer a huge source of structured information in different domains.
 - Most of the articles are multilingual and provide comparable corpora for translational purposes
- e. *Terminological Databases*. These resources are databases that contain the specific terminology of one or several domains of knowledge. They are similar to glossaries but usually contain additional data regarding the source from which the data has been obtained, language

⁶<http://www.wordreference.com>

⁷<http://es.wiktionary.org>

⁸<http://www.merriam-webster.com/>

⁹<http://dict.leo.org/>

¹⁰<http://en.wikipedia.org/wiki/Wikipedia>

¹¹<http://dbpedia.org/About>

¹²<http://www.freebase.com/>

usage examples, synonyms and related terms, etc. For the purpose of ontology localization, the multilingual terminology databases are interesting resources. Most international organizations maintain terminology databases to support the writing of technical documentation and its translation, as well as the communication between specialists. An example of a multilingual terminology database is IATE¹³, created and maintained by the European Union (EU).

- f. *Lexicons*. In a restricted sense, a computational lexicon is considered as a list of words or lexemes hierarchically organized and normally accompanied by meaning and linguistic behaviour information [Hirst, 2003]. One of the best known online English lexicon is WordNet. In addition to this, the EuroWordNet¹⁴ lexicon draws on WordNet structure to create wordnets in other languages and link them through a so-called Interlingual index, a list of unstructured meanings that provide the mappings across the wordnets. This kind of resources is very useful because its structure helps in disambiguating the different senses associated to words. In the case of EuroWordNet, it also provides translation candidates. The major drawback is that such resources contain general-purpose lexical entries, although in recent projects drawing on WordNet, wordnets containing the specific terminology of a domain are being developed (see the KYOTO¹⁵ project).
- g. *Thesauri*. Thesauri are controlled vocabularies of terms in a particular domain with hierarchical, associative, and equivalence relations between terms. Thesauri are mainly used for indexing and retrieving articles in large databases [ISO, 1986]. More specifically in the computer science domain, a thesaurus is defined as “a controlled and dynamic documentary language containing semantically and generically related terms”, which comprehensively covers a specific domain of knowledge. Two well-known multilingual thesauruses are Agrovoc¹⁶ and EuroVoc¹⁷. Both AGROVOC and EuroVoc have been migrated to semantic web technologies making use of the SKOS (Simple Knowledge Organization System) language. In this way, thesauri are easily queriable in the Web.

¹³<http://iate.europa.eu>

¹⁴<http://www.illc.uva.nl/EuroWordNet/>

¹⁵<http://xmlgroup.iit.cnr.it/kyoto/>

¹⁶The multilingual thesaurus of the Food and Agriculture Organization of the United Nations <http://aims.fao.org/website/AGROVOC-Thesaurus/sub>

¹⁷<http://eurovoc.europa.eu/drupal/>

Semantic resources

The semantic resources that can be used to perform a particular translation technique are the following:

- a. *Taxonomies*. These resources comprise an organized list of concepts that are drawn from diverse data sources and organized according to an expert in the domain [Boiko, 2005]. Taxonomies are very common in the biological domain, but we also find many about economical or industrial activities, occupation, etc. See for instance, the Standard Industrial Classification¹⁸ (SIC) of the United States, ESCO¹⁹ taxonomy for employment in Europe. These classifications can be useful resources to ontology localization because they contain the specific terminology of a certain domain, and with a certain degree of structure.
- b. *SKOS*. The Simple Knowledge Organization Systems (SKOS) [Miles et al., 2005] is a W3C recommendation designed for representation of thesauri, classification schemes, taxonomies, subject-heading systems, or any other type of structured controlled vocabulary. Using SKOS, each term in each taxonomy can be represented in a machine readable format containing definitions, labels, and related concepts for the term expressed in SKOS. The SKOS framework allows associating labels and definitions in multiple languages to any concept. This means that we can associate the labels “Dispositivos móviles”@es, “appareils mobiles”@fr or “Mobile Geräte” to the concept “Mobile_device” to include the Spanish, French and German labels. Well-known controlled vocabularies such as EuroVoc have been expressed using an ontology that extends SKOS. All data objects supported by SKOS for handling labels, can be useful to concept identification, disambiguation and translation in ontology localization.
- c. *Linked Data*. The term Linked Data refers to a set of best practices for publishing and connecting structured data on the Web [Bizer et al., 2009]. The main objective of Linked Data initiative is connecting data from diverse domains to enable new types of applications. Thanks to the links created between the data, these data can be browsed and queried starting in one data source and navigating along the links to other related data sources. This potentially augments the possibilities of obtaining relevant data. The Linked Data can contain information of many domains of knowledge, being the most represented nowadays: media, geography, publications, e-Government, and live sciences. There are also resources that contain general information, such as DBPedia.

¹⁸<http://www.sec.gov/info/edgar/siccodes.htm>

¹⁹<http://esco.tenforce.com/esco-browser/>

These resources in the Linked Data format, and specifically DBPedia, may have an enormous potential for the ontology localization activity. They do not only offer structured information of a certain domain of knowledge, but also numerous links to related information. Those typed links are very useful in the disambiguation process. Although most of the resources on the Linked Data cloud are monolingual in English, in the near future we expect many of them to be in other languages.

- d. *Ontologies*. Formally, an ontology consists of terms, their definitions, and axioms relating them [Gruber, 1995]; these resources can be viewed as computational knowledge organization systems for domain specific text and domain specific knowledge. The ontologies have become crucial instruments of knowledge management processes, since they provide a formalized, hence conceptualization of a specific knowledge area that is usually contained in domain specific text corpora. In localization, and particularly in machine translation, ontologies have been used to improve the performance of translation systems [Hovy et al., 2001] by enhancing the knowledge base that supports the linguistic algorithms of source language text analysis and target language text generation. The way of accessing ontologies which are available on the Web is by means of Semantic Web search engines, such as Watson²⁰, or Swoogle²¹. This allows us to see how a certain concept has been described (by means of properties and relations) in a certain ontology. The ontology localization activity would greatly benefit from the availability of multilingual ontologies to obtain translation candidates. However, multilingual ontologies are still scarce on the Web, and mechanisms should be developed to access and query multilingual ontologies.

Once we have presented a brief description of the types of resources that can provide valuable information when localizing an ontology, we now discuss in more detail the main classes of Translation Techniques according to the above classification.

5.2 Basic Translation Techniques

In this section, we introduce the main characteristics of the different translation techniques and methods shown in Figure 5.1 (see middle layer). To facilitate the analysis of these techniques we have designed a framework which covers their main aspects:

²⁰<http://watson.kmi.open.ac.uk/>

²¹<http://swoogle.umbc.edu/>

5.2. BASIC TRANSLATION TECHNIQUES

1. *Aims and scope.* This section includes information about aims and purpose of the translation technique, core resource, and core MT technology used.
2. *Methods employed.* This one is the largest section of the evaluation framework, in which the translation technique is detailed, describing the main methods available to obtain the ontology translations. Different methods have been proposed in literature to discover the translations of simple and compound words and short phrases. However, in this thesis we focus on those approaches that incorporate source context information to improve the translation quality. Thus, the proposed techniques will differ in the way that term contexts are detected and exploited.

In this section, we focus the discussion on the following three basic research tasks:

- *Resource pre-processing*²²: Each one of the MT techniques identified in the middle layer of figure 5.1 bases its operation on some lexical resource (e.g., dictionaries, corpora, etc). In some cases, these resources need some mechanism of processing before being used; this section describes this process.
- *Translation candidate extraction*: This section describes the process used to identify possible candidate translations.
- *Translation selection*: Once the candidate translations have been identified, this section describes ranking methods based on word sense disambiguation to select the more appropriate translations.

In some cases it is possible to generalize the technique by using only one mechanism of translation. For all the methods we use the terms: word, text, segment, phrase, short phrase sentence, and paragraph in order to refer to the different elements that can be translated in one ontology. These elements can be: simple ontology labels, compound labels, or term annotations.

3. *Advantages* In this section we specify the main advantages of the translation technique in the localization process.
4. *Disadvantages.* The main limitations of the technique under consideration are described in this section.

As we will explain in the following sections, all techniques have drawbacks that could be overcome by combining some of the strategies that we will introduce in the section 5.8.

²²The absence of this section in the method means that it does not require a resource pre-processing.

5.3 Online MT-based Techniques

Online MT-based techniques take advantage of the recent availability of on-line translation services such as Bing Translator²³, GoogleTranslate²⁴, or Yahoo Babelfish²⁵. These systems use rule-based engines and/or statistical-based engines for its operation, so, they rely on countless built-in linguistic rules, bilingual dictionaries and glossaries for each language pair and monolingual and bilingual corpora.

5.3.1 Methods Employed

Depending on the online MT systems used to discover the translations of the elements of an ontology, we have three methods: *statistical-based*, *rule-based* and *hybrid-based systems*. The hybrid-based systems leverage the strengths of statistical and rule-based translation methodologies. In all cases the translation process is as follows:

- *Translation candidate extraction:* The basic approach to use an online MT system in ontology localization is simple: one just has to submit the ontology element to an MT system to obtain a translated version. In the literature some works show that better performance in terms of output quality can be achieved when these systems can process the texts that they are required to translate into smaller chunks [Wu et al., 2008a]. Our own experience with these systems suggests a high efficiency in the translation of compound labels and short phrases. However, if a simple word is submitted, then there is a high chance that the word will be translated by its default translation.

To solve the translation of simple labels we have investigated the use of *term verbalization context* as translation input. Remember that *term verbalization context* produces a short natural language phrase of the term in the ontology. However, this solution involves the use of different word/phrase alignment tools and algorithms for identifying translation relationships among the words in a bitex²⁶ as used in statistical machine translation (see section 5.5.1). Of course, further investigation is necessary to evaluate this approach as a plausible solution in order to mitigate the lack of context in simple labels.

- *Translation selection:* Each online translation service uses its own disambiguation method which does not allow to be customized. In other

²³<http://www.microsofttranslator.com/>

²⁴<http://translate.google.com/>

²⁵<http://babelfish.yahoo.com/>

²⁶In the field of translation studies a bitext is a merged document composed of both source and target language versions of a given text.

5.4. KNOWLEDGE-BASED TECHNIQUES

words, these methods do not use the context of the term to be localized to rank their translations. Despite the fact that we are interested in translation methods that use source context information to improve translation quality, we include these systems as part of the classification due to improvements in quality reported in similar domains (e.g., translation of metadata records [Chen et al., 2012] or query translations [Wu et al., 2008b]).

5.3.2 Advantages

Online MT services have proven to be an elusive goal in localization, but today a number of systems are available which produce output which, though not perfect, is of sufficient quality to be useful in a number of specific domains. Also, these services save time while translating large texts and often allow for customization by domain or user-specific settings e.g., choosing between American English and British English.

5.3.3 Disadvantages

Normally the output of these systems is limited to one per word, while there are multiple expressions for it in the target language. For example, both “drogue” and “stupfiant” are correct French translations of “drug” in the sense of illegal substance, but both Babelfish and Google only choose “stupfiant” in their translations of “drug traffic”. These translation services do not suggest non-translation, but strongly related words in the translation results. However, strongly related words can be very useful in ontology localization, even if they are not translation words. For example, it may be useful to “translate” the word “computer” by the French word “programme” even if the latter is not a literal translation of the former. This latter term may help retrieve other related terms, which could be relevant.

Another problem with these systems is the difficulty to translate unknown words, or out-of-vocabulary words (often referred to as OOV [Qu et al., 2012]). A typical case is the translation of ontology entities representing names of persons or organizations. Also, no extensions can be made to these systems, e.g., addition of very specific domain terms or proper names. Finally, the translation produced by these translation services is often limited to a certain number of characters per day.

5.4 Knowledge-based Techniques

Knowledge-based techniques rely on dictionaries, terminology databases, glossaries, encyclopedias, thesauri or lexical knowledge bases, without any corpus evidence to generate the target translations. These resources provide information such as examples, definitions, or semantic hierarchies and

associations could be used to help select more appropriate translations in the context. Basically these methods use a direct translation approach and they rely on similarity measures computation to disambiguate the candidate translations (e.g., Pedersen et al. [Pedersen et al., 2005], Resnik [Resnik, 1999]).

5.4.1 Methods Employed

In the literature the knowledge-based techniques are normally classified into: *dictionary-based* and *thesauri-based* approaches. This distinction takes into account the grade of structured information contained in the resource. For our purposes we use the same categorization:

Dictionary-based

Dictionary-based methods take advantage of the multilingual linguistic information available in machine readable dictionaries, glossaries, encyclopedias or terminological databases to discover the translations. The main difficulty of the dictionary-based techniques is to select the correct translation of a term among all the translations provided by these resources.

- *Translation candidate extraction:* Some of these resources require a normalization process before each word is submitted as a query to the translation source. The normalization process involves for example, transform the word to singular form, verbs in the infinitive form, and adjectives in their positive form. After the normalization, the resource returns a set of translations any time the label exactly matches a word in the source entries.

All these resources offer an exact or fuzzy search mechanism to extract the candidate translations. To prevent an explosion of nuisance matches, the *term POS taggig* context can be used. POS information has shown to solve 87% of all word ambiguities [Wilks and Stevenson, 1997]. This is useful, since dictionaries in general have separate hierarchies for words of different POS, and contemporary POS-taggers are of high accuracy [Brill, 1995]. With this context information we can only retain those translations whose POS exactly match with the POS of the search term. This assumption is accomplished in the majority of languages.

- *Translation selection:* In spite of the process of selection performed in the previous step, it is common for a single word to have several translations, some with very different meanings. To disambiguate the senses of a source term, we can employ mainly the example sentences, definitions or related terms listed for each sense division of a source

word. The Lesk algorithm [Lesk, 1986], in which the most likely meanings for the words in a given context are identified based on a measure of contextual overlap among dictionary definitions pertaining to the various senses of the ambiguous words, provides reasonable disambiguation precision for these types of resources. If additional resources are available (e.g., a set of semantic relations from a semantic network or a minimal set of annotated data) other methods can be applied (see *Translation selection* section in the Thesauri-based techniques).

Thesauri-based

Thesauri-based methods take advantage of resources with semantic hierarchies and associations (such as lexicon or thesaurus), to generate the target ontology translations.

- *Translation candidate extraction:* The process used to discover candidate translations is similar to the method introduced for the *Dictionary-based techniques*.
- *Translation selection:* In addition to Lesk algorithm, we can disambiguate the candidate translations using the implicit relations such as synonym, hypernym, hyponym, etc., found in this type of resources. In effect, the senses of surrounding words in the context can be expanded to include the senses of these related words to which they are semantically related through extended relations. Different measures of semantic similarity can be used for ranking the candidate translations. In the next part we list the measures that are more relevant for the purposes of this thesis:
 - *Variations of the Lesk Algorithm:* Among all variations of this algorithm, the simplified Lesk method [Kilgarriff and Rosenzweig, 2000] is the one that improves most in comparison to the original algorithm both in terms of efficiency (it overcomes the combinational sense explosion problem) and precision (comparative evaluations have shown that this alternative leads to better disambiguation results). In this simplified algorithm, the correct meaning of each word in a text is determined individually by finding the sense that leads to the highest overlap between its dictionary definition and the current context. Another variation of the Lesk algorithm, called the adapted Lesk algorithm, was introduced by Banerjee and Pedersen [Pedersen et al., 2005], which extended gloss overlaps through the rich network of word sense relations in Wordnet rather than simply considering the glosses.
 - *Measures of semantic similarity computed over semantic networks:* There are a number of similarity measures that were developed to

quantify the degree to which two words are semantically related. Most such measures rely on semantic networks and follow the original methodology proposed by Rada et al. [Rada et al., 1989] for computing metrics on semantic nets. These measures include methods for finding the semantic density/distance between concepts. A comprehensive survey of semantic similarity measures is reported by Budanitsky and Hirst [Budanitsky, 2001].

More detailed reviews about different knowledge-based disambiguation techniques can be found at [Agirre and Stevenson, 2006].

5.4.2 Advantages

The resources that establish these techniques are widely available, dictionary-based approaches are easy to implement, and these resources have the ability to produce consistent, high-quality translations (conditional to the quality of the original bilingual resource).

5.4.3 Disadvantages

One of the main problems associated with dictionary-based techniques is untranslatable words due to the limitations of general resources. The category of untranslatable words involves new compound words, special terms, and cross-lingual spelling variants, i.e., equivalent words in different languages which differ slightly in spelling, particularly proper names and loanwords. The problem of missing translations can be addressed by automatically mining additional translation relations. We leave this problem to Section 5.5.1.

5.5 Corpus-based Techniques

These methods use a parallel or comparable corpus of aligned documents to discover translations. The criteria used for alignment combine linguistic and statistical information.

5.5.1 Methods Employed

Many approaches have been proposed to extract translation relations from parallel or comparable corpora. In the following, we will describe some representative approaches:

Example-based MT

The philosophy of example-based machine translation (EBMT) [Nagao, 1984] combines the features of rule-based and statistical approaches in a manner that seems favorable for the task at hand. The main idea behind EBMT

is that a given input phrase in the source language is compared with the example translations in the given bilingual parallel text to find the closest matching examples that can be used in the translation of that input phrase.

One of the main approaches in the EBMT paradigm is to use pattern matching techniques. First, these approaches collect word sequences from each corpus using translation patterns to acquire candidates for bilingual expressions. Second, a search for pairs of words that satisfy the correspondences of the sequences is performed. Therefore, a pre-processing step such as part of speech tagging and syntactic category identification is necessary to apply this method.

- *Resource pre-processing:* Before discovering candidate translations, a bilingual template acquisition from a simple monolingual corpus or parallel corpora has to be completed. In general, this process involves three phases: retrieving local patterns, assigning their syntactic categories with part-of-speech (POS) templates, and making translation patterns.
 - *Retrieving local patterns.* In order to retrieve local patterns any method for retrieving word sequences may be used [Kansai et al., 1996, Sato and Saito, 2002]. These methods generate all n-character (or n-word) strings appearing in a text and filters out fragmental strings with the distribution of words adjacent to the strings. This is based on the idea that adjacent words are widely distributed if the string is meaningful, and are localized if the string is a substring of a meaningful string.
 - *Identifying syntactic categories.* Since the strings are just word sequences, this task gives them syntactic categories. Thus, this task involves the assignation of part-of-speech tags for each component word discovered in the previous step. A syntactic category can be used to group similar tagged words. For example, the syntactic category NN can be used to group the following sample POS templates, (word) (word) or (word) (preposition) (word). In the example NN represent a noun phrase.
 - *Making translation patterns.* The final process is to generate the bilingual translation patterns. In the case of using a monolingual corpus as base to discover the patterns, we need to translate each word (identified in step one) as previous step to identify its syntactic categories.

The output of this process is a repository of lexical templates for MT.

- *Translation candidate extraction:* The term POS tagging context could be valuable to filter and prune the extracted candidate translations.

To retrieve candidate translations, we can collect the n-grams of POSs appearing in a translation pattern (e.g., NN, JN, etc.) from each corpus. As this method simply extracts word sequences according to POS tags, it also collects noisy sequences. However, most meaningless sequences can be eliminated, estimating different types of word similarity correspondences.

- *Translation selection:* After generating a ranked list of translation candidates for each source term, ranking techniques must be used to estimate the coherence of the translated label and decide the best translation. The ranking factor can be estimated using one of the techniques described below:
 - *Ranking through Web.* The Web can be considered as an exemplar linguistic resource for decision-making [Grefenstette, 1999, Li et al., 2003]. In this approach, each candidate translation is sent to a Web search engine (e.g., Google) to discover how often the combination of translation alternatives appears. The number of retrieved Web pages in which the translated sequence occurred is used to rank the translation candidates.
 - *Ranking through a test collection.* Large-scale test collections could be used to rank the translation alternatives and complete a final translation. We can follow the same steps as the previous technique, replacing the Web by a test collection and a retrieval system to index documents of the test collection.
 - *Ranking through an interactive mode.* An interactive mode [Ogden and Davis, 2000] could help solve the problem of identifying final translations. The interactive environment setting should optimize the label translation, select best translation alternatives and facilitate the information access across languages. For instance, the user can access a list of all possible candidates ranked in a form of hierarchy on the basis of word ranks associated to each translation alternative.

Statistical-based MT

These approaches analyze large collections of texts on a statistical basis and automatically extract the most probable translations in the target language [Peters and Sheridan, 2000]. The recent progress in SMT suggests interesting future development for ontology localization [Stroppa et al., 2007, Gimpel and Smith, 2008]. In particular, phrase-based translation approaches have become the state of the art in SMT, while these approaches have not yet been widely investigated in localization. A recent work [McCrae

et al., 2011a] analyzes different translation strategies using statistical machine translation approaches that also utilize the semantic information beyond the label or term describing the concept, that is relations among the concepts in the ontology, as well as the attributes or properties that describe concepts:

- *Resource pre-processing:* Bilingual word/phrase alignment is the first step of most current approaches to SMT. Alignment is a vital issue in the construction and exploitation of parallel corpora. The alignment methodology tries to identify translation equivalence between sentences, words and phrases within sentences. In most literature, alignment methods are either categorized as association or estimation approaches (heuristic and statistical models). Association approaches use string similarity measures, word order heuristics, or co-occurrence measures (e.g., mutual information scores). The major distinction between statistical and heuristic approaches are that statistical approaches are based on well-substantiated probabilistic models while heuristic ones are not. Most current SMT systems use a generative model for word alignment such as the one implemented in the freely available tool GIZA++ [Och and Ney, 2003]. GIZA++ is an implementation of the IBM alignment models [Brown et al., 1993]. These models treat word alignment as a hidden process, and maximize the probability of the observed (e, f) sentence pairs using the Expectation Maximization (EM) algorithm, where e and f are the source and the target sentences.
- *Translation candidate extraction:* To discover candidate translations, SMT-based methods generally use the occurrence frequencies of substrings of the sentence in target-language corpora. The score assigned to each candidate translation depends on both: i) the extent to which the source sentence meaning is also expressed in the candidate translation, and ii) the extent to which the candidate translation is likely to be a valid sentence in the target language regardless of whether or not its meaning bears any relationship to the source sentence. Details of how this score is computed is out of the scope of this thesis, however this information can be consulted in [Hearne and Way, 2011].
- *Translation selection:* In order to discover the final translations, the approach introduced in [McCrae et al., 2011a] uses word sense disambiguation by comparing the structure of the input ontology to that of an already translated reference ontology. We found this method to be very effective in choosing the best translations. However it is dependent on the existence of a multilingual resource that already has such terms. As such, we view the topic of taxonomy and ontology translation as an interesting sub-problem of machine translation and believe

there is still much fruitful work to be done to obtain a system that can correctly leverage the semantics present in these data structures in a way that improves translation quality.

Translation Memory tools

The essential idea behind these techniques is the use of a linguistic database (also called translation memory) in order to reuse previously translated words. These techniques are often used in order to compare segments in the source text with the translated segments in the translation memory. For our purposes, a segment can consist of simple ontology labels, compound labels, or term annotation paragraphs.

- *Translation candidate extraction:*. Linguistic databases provide a number of efficient search options to extract candidate translations:
 - *Fuzzy matching.* This is the dominating approach for the retrieval of similar segments from translation memories, because the possibility of exactly repeated segments is small, except in the context of re-translating the labels of a modified resource (in our case an ontology). The method can be based on orthographic similarities, which can be efficiently computed by comparing the number of corresponding substrings (e.g., bi- or trigrams) of two segments [Willett and Angell, 1983, Rapp, 1997]. Another option to measure the distance between two fuzzy matching content segments is to use the Levenshtein algorithm [Levenshtein, 1965]. The Levenshtein distance between two strings is given by the minimum number of operations needed to transform one string into the other, where an operation is an insertion, deletion, or substitution of a single character.
 - *Syntax trees.* This approach requires natural language parsers for both languages to be considered. The parse tree of the segment to be translated is compared to the parse trees of all source language sentences in the linguistic database. If an identical parse tree is found, it is assumed that the parse tree of the correct translation should be identical to the parse tree of the corresponding target language sentence retrieved from the linguistic database [Maruyama, 1992]. The main problem with this approach is that high quality parsers for unrestricted languages are not available for many languages. Also, the disambiguation of semantically ambiguous words is not always possible by only considering the syntax.
- *Translation selection:* Basically, the process of selection is a manual labor, in which the user performs the dominant role and makes the

final decisions concerning the chosen translations. A solution to this problem is to describe the linguistic database data with the Translation Memory eXchange format²⁷ (TMX). TMX is an open standard that uses XML for the archiving and mutual exchange of the Translation Memories (TM). In TMX a translation unit²⁸ can contain markup content elements, which can be used to disambiguate the candidate translations. For example, using the *term POS tagggig* context, we can to select only those annotated translations whose POS match exactly with the POS of the searched term. This assumption is accomplished in the majority of languages.

Web-based

Particularly for domains where sufficiently large text corpora are not available, or accuracy and coverage of translation dictionaries are rather low, Web-based translation methods are a good alternative.

These models propose to mine translations from Web corpora specially for discovering OOV term translations. OOV terms principally consist of short phrases such as named entities (person, location or organization names), book and movie titles, science, medical or military terms and others²⁹. Therefore, we consider that the Web-based methods can be used to discover the translations of instances and very specific domain terms.

Anchor-Text Mining Method

This approach searches the Web for parallel text and extracts translation pairs among anchor texts pointing together to the same webpage [Lu et al., 2002]. An anchor text is the descriptive part of an out-link of a Web page used to provide a brief description of the linked Web page. For instance, the text part “Apple” is the anchor text in the example below.

```
<a href="http://en.wikipedia.com/wiki/Apple_Computer">Apple</a>
```

This method supposes that for a source term appearing in the anchor text of a Web page, it is likely that its corresponding target translations may appear together in other anchor texts linking to the same page.

- *Resource preprocessing:* As a first step, this method needs to extract the Web pages whose anchor-text sets contain both source and target terms. In order to collect large numbers of pages from the Web and

²⁷<http://www.lisa.org/fileadmin/standards/tmx1.4/tmx.htm>

²⁸In TMX, an entry consisting of aligned segments of text in two or more languages is called a Translation Unit.

²⁹Some names are single word, which could be regarded as one-word phrases.

build up a corpus of anchor-text sets, a Web crawler³⁰ needs to be implemented.

- *Translation candidate extraction:* Considering that an anchor text might be a short text, heading, phrase, or URL, the term extraction process needs to extract key terms as translation candidates from the anchor-text corpus. Different methods can be used to extract translation candidates:
 - *PAT-tree-based:* The PAT-tree-based keyword extraction method is an efficient statistics-based approach that includes n-gram modeling, and completeness and significance analysis of semantics [Chien, 1997]. The advantage of this method is the ability to extract many significant terms and phrases without the limitations of string length and that of using a dictionary.
 - *Query-set-based:* This method takes user queries from real-world search engines as vocabulary sets to segment key terms in anchor-text sets. All the query terms in the target language are taken as translation candidates and their similarity to the source query is estimated.
 - *Tagger-based:* This method uses a tagger system, to segment the texts into meaningful words and to extract unknown words such as proper nouns and new terms. This method is different from the PAT-tree-based method in that it is more linguistically-based.
- *Translation selection:* The process of selection assumes that a translation candidate has a higher chance of being a translation only if it frequently co-occurred with the source term in the same anchor text sets. To estimate the degree of similarity between a source term and each translation candidate that co-occurs in the same anchor-text sets, any symmetric similarity measure can be used. In literature, different works use a function based on the probabilistic inference model [Wong and Yao, 1995] for these purposes.

Search-result mining.

These methods are based on the observation that for many source language search-result pages, there are rich snippets of summaries with a mixture of source and target texts. Given an input term in a source language, the search engine searches the translation terms in documents written in other

³⁰ According to Wikipedia, a Web crawler is a computer program that browses the Web in a methodical, automated manner or in an orderly fashion. Other terms for Web crawlers are ants, automatic indexers, bots, Web spiders, Web robots, or -especially in the FOAF community-Web scutters.

5.5. CORPUS-BASED TECHNIQUES

languages. The returned snippets containing the term are collected and translations are extracted from the snippets.

Although a quite large amount of term translations can be acquired using a search snippet-based mining scheme, the scheme may fail to extract low frequency term translations. If a term translation pair occurs only a few times on the Web, the translation of the term may not be retrieved by the search engine since the search engine ranks Web pages based on the PageRank algorithm which is irrelevant to the occurrence of its translation. As a result the top-n returned snippets may not contain the translation.

- *Resource-preprocessing:* The collection process of Web pages is performed using a Web search query. Basically, there are two approaches for building the query: i) using a monolingual query for source language pages containing the target language terms [Cheng et al., 2004], or ii) using cross-lingual query expansion [Zhang et al., 2005]. In the last approach to search for pages containing the term to be translated and its translation, the Web search query contains the term and one hint word generated by cross-lingual query expansion.
- *Translation candidate extraction:* The terminology translation mining performs a preprocessing on Web snippet texts by filtering out HTML tags, punctuation marks and non-query source words. Then, it extracts the translation from the processed top-N snippets, and provides confidence scores for each translation candidates.
- *Translation selection:* In order to select the translations these methods rely on different term similarity estimation techniques. Different measures have been proposed in literature for estimating the association between words/phrases based on co-occurrence analysis, including mutual information, the DICE coefficient, and statistical tests, such as the chi-square test and the log-likelihood ratio test.

5.5.2 Advantages

Translation corpora are an ideal resource for establishing equivalence between languages since they convey the same semantic content. Also, these techniques can be quickly adapted to new language pairs since the algorithms are almost language independent and most language specific information is automatically derived from parallel corpora. Finally, most of these methods can achieve high translation accuracy.

5.5.3 Disadvantages

While this method alleviates the problem of limited scalability found in the previous approaches, it relies on the existence of a parallel corpus in the

desired domain, which is often an unreasonable requirement. It is not always possible to find corpora of different languages and domains, together with the fact that corpus annotation requires a lot of effort and resources. In case of Web-based translation methods, there are some issues that need to be solved before using the Web information to mine terminology translation: i) how to find more comprehensive results, i.e. mining all possible forms of annotation pairs in the Web, and iii) how to remove the noises formed in the statistics and rank the remaining candidates.

5.6 Semantic-based Techniques

The semantic-based techniques take advantage of linked data and ontologies resources that provide a formal description of concepts, terms, and relationships within a given knowledge domain. In this work we have mainly used these techniques to disambiguate the identified candidate translations from the other approaches.

5.6.1 Methods Employed

Our intuition in approaching the ontology translation is that the comparison of ontology or taxonomy structures containing source and target labels may help in the disambiguation process of translation candidates [McCrae et al., 2011a]. A prerequisite in this sense is the availability of equivalent (or similar) ontology structures to be compared; we briefly summarize the main steps of our approach, named *ontology structure comparison*.

- *Translation candidate extraction:* As a first step, this method tries to discover the different senses of a term, using for this purpose semantic descriptions available in different sources of knowledge. The semantic knowledge can be obtained from available online ontologies accessed by means of ontology search engines. These tools crawl the Web to obtain different types of semantic information such as ontologies, instance data, and specific terms i.e., URIs that have been defined as classes and properties.

Some ontology search engines require a normalization process before each word is submitted as a query. The normalization process involves rewriting the words in lower-case, removing hyphens, etc. After normalization, the ontology search engines return different ontological terms that match those normalized keywords. The main advantage of using a pool of ontologies instead of just a single one is that many technical or subject-specific senses of a term cannot be found in just one ontology. For each term obtained from the ontology search engines, a sense is built. Each sense is represented by means of the hierarchical graph of hypernyms and hyponyms of synonym terms found in

one or more ontologies. Thus, senses are built with the information retrieved from matching terms in the ontology pool. Notice that the more ontologies or knowledge bases accessed the more chances to find the semantics of a term. As matching terms could be ontology concepts, attributes or instances, three lists of candidate keyword senses are associated with each normalized keyword: concepts, attributes and instances. The result of this process is a list of possible senses for each word.

- *Translation selection:* This step uses the *structural context* information for ranking the different senses obtained in the previous step according to the similarity with its lexical and semantic context. To estimate the probability of synonymy, in other words the degree in which the words are related, any semantic relatedness measure can be used. These measures consider not only similarity between the words, but any possible semantic relationship between them [Gracia and Mena, 2009]. To avoid the use of a cross-language semantic measure as the source and target senses are expressed in different natural languages, the external ontologies can be limited to those resources that have linguistic information in other languages.

5.6.2 Advantages

The main advantage of this approach is the increased availability of online semantic resources. Making the best use of such resources leads to a higher quality translation with lower development costs.

5.6.3 Disadvantages

While these techniques allow for the obtaining of more exact translations, the lack of ontologies enriched with linguistic information into different natural languages implicates the use of cross-lingual semantic disambiguation measures. As these measures generally use the Web as multilingual corpus to establish the similarity, the translation process can be very slow.

5.7 Analysis of Translation Techniques

Taking into account the advantages and shortcomings of the different techniques introduced in the previous sections, we believe ontology translation techniques do not always clearly fall into one or the other of the four broad categories – online MT-based, knowledge-based, corpus-based and semantic-based; many techniques could combine features of different approaches. For instance, online MT-based techniques may be used to generate candidate

translations and corpus-based or semantic-based techniques to disambiguate translated ontology terms – these could be thought of as hybrid approaches.

In fact, in this work we propose as hypothesis that an appropriate combination of the previous translations techniques leads to better localization results than only using one at a time. Attempts at combining outputs from different systems have proven useful in many areas. For example, people in the speech community pursued the idea of combining off-the-shelf Automatic Speech Recognizers (ASRs) into a super ASR for some time, and found that the idea works (Fiscus [Fiscus, 1997], Schwenk and Gauvain [Schwenk and Gauvain, 2000], Utsuro et al. [Utsuro et al., 2003]). In Information Retrieval (IR), we find some efforts going (under the name of distributed IR or meta-search) to selectively fuse outputs from multiple search engines on the Internet (Callan et al. [Callan et al., 2003]). In Ontology Engineering, some ontology matching systems are using the combining of different matchers to produce a more efficient matching algorithm. In Machine Translation, different multi-engine MT systems have been designed as an attempt to integrate the advantages of different translation systems without accumulating their shortcomings.

In the next section we present at the strategic level, some natural ways to compose different translation algorithms to localize an ontology.

5.8 Ontology Localization Strategies

In the previous section we described a variety of translation techniques that can be used to localize an ontology to the linguistic level. We also showed that all these approaches have some advantages and disadvantages with regard to discovering the more appropriate translation of an ontology element. With such a wide range of term translation approaches, it would be beneficial to have an effective strategy for combining these models into a localization system that carries many of the advantages of the individual techniques and suffers from few of their disadvantages.

From a technical point of view, the different translation models can be seen as the building blocks on which a ontology localization solution is built. In particular, the following aspects of building a working localization system are considered in this section:

- organizing the composition of various translation algorithms (section-5.8.1).
- combining the results of the basic translation algorithms to discover the more appropriate translations for each ontology element (section-5.8.2).

As the different translation methods focus on the same objective there are several dependencies between them. Nevertheless, certain combinations

are not reasonable: basically because the output of a translation algorithm in some cases can not be used as input for other methods.

5.8.1 Translation Composition

In this section we present at the strategic level, some natural ways to combine different translation algorithms. We choose to use multiple translation techniques to localize an ontology based on the following assumptions:

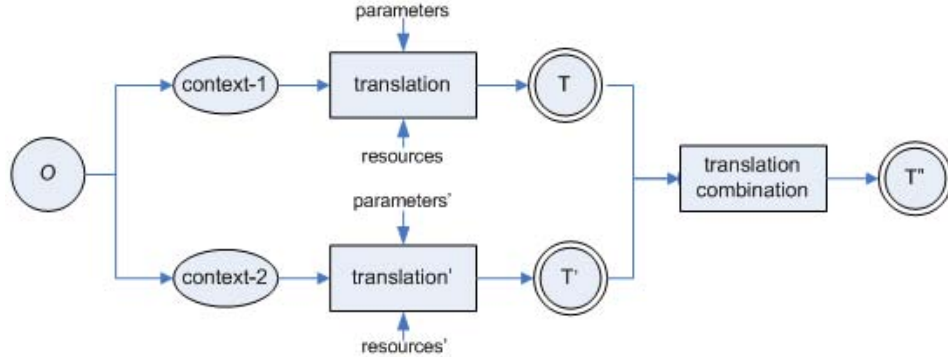
- Using more than one translation approach/resource gives us a wider range of word translation candidates to choose from, and the correct translation is more likely to appear in multiple translation resources than a single translation resource. Translating the ontology elements using multiple different translation resources gives us the possibility of minimizing very pronounced outliers.
- Using multiple translation approaches/resources gives us the possibility of maximizing the use of all available resources, allowing broad applicability to a range of operational settings. In fact, there are few resources that cover special terminology and fashionable terms. So, an ontology localization system should be ready to use whatever is available.

The translation composition proposed in this thesis is inspired on all empirical studies described in literature about multi-engine MT or MEMT architecture which operates by combining outputs from different translation engines. In order to classify the translation strategies, a distinction can be made as to whether translation paradigms are triggered in *parallel* or *sequential*.

Parallel composition

In a parallel strategy, each translation algorithm is fed with the source ontology element and generates an independent translation. The translations are then collected from their output and (manually or automatically) recombined. While there is an element of redundancy in such approaches given that more than one algorithm may produce the correct translation [Way, 2001], one might also treat the various outputs as comparative evidence in favor of the best overall translation.

The parallel combination of translation algorithms to localize an ontology is illustrated in Figure 5.3. In the figure, the ovals represent the context extracted from lexicon and core ontology, and the translation results are represented as concentric circles. Notice that the term context used for disambiguate the candidate translations depends on the techniques used to obtain the translations.



Sequential composition

In this approach, two or more translation algorithms are triggered on different sections of the same source ontology element. The output of the different techniques is then concatenated without the need for further processing. For instance, one would like to first use a dictionary based translation (section 5.4) to discover candidate translations, before running one translation based on corpus (section 5.5) or ontologies (section 5.6) to select final translations. The reasoning behind this approach is that if one knows the properties of the translation algorithms involved, reliable translations can be produced by using fewer resources than in a parallel approach. Integration of knowledge-based techniques with corpus-based techniques is a common strategy in commercial translation.

The sequential combination of translation algorithms for localizing an ontology is illustrated in Figure 5.4. Note that this sequential process can be used to eliminate the need of multilingual resources in the final stages of the localization process. Thus, in this setting, the final translation decision benefits from the candidate translations obtained by the first algorithms. Indeed, the second translation algorithm (*translation'*) can use only a monolingual resource to select the more appropriate translations. As in the parallel composition of translation algorithms, the term context (*context*) extracted from the ontology depends on the technique used in each step of the sequential translation process.

5.8.2 Translation Combination

When several translation algorithms or resources are combined, a crucial problem is to choose a translation among multiple translations produced for each algorithm. Note that this problem must be considered even though we combined algorithms that follow the same translation process. Translation algorithms adopting the same paradigm usually produce different transla-

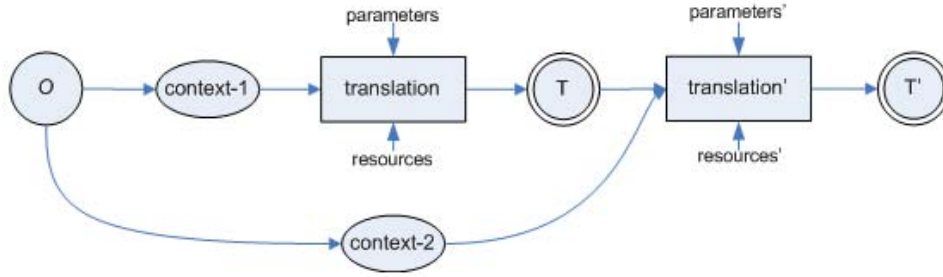


Figure 5.4: Sequential composition of translation algorithms.

tions for the same input, due to their differences in training data, or preprocessing strategies. Therefore the question we want to address in this section is, how do we go about choosing among translation algorithm outputs so that we end up with the best one?

Traditionally, translation combination has been conducted in two ways: *black-box combination* and *glass-box combination* [Huang and Papineni, 2007]. To choose a specific translation, the black-box combination method basically uses the external information of each translation approach. The information can be extracted from the general confidence that we have in the method, the input text to be translated, or the output produced by the translation method. This approach can be particularly useful when it is not possible to have access to internal features of the translation approaches (e.g., online MT systems).

In the glass-box combination, each translation algorithm provides detailed decoding information, such as translation model score, phrase and word probabilities, segmentation lattices³¹, or alternative translations per source word [Huang and Papineni, 2007]. This information is used to recombine the best parts from multiple candidate translations into a new utterance that will be better than the best of the given candidates. The main advantage of these approaches is that a possibly new translation can be generated that includes “good” partial translations from each of the involved algorithms.

Some of the well-known combination methods used in MT, such as linear combination, hypothesis selection, noisy channel models, confusion networks, and lattice combination can be used in ontology localization. The first two methods use a black-box combination, while the other approaches use a glass-box combination. A description of all the combination methods used in the field of MT is out of the scope of this thesis. However, for more details, cf. Nie et al. [Nie et al., 2001]; Callison-Burch and Flounoy [Callison-Burch and Flounoy, 2001], Nomoto [Nomoto, 2004], Paul et al. [Paul et al., 2005]; Brown et al. [Brown et al., 1990]; and Park [Park, 2001], Matusov et al. [Ma-

³¹alternative ways of breaking the input to an MT system into words

tusov et al., 2006], Sim et al. [Sim et al., 2007], Rosti et al. [Rosti et al., 2007b, Rosti et al., 2007a, Rosti et al., 2008].

5.9 Classification Guidelines for Ontology Localization Approaches

There are many high level factors that can be used to classify the approaches used to localize an ontology into different natural languages. In this section we provide some classification guidelines for localization approaches. The classifications discussed above provide a common conceptual basis for organizing them, and, hence, can be used for comparing (analytically) different existing ontology localization systems as well as for designing new ones. We now explain the four main factors: type of localization, localization process, output, and use case.

5.9.1 Type of Localization.

Localization approaches may vary in the type of localization chosen for ontology. We have identified two dimensions involved which determine the type of localization that needs to be performed:

- *Level of Localization.* According to the level of localization, approaches can be classified by: *linguistic* and *cultural*. In the linguistic level, ontology localization will affect only the lexical layer of the ontology. In some cases the conceptualization can be maintained and only the lexicon of the ontology (in particular the labels of it) is translated. The cultural level includes the ontology adaptation to a particular culture. When adopting the ontology to a new culture, this will typically lead to changes at the three levels (language of the labels, terminology and conceptualization).
- *Localization Purpose.* Depending on localization purpose, localization approaches can be categorized by: *instrumental* and *documental* [Montiel-Ponsoda, 2011a]. In the first case, the goal of the target ontology can be to have the same function in the target community as the original ontology in the source ontology. The purpose of the localization can also be “to document” the ontology in another language to make it accessible to a community which speaks another language.

5.9.2 Localization Process.

The main factor is the process of localization. In fact, the processes of different ontology localization approaches may differ most, when a large number of parameters and methods exist.

5.9. CLASSIFICATION GUIDELINES FOR ONTOLOGY LOCALIZATION APPROACHES

- *Ontology Elements Supported.* A first dimension to be considered is the type of ontology elements that can be localized. Depending on the considered ontology elements, the algorithms of localization can be more or less complex. For example, the localization of ontology concept and relations is more complex than the localization of ontology instances, because a big part of the instances are represented by a name, and therefore should not be translated (e.g., a label containing “Michael Schumacher”).
- *Degree of Automation.* Ontology localization approaches may vary in the degree of automation, ranging from fully manual through automatic recommendations to fully automatic.
- *Resources Used.* The grade of consensus, coverage and precision of the used resources to discover the translations are three important factors to take into account in order to compare localization approaches. These three parameters can be used to estimate in some degree the quality of different localization approaches. In fact, we believe that the heterogeneity of the used resources can be the principal reason why different approaches cannot be directly compared.
- *Ranking Method Used.* Another important factor to consider is the ranking mechanism used to pick a correct translation among multiple candidates. Depending on the degree of automation we may classify the translation ranking methods on supervised and unsupervised. The supervised ranking methods will require intensive human manual labor to pick the most appropriate translation, while that in the unsupervised ranking methods the choice of the final translation will be an automatic process.

5.9.3 Output.

The output may differ considerably for different approaches.

- *Degrees of equivalency.* Apart from the information that localization algorithms exploit and how they manipulate different tools and resources, an other important class of dimensions concerns the form of the result these systems produce. The kind of equivalence between the ontological terms and its translations might be of importance. For example, ISO 5964 [ISO, 1985] defines a classification scheme for different types of equivalence between terms: exact equivalence, partial equivalence, single-to-multiple equivalence, inexact equivalence and non-equivalence. The simplest cases are one-to-one translations. However, in real world, one will often encounter n-to-m translations instead.

- *Confidence.* Another significant distinction in the output results, concerns the confidence measures of the translations. Only recently have researchers started to investigate confidence measures for machine translation [Ueffing et al., 2003, Gandrabur and Foster, 2003, Blatz et al., 2004, Quirk, 2004]. Possible applications of the confidence measures include: i) post-editing, where words with low confidence could be marked as potential errors, ii) improving translation prediction accuracy, iii) combining output from different machine translation systems: hypotheses with low confidence can be discarded before selecting one of the system translations [Akiba et al., 2004], or the word confidence scores can be used for generating new hypotheses from the output of different systems [Jayaraman and Lavie, 2005], or the confidence value can be employed for re-ranking [Blatz et al., 2004]. We consider the confidence measure a factor essential for any ontology localization system.
- *Provenance.* The knowledge of provenance and its effects on the localization activity is another important factor to consider. Although there are no conclusive studies on whether provenance information about translation suggestions that combine different techniques has an impact on quality and speed of revision [Teixeira, 2011], we believe that this information should be taken into account when analyzing and comparing the results of different ontology localization systems. Different dimensions could be proposed taking into account the levels of provenance information that a system could provide to stakeholders, for example, the resources or algorithms used.

5.9.4 Use case.

The ontology localization activity can contribute as a plausible solution to different applications. For example, a typical case of the ontology localization activity is the multilingual ontology matching (MOM) application. MOM refers to the process of establishing relationships among ontological resources from two or more independent ontologies where each ontology is labeled in a different natural language [Fu et al., 2009b, Trojahn et al., 2008]. This activity requires support of ontology localization because MOM is achieved by first localizing the labels of a source ontology into the target natural language. Then by applying monolingual ontology matching techniques to the translated source ontology and the target ontology it is possible to establish matching relationships.

We believe that even though the cases might not directly be reflected in input, process, or output, they definitely influence the complete setting of the localization process. Therefore, the cases must be considered a factor for distinction.

5.10 Summary of the Chapter

Ontology localization has different facets; one of these facets is the translation. To automatize the translation task, a variety of techniques can be used. The classifications discussed in this chapter provide a common conceptual basis to analyze the advantages and shortcomings of each technique with regard to the localization activity.

We have provided such classifications based on a way of modeling the context used for the translation on one side and the kind of technology used to localize an ontology into different natural languages on an other. Once the different translation techniques have been identified, we have presented the strategic issues involved in creating localization solutions. In particular, this involves the composition of basic translation techniques and the combinations of their results.

We have finished this chapter describing some high level factors that can be used to classify the approaches used to localize an ontology into different natural languages.

Chapter 6

Lyfe-Cycle Model and Architecture

In this chapter we discuss two important issues related to ontology localization activity: life-cycle and system architecture. As we discussed in the introduction chapter, a typical localization project involves several tasks that extend far beyond the translation process itself. This is why the first goal of this chapter is to describe the life-cycle model by means of the representation of the major components of this activity and their interrelationships in a graphical framework that can be easily understood and communicated. As second goal of this chapter, we outline our approach to the definition of a system architecture that supports the ontology localization activity. The proposed model comprises the system components, the externally visible properties of those components, the relationships (e.g., the behavior) between them, and provides a base from which localization systems can be developed.

First, we give an intuitive view of the whole localization activity, including the translation phase, which was extensively described in the previous chapters. Later in this chapter, we introduce some basic requirements for an ontology localization system. Then, we will propose a system architecture based on the ontology localization life-cycle model, considering also the system requirements identified from different works in related areas. After defining the architecture, we will see the main modules needed to allow such an ontology localization approach in distributed and collaborative environments. Finally, we describe general comments and different technical details related to the LabelTranslator system, our approach to perform an automated localization in distributed and collaborative environments.

6.1 Ontology Localization Life-Cycle

Localization is a very effort intensive activity and requires a systematic approach covering the entire life-cycle of the localized product [Mudur and Sharma, 2002]. However, based on our investigation of existing academic projects and commercial systems [Esselink, 2000, Müüller, 2009, Jevsikova, 2009], we have identified that the current R&D efforts on localization (especially in the software area) suffer from the lack of a comprehensive life cycle model. We consider that the ontology localization is not a once-in-a-lifetime activity. It should be viewed as a continuous, iterative activity in which the localization outcomes of the current and past localizations can and should affect the future choice of localization policies and strategies and, thus, the behavior of an automated localization system. A comprehensive localization life cycle model is needed to clearly define the different phases of a localization process and to show:

- what information and knowledge should be specified or defined at different phases, and,
- how the results of the ontology element translations provide the feedback to other phases of the life cycle.

In this section we present an ontology localization model, which identifies the key concepts and elements needed to build an automated ontology localization system. One of the elements in the model is the *translation phase*, which in many analogous implemented software localization systems is not automated. In the previous chapters of this thesis, we study the key elements of the translation phase, with the dual aim of reducing the localization effort and identifying the steps to produce a general ontology localization model. In fact, the translation phase used to localize an ontology has been the core of our ontology localization life-cycle model.

6.1.1 The Automated Ontology Localization Model

The ontology localization life-cycle model is presented in Figure 6.1. This generic model depicts the major issues involved in the automating ontology localization activity. Our approach is inspired on different software life-cycle models [Sheu, 1997, Rajlich and Bennett, 2000, Ruparelia, 2010, Wright, 2011], which are used to illustrate the significant phases or activities of a software project from conception until retirement.

Although the order of steps presented in the model is logical, we believe that different ontology localization systems may use a different order, may group two or more steps into a single step or may not implement certain steps at all. The model is also independent of who actually performs the work. For example, if the ontology developer is using a distributed and

6.1. ONTOLOGY LOCALIZATION LIFE-CYCLE

collaborative team for localizing an ontology, many steps will be performed by the developer and others by the localization team. The value of the model is that it covers the major issues involved in this activity and provides a vocabulary to discuss these issues.

In the figure the main phases are represented by a rectangle, whereas the sub-phases are represented by ellipses. The thick line represents the main process flow; the secondary process flow is represented by a solid line. The data access is shown as a dotted line in the figure.

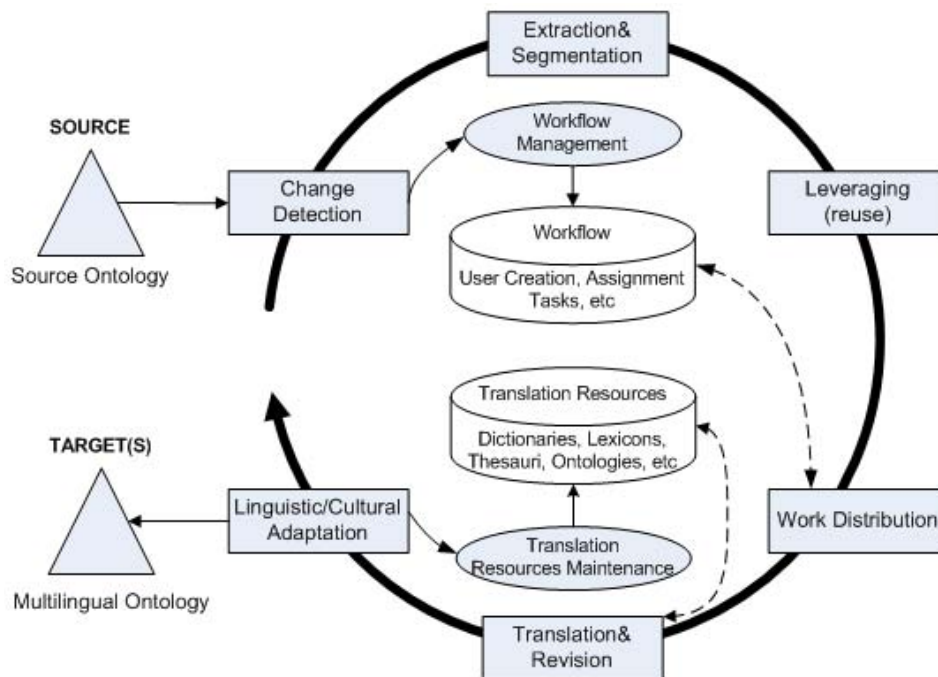


Figure 6.1: The Automated Ontology Localization Life-Cycle Model.

The proposed ontology localization life-cycle model is concerned mainly with managing the translation and localization of the ontology content into any number of target languages. In the following section we describe the main components involved in the model:

6.1.2 Automated Localization Cycle

The ontology localization cycle describes phases of the localization activity and the order in which those phases are executed. Each phase produces deliverables required by the next phase in the life cycle:

- *Change Detection*. This phase monitors the *source ontology* content and it is responsible for detecting changes and initiating actions. We believe that change monitoring may operate continuously or at regular

intervals. In addition, this phase starts the *Workflow Management* process, which is responsible for distributing the work to one or more translators/reviewers in one or more localizations.

- *Extraction.* Each ontology term requires its own extraction method from the ontology. The extraction method is responsible for extracting the ontology labels (representing any ontology element) and its context from the ontology.
- *Segmentation.* Once the label of the ontology element is extracted, it must be segmented into individual short phrases or multiword units¹ (MWU) in order to be translated appropriately.
- *Leveraging.* The Leveraging phase tries to translate all source labels using the translations stored in previous ontology localizations. It may use one or more translation memories to store the pre-translated ontology labels. This phase can be performed only when ontologies to be translated have a similar domain to ontologies previously translated.
- *Work Distribution.* Once the ontology localization activity has been initiated, the work must be distributed to one or more translators/reviewers in one or more localizations. This process is carried out by the *Workflow Management* process. The systems should provide some form of database which stores a list of translators and reviewers along with the language pairs they can handle. We consider that when the ontology to be localized is small and the target languages are known the own ontology editor may execute all tasks.
- *Translation.* In this phase, the translator actually translates the ontology labels received using the *localization resources* provided by the system or its own tools if the system can interface with them. This is likely the most important step since the main cost of localization is translation and the cost of translation is largely determined by the efficient of an environment provided to the translator. The translator may work online with a browser-based tool or offline on his desktop PC. However, the offline method requires some mechanism to update the realized work.
- *Review.* The translation work is then routed for reviewing (editing and proofing). The work is checked for translation accuracy and for overall term correctness. The system should allow any way of measuring the translation quality.

¹A multiword unit (MWU) is a connected collocation: a sequence of neighboring words “whose exact and unambiguous meaning or connotation cannot be derived from the meaning or connotation of its components” [Choueka, 1988].

- *Linguistic/Cultural Updating.* The goal of this task is to update the ontology with the linguistic information obtained for each ontology term in the target language. The result of this process is a *multilingual ontology*, which expresses the correspondences between entities belonging to the *source ontology* and the multilingual terms pertaining to a natural language. This phase may require only the adaptation of the ontology to a particular language or an ontology re-engineering process for transforming the conceptual model of an existing and implemented ontology into a new, more correct and more complete conceptual model which is re-implemented. It is at this time that the *localization resources* (translation memories, glossaries, etc) are updated and that *Localization Resource Maintenance* is best performed.

6.1.3 Data Structures.

All steps shown in the model revolve around two major data structures: *Workflow* and *Localization Resources*. The aim of the *Workflow* repository is to help manage, monitor and control the localization activity, while the *Localization Resources* repository helps to reduce the cost, increase the quality and increase the consistency of the translation work. They store the basic objects of the ontology localization activity: the participants, and the tools and resources, respectively. These objects require management and maintenance with the appropriate activities:

- *Workflow Management.* This activity refers to the process of defining and maintaining the workflow templates that specify which steps are to be processed by users or by the system, and the conditions under which they are processed. Some ontology localization systems will have wizards with only a few questions to answer, others will require several pages of options to be set, still others will have graphical interfaces that allow for a process to be defined as a flowchart.
- *Translation Resources Maintenance.* The more work that is routed through the ontology localization system, the more translation knowledge is accumulated, promoting more re-use. But as more and more data is accumulated, the system will also accumulate different translations for the same ontology elements. As translation knowledge grows, it becomes less precise and contains more “noise”. Therefore translation resources maintenance is required to avoid the chaotic growth of translation knowledge and ensure that the captured data can be leveraged in a meaningful way.

All steps above described are the base of our generic architecture for localizing ontologies and distributed and collaborative environments. The details of our approach will be described in section 6.3.

6.2 Key Requirements for an Ontology Localization Infrastructure

In this section we describe some desirable requirements for an ontology localization system, motivated by works in related areas and own experiences. To define infrastructure requirements, we took key factors as our starting point. We have been collecting factors from different software localization systems, comparing them with our own observations in the field of ontology localization, and grouping them according to their nature and relationship into three groups: i) collaboration and distribution of the tasks; ii) translation; and iii) extensibility. This grouping has given us a clearer idea of how to convert some of these identified factors into positive influences on ontology localization. This has inspired the definition of the infrastructure requirements. In the following sections we briefly describe the three groups of requirements.

6.2.1 Requirements for Collaborative and Distributed Localization Activity

In this section we present the most relevant requirements to support a distributed and collaborative ontology localization based on the analysis of the process (e.g., workflow) typically followed by organizations in the development and localization of ontologies. First, for our analysis, we considered existing processes for collaborative localization used in international institutions. As a case study, we focused on the collaborative localization process followed at FAO for localizing the AGROVOC Concept Server. Secondly, we observed how different software development paradigms and approaches deal with issues like cooperation among distributed team members. Finally, we discuss the main features identified as core requirements to support a collaborative and distributed ontology localization.

Collaborative Ontology Localization: AGROVOC Concept Server

One of the most important resources for covering the terminology of all subject fields in agriculture domain is the AGROVOC thesaurus (introduced in section 2.3.3), which evolved into a semantic system in order to provide ontology services. This newly reengineered system is called the “AGROVOC Concept Server (ACS)”.

The development of the ACS was based on: i) the need of making the development and maintenance of the AGROVOC thesaurus more collaborative and especially more direct for users without the intermediate actions of FAO staff, and ii) the idea to convert AGROVOC into a more complete structure allowing for the representation of more information (such as additional linguistic information, or the ability to have multiple translations for

6.2. KEY REQUIREMENTS FOR AN ONTOLOGY LOCALIZATION INFRASTRUCTURE

a specific term, etc.). This new infrastructure proposes a system, in which all actors interact collaboratively and concurrently.

The collaborative aspect and the number of people that eventually interact via the ACS calls for well defined and well managed workflows to avoid confusion, data inconsistency and assure quality control. Therefore, in the collaborative aspects of the creation and localization of an ontology we need to consider at least:

- collaboration over different steps performed by different people, and
- collaboration among several participants for every single step

Related Software Development Approaches

In addition to the case analyzed in the previous section, we have observed how different software development paradigms and approaches deal with issues like cooperation among distributed team members. We have focused our research on techniques from different domains that extensively rely on communication, collaboration and/or coordination techniques. Important influences to our proposal of requirements are:

- *Collaborative ontology development.* The collaborative development of ontologies within an organization usually follows a pre-defined process that specifies who (depending on the user role), when (depending on the ontology state) and how (what actions/operations) an ontology can change. To support this process some authors [Palma et al., 2011, Tudorache et al., 2008] propose the use of workflows to formalize the collaborative ontology development. This same idea could be applied in the collaborative localization of ontologies.
- *Distributed software development (DSD).* We have transposed major characteristics of DSD to the ontology localization context, exploring how they could be extended to localization of ontologies. We have considered reported issues related to global DSD in the wider context [Carmel and Agarwal, 2001, Prikladnicki et al., 2008]. Some issues that arise may apply to collaborative and distributed ontology localization. The following characteristics, already adapted to deal with process improvement have been preserved for our purposes: i) the process management is distributed by the Internet and ii) the process improvement is collaborative and decentralized.
- *Bug tracking tools.* As in software maintenance, it is possible to identify and deal with the weaknesses (translation errors) of a localized version, converting them into improvements to the next version. In this way, one can relate error-handling management to ontology localization. Both approaches follow a similar workflow including submission

(proposal), evaluation, and approval (or rejection). In the software testing context this error handling is being supported by bug-tracking tools. These tools could be customized to handle ontology translations.

- *Knowledge Management (KM) practices.* The development of software can benefit from many KM practices, and indeed several aspects of KM employed in software development have been studied. There are many tools to support KM practices (e.g., contribution, knowledge dissemination, and collaboration) that can be useful to ontology localization. We are particularly interested in how to promote collaboration and improve participation and as such benefiting from different skills. Building networks and “knowledge communities” powered by accumulated translation knowledge can be a good strategy to facilitate the localization of ontologies.

Summary of the Main Features

In this section we discuss the main features of both the FAO localization workflow and the distributed software development paradigms. The goal of this discussion is to identify the core requirements to support a collaborative and distributed ontology localization activity. Five main requirements have been identified:

- *Flexible workflow support.* The main common thread for the process that we described in the FAO case is that many steps in these workflows require human actions. Human-centred workflows are different from service workflows that combine software services for automatic execution. For our purposes we consider that a combination of these workflow approaches is a good alternative. We envisage a service workflow that enforces and automatically executes critical ontology localization steps such as ontology submission, change detection, e-mail notification of localization tasks and events, and real-time tracking and reporting of individual localization works. Localization activities such as the selection of ontology labels or reviewing of translations may be controlled by a human-workflow.
- *User management and provenance of information.* With multiple users contributing to the localization of an ontology, it is critical for users to understand where information is coming from. Thus, users must be able to see how localization participants reach consensus on ontology label translations, who can perform translations, who can comment on them, when ontology label translations become public and so on. Any ontology localization system must include these features.
- *Centralized control.* A centralized view on all localization projects should be provided by all ontology localization systems, giving local-

6.2. KEY REQUIREMENTS FOR AN ONTOLOGY LOCALIZATION INFRASTRUCTURE

ization manager's easy access to critical information, such as assigned ontologies and workflow tasks, required to manage and coordinate the ontology localization activity effectively.

- *Collaborative localization support.* As we have described in the FAO case, the localization activity is a collaborative effort, where the number of users participating in localization ranges from a handful to a couple of dozens. With larger groups of users contributing to ontology localization, we believe that it is necessary to define appropriate workflows, strategies and an infrastructure to support the process that coordinates the collaborative ontology localization within an organizational setting.
- *Distributed localization support.* The most obvious element of complexity in the cases that we described in the previous section is the number of parties involved and their geographic distribution (see the typical localization scenario in Figure 6.2). The geographic distribution of the parties is motivated by the fact that localization in most cases requires in-country reviews to check the content of the translation.

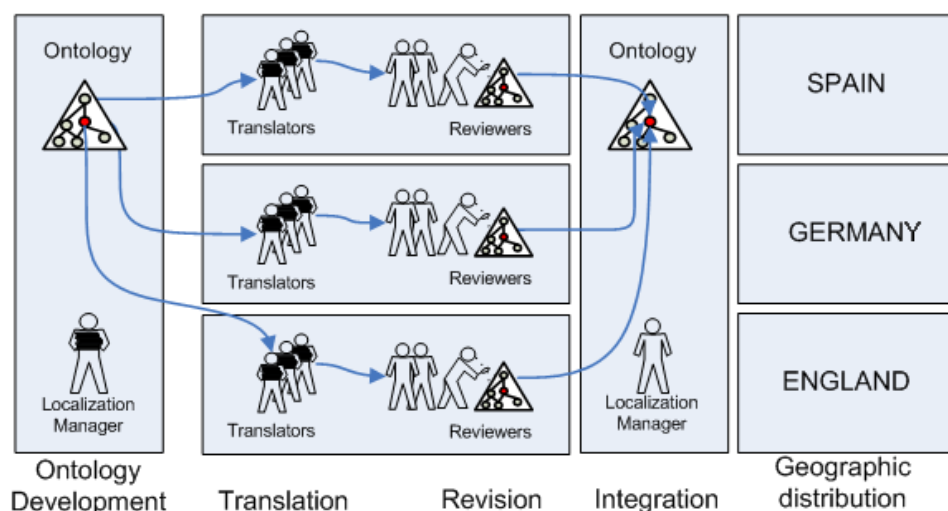


Figure 6.2: Typical Ontology Localization Scenario.

Managing a large number of parties presents a lot of challenges. However, in this thesis we focus only on two challenges found in the FAO case and the software development approaches: *communication* and *version tracking*.

In order to reach consensus between stake-holders on a specific action to be performed, an option is to use e-mail as a communication mechanism. Because it is simple and readily available, e-mail is usually

the primary communicating tool, which includes handing off projects and files that contain versions of documents, as well as tracking status [Duhl, 2008]. Nevertheless, the use of e-mail frequently results in confusion over which version of which document is current, the status of each document, and who is currently doing what, all of which leads to inefficiency and waste of time [Duhl, 2008]. We agree with this appreciation, in fact, we believe that the ontology localization activity must support a efficient mechanism for managing versions of localized ontologies (as in the FAO case), controlling ontology access through some form of check in/check out and file locking, and enabling remote or distributed access.

The second main source of complexity is due to the high number of documents that need to be distributed across project participants. This situation is even further complicated when changes are made to the original ontology elements during the course of the localization activity. These late modifications need to be introduced into the translation chain and lead to the existence of multiple versions of the same set of files, which in turn leads to frequent errors and substantial management overhead. In the case of FAO, a unique and homogenized maintenance tool for the collaborative management of AGROVOC is employed. We consider that a similar approach that uses a workflow model integrated with a content management component for providing greatest flexibility and power can be used.

Along with the features above described, the requirements introduced in the next sections are considered in our approach to support the automated localization of ontologies in distributed and collaborative settings. These lists are not exhaustive, but they provide a basis for fostering further progress in the building of ontology localization systems.

6.2.2 Requirements to Support Automatic Ontology Translation

The major requirements identified are the following:

- *To cope with specificities of translation.* A key issue for any localization system is to handle some traditional problems in translation: some concepts from the source language have no equivalents in other languages, polysemous words and homographs, quasi-synonymy, etc.
- *To deal with transcription of domain terms.* Unlike the translation of textual material (technical manuals, etc.), translation of ontology elements must address the difficulties of unknown words (proper names, technical names, etc.), fragmented input (partial sentences), transcription errors (typographical errors, omissions), etc.

6.2. KEY REQUIREMENTS FOR AN ONTOLOGY LOCALIZATION INFRASTRUCTURE

- *To avoid the use of pre-editing.* Pre-editing is not feasible, because the ontology elements must be preserved exactly as they have been designed. Pre-editing is possible only when the ontology editor has authorized the editing.
- *To reduce the imprecisions in the translations.* All ontology elements should be translated accurately to avoid misinformation on the multi-lingual ontology.
- *Support for different natural languages.* Although most ontologies that have been built so far have their labels in English, an ontology localization system should allow for the translation between any pair of natural languages.
- *To rank the obtained translations.* Since the translation resources can provide multiple candidate translations, the localization system should provide a ranking method to order the translations according to the context of the term under consideration.
- *To reduce the user's intervention.* As there are many available resources that can be used to localize and to acquire knowledge of a domain, the localization system should automatically translate the labels, minimizing the requests to the user.

Notice that although some of the translation requirements are challenging for MT, ontology elements do have characteristics which make it amenable to automatic localization using MT techniques, as we describe in the section 4.1.2 .

6.2.3 Requirements for an Extensible Ontology Localization Infrastructure

It is a challenge to design an extensible platform for ontology localization. We consider that extensibility is a key factor for leading to success of these systems. From our point of view, the following items ensure even partially the extensibility of the localization systems:

- *Independence of ontology language.* The system has to support localization of ontologies described in different ontology languages e.g., OWL, RDFS or FLogic.
- *Independence of domain.* The system has to support ontology localization in any domain of knowledge.
- *Broad coverage.* The localization system should translate not only concept labels, but also attributes and relations.

- *Easy to extend to new languages.* The localization system should have the ability to include new languages in the system with a minimum level of effort required to implement the extensions. This means that no alterations of the knowledge representation in the ontology should be needed when the localization process is applied to a new language. This characteristic should be a key design criterion for any system.
- *Representation of multilingual information.* The system should provide different ways to store multilingual information within the ontology. One way would be the inclusion of multilingual labels in the ontology. Other options to represent the multilingual information could be the combination of the ontology with a mapping model that establishes links among the conceptualizations or the association of the ontology with an external linguistic model [Montiel-Ponsoda et al., 2011, McCrae et al., 2011b]. The appropriateness of each approach would be principally determined by the domain type of the ontology and the final function of the resulting ontology [Espinoza et al., 2009b].
- *Automated content extraction.* The system should automatically highlight localizable content within a ontology prior to routing to the translation process using prepackaged or custom-developed ontology label filters, and therefore reducing the time spent on ontology content preparation.

All requirements above described have to aim to optimize the performance of the ontology localization activity, however we consider that the accomplishment of all these requirements is a very complicated task.

6.3 Global Description of the Architecture

In this section we present a generic architecture for localizing ontologies in distributed and collaborative environments. The system architecture is based on the life cycle model proposed in the Section 6.1.

The philosophy used in the design of our general architecture was to create a system that supports a great part of the system requirements described in the previous section. *Translation*, *Collaboration*, and *Distribution* have been the key factors in the design since we think of an Ontology Localization System as a tool to reduce the cost of translation and minimise the time to localize an ontology. Also, to solve the quality problem in automatic localization systems we propose a collaborative and distributed approach to: i) enhancing the communication and collaboration among localization stakeholders; and ii) increasing ontology user's participation in improving the ontology localization process.

6.3. GLOBAL DESCRIPTION OF THE ARCHITECTURE

Our approach relies on three different modules, namely: *Ontology Management*, *Localization Management* and *Ontology Translation*. These modules also solve the difficulties faced by the three primary groups of people involved in ontology localization: ontology editors and managers, localization managers, and individual translators/reviewers, groups, and communities responsible for localizing ontologies. A more detailed description of these localization roles will be introduced in Chapter 7. The high-level modules of the architecture are illustrated in Figure 6.3 and their features will be explained along this section.

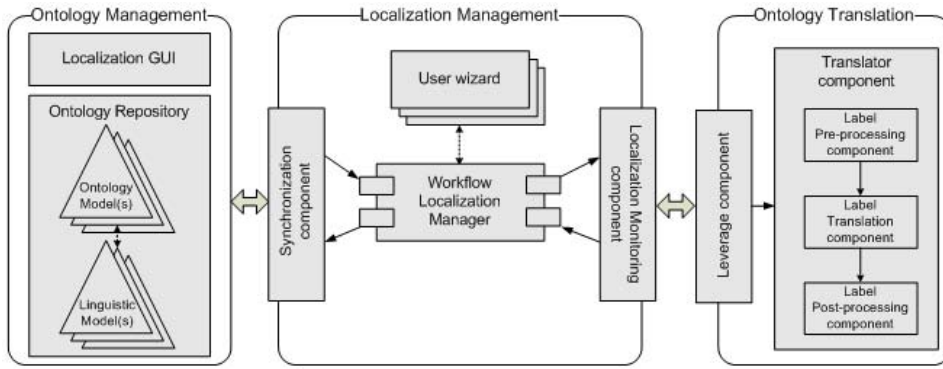


Figure 6.3: General architecture to support Ontology Localization.

- The *Ontology Management* is the module which enables ontology editors and managers to automatically and securely access, as well as manage multilingual content for localization. The *Ontology Management* also provides other functionalities such as: i) to create and update ontologies and instances, ii) to allow for the selection of ontology terms that need to be localized, iii) to handle different versions of the ontologies, and iv) to manage the access and the visualization of the translated labels with the aim of ensuring that the labels of the linguistic model are a mirror of the terms in the ontology.
- The *Localization Management* is the module designed to help manage, monitor and control the localization activity. It is the key for the interoperation between the different ontology localization stakeholders. The *Localization Manager* uses workflow technology to automatically detect new or modified source ontology content, to automatically route the ontology elements to the localization stakeholders, to support the different localization tasks, and to automatically deliver the translated ontology elements back to the *Ontology Management*. Also, the *Localization Manager* detects the changes in the ontology model and then automatically propagates those changes to the linguistic model using the *Synchronization component*. Finally, the *Localization component*

is responsible for controlling and managing the tasks that the ontology stakeholders are allowed to perform depending on their roles and the status of the ontology elements to be localized.

- The *Ontology Translation* is the core of the whole system. It allows for the automatical discovery of the most appropriate translations of each ontology element. The *Translation Leverage component* tries to translate a source label using the translations stored in previous ontology localizations. In the absence of previous translations available, the *Ontology Translator component* performs three steps: *label pre-processing*, *label translation* and *label post-processing* to provide the best possible translation of each input. To achieve more accurate translations, the *Translator component* has access to different linguistic and semantic translation resources.

The modules above described have been used and/or implemented on an ontology localization system, with the translation of ontologies between English, German and Spanish [Espinoza et al., 2008b, Espinoza et al., 2009a]. In the following section we explain these modules in detail and provide examples obtained from our system.

6.4 The Ontology Management Module

The Ontology Management Module is a module designed to help editors and managers to automatically and securely access and manage multilingual content for localization. The main uses of the Ontology Management module are: i) development and storing of the ontologies that need to be localized, ii) selection of the ontology elements to be translated into different natural languages, iii) handling of different versions of the localized ontologies, and iv) deploying of all multilingual information associated with ontology elements.

Analyzing the state-of-the-art ontology management tools², we observe that the evolution of semantic technologies has led to a number of concrete implementations to support specific ontology engineering activities and that in particular the first three functionalities used in the ontology localization activity are well supported. However, popular tools available today for ontology development are limited with respect to how to model multilinguality in ontologies. While typically today's ontology management infrastructures include multilingual data in the ontology meta-model, (see Section 1.4.3 for more details), we require an environment that adequately supports the inclusion of as much linguistic information as wished, as well as the possibility

²Readers interested in the state-of-the-art of ontology management tools can refer to [Martin et al., 2008].

6.4. THE ONTOLOGY MANAGEMENT MODULE

to establish links between the linguistic elements within one language or across languages.

The NeOnToolkit³ is an ontology management tool for engineering contextualized networked ontologies and semantic applications. With NeOnToolkit, we aim to start state of the art ontology localization by developing an ontology localization system. Particularly, we aim at improving the integration of multilinguality in ontologies, using a repository which keeps ontology knowledge and linguistic (multilingual) knowledge separate and independent.

The *Ontology Repository* (OR) is the critical component which supports the association of the ontological model(s) (sources ontologies to be localized) with a multilingual linguistic model. Thus, the ontology repository relies on the combination of two independent modules, the ontological and the linguistic one. In our system, the linguistic information needed to build a multilingual ontology is generated automatically by the *Ontology Translator* module, which will be explained later on.

The rationale underlying OR is not to design a lexicon for different natural languages and then establish links to ontology concepts, but to associate multilingual linguistic knowledge to the conceptual knowledge represented by the ontology. What the Linguistic Repository (LR) does is to associate word senses as defined by Hirst [Hirst, 2003]- in different languages to ontology concepts, although word senses and concepts can not be assumed to overlap. The LR goes along the line of what Pustejovsky [Pustejovsky, 1991] defined as Sense Enumeration Lexicon, in which a unique sense is associated with a word string. It enhances the scalability of the ontology localization approach by avoiding the need for investing time and energy in the development of a multilingual ontology for each target language.

The linguistic information stored in the OR is initialized when new ontologies join the Ontology Management module. Also, for each ontology element localized, the OR automatically establishes a link with the ontology term under consideration.

To support the translation resources maintenance phase identified in the localization life-cycle model (see section 6.1), we believe that the multilingual ontology, the result of this activity, should be automatically/manually added to the list of resources managed by the Ontology Translation module. The incorporation of this new resource will ensure that the translated data can be leveraged in a meaningful way.

³The NeOnToolkit is the heart of the infrastructure of the NeOn project. NeOn is an large European Research project developing an infrastructure and tool for large-scale semantic applications in distributed organizations.

6.5 The Localization Management Module

This section describes in detail the goal and functionalities of the Localization Management Module which is the key module for localization managers to monitor, manage and control the localization activity. The Workflow Localization Manager (WLM) is the core component of this module. The main goals of the WLM are:

- To manage the timely flow of the localization activity from initiation to delivery,
- To detect changes in the ontology model and propagate those changes to the linguistic model, and
- To manage the individual localization task performed in the Ontology Translator module.

In order to support the first goal the WLM includes a *collaborative workflow*, which implements the necessary mechanisms to allow the ontology stakeholders to perform the activities of the ontology localization life-cycle. Thus, the collaborative workflow is responsible for the coordination of who (depending on the user's role) can do what (i.e. what kind of actions) and when (depending on the status of the ontology elements).

From a technical point of view, the collaborative workflow is associated with a set of initialization parameters (e.g., user roles, assigned tasks, etc), source and target languages, and a partially ordered set of activities or states. The WLM individually stores the initialization parameters of each ontology. However, the information about user, roles and skills are stored in a shared database, which have two benefits:

1. *Improved Project Staffing.* The Localization Managers can see all the information related with a participant (e.g., language skills). This saves time and allows for better decisions when staffing a new ontology localization project.
2. *Shared Information Across Ontology Projects.* The shared information provides a particular benefit to ontology projects that need to localize several ontologies. Maintaining a single user database allows to share users in different ontology projects.

Coming back to the description of the workflow, the activities supported are: selecting the ontology elements to be translated, translating the selected elements, reviewing the translations, and updating the ontology with the linguistic information obtained. These activities summarize the localization tasks commonly followed by different organizations (see section 6.2.1 for more details). In the following we summarize the main steps in the workflow process to localize an ontology (see Figure 6.4):

6.5. THE LOCALIZATION MANAGEMENT MODULE

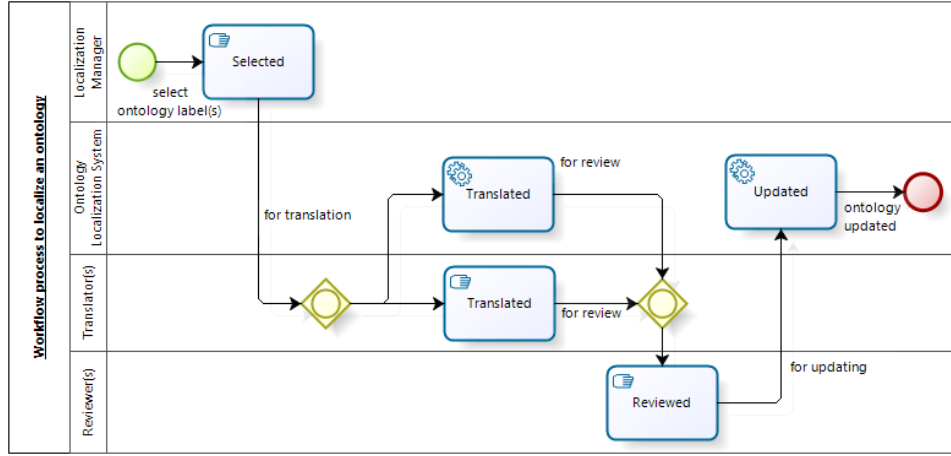


Figure 6.4: Workflow process used to localize an ontology.

- An ontology is passed to the Localization Manager for localization.
- The Localization Manager manually selects the ontology labels to be localized and sends the selected labels for translation.
- A translator downloads the selected labels to be localized and (s)he performs the translations using an automated localization tool (as proposed in this thesis) or an intensive manual process.
- Once translation activities have been accomplished, the translators upload the translated ontology labels and send them for review.
- The reviewers download the translated labels and check for possible errors.
- Finally, the Localization System updates all linguistic information of each localized label.

In the next sections we explain the rest of the associated components.

6.5.1 Synchronization Component

The *Synchronization component* supports the second goal of the workflow localization manager which is detecting the changes in the ontology management module. This component listens the changes in the ontology model and then automatically propagates those changes to the linguistic model using synchronization techniques. Remember that our system follows the current trend in the integration of multilinguality in ontologies, which suggests the suitability of keeping ontology knowledge and linguistic (multilingual)

knowledge separated and independent [Montiel-Ponsoda et al., 2008, Buiteelaar et al., 2009]

In order to keep both models, the ontology model (OM) and lexical model (LM), synchronized we first need to find out exactly what has been changed in the ontology model, then find the equivalent places in the linguistic model and only then start the updating. In a previous work [Espinoza et al., 2009a] we introduced the module for managing the conceptual knowledge and the linguistic knowledge by means of synchronization techniques. Hence, we briefly highlight the main features.

Addition of new terms in the ontology, or deletion of an existing term can be controlled by some mechanism of change tracking. Change tracking in our approach enables the system to obtain only changes that have been made to the ontology terms, along with the information about those changes. By adopting this feature, our system can accurately identify the minimum set of changes needed to adjust the structure of the linguistic model, a critical first step to ensure that a change is made in the localized ontology. To correctly update the linguistic model, the system needs to identify:

1. all ontology terms in the original ontology whose labels have changed in the updated ontology,
2. any ontology term that has been added to the updated ontology,
3. any ontology term which has been removed from the original ontology, and
4. any ontology term whose position in the updated ontology differs from that in the original ontology.

Finding where a translation is required is only part of the problem. We also need to ensure that changes in the ontology structure are accurately propagated to the linguistic information. This requires that elements whose structure need to be updated are clearly flagged in the linguistic model, and that the relevant structural changes are indicated in a form that turns the updating of the translation into a simple process, thus involving minimum work on the part of the linguist user or domain expert. Figure 6.5 illustrates the process used in our system to synchronize the conceptual and linguistic information. In the following we analyze the process in more detail, describing the actions performed by each actor of our scenario.

- *Ontology expert.* (S)he is responsible for editing the changes in the ontology model. All the changes executed in each user session are stored in a repository as a new version. The types of changes that our system can manage are the following: changes of the label content (e.g., ontology label rename) and ontology structure changes (e.g.,

6.5. THE LOCALIZATION MANAGEMENT MODULE

delete or add operations). For each case, the system stores the type of operation executed and its additional information (e.g., the name of the renamed label). This information is used in our system to synchronize the conceptual and linguistic information.

- *Linguist expert(s)*. The linguist expert in a specific target language is responsible for performing the localization process. Notice that this process always uses the last version of an ontology. When the linguist needs to update the linguistic model (LM), our system tries to synchronize both models, performing the following actions: (1) obtaining the current version of the LM to be updated, (2) extracting the last version of the changes in the ontology model (OM) from which the last localization was taken (normally the one with the same number as the LM), (3) performing all the actions of the file of changes in the LM, and (4) updating the LM version in the repository.

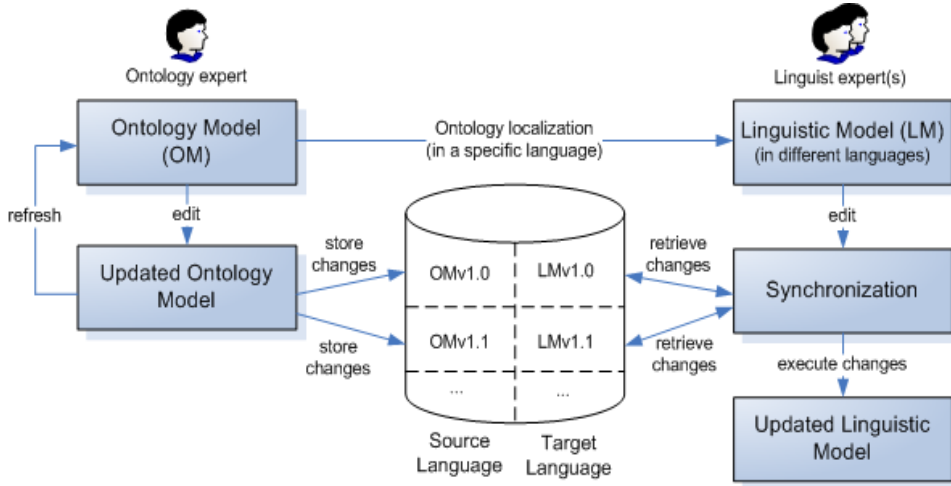


Figure 6.5: Synchronization of ontology and linguistic model.

6.5.2 Localization Component

The last goal of the workflow localization manager is supported by the *Localization component*. This component controls and manages the tasks that the ontology stakeholders are allowed to perform depending on their roles and the status of the ontology elements to be localized. The possible tasks in the collaborative workflow (described previously) apply to different abstraction levels. In our solution we consider two levels: ontology element level and translation level. Although the workflows can be used independently of the underlying ontology model, the specific set of ontology terms depends on the ontology model. In our approach, we are mainly considering the OWL

ontology model, in which an OWL ontology consists of a set of axioms and facts. Concepts, properties, instances and ontology term comments are the set of ontology elements we are taking into account.

The possible states that can be assigned to ontology elements are:

- *In Use*: This is the status assigned to any element when it first passes into the collaborative workflow, or when it was localized and then updated in the Ontology Repository.
- *New*: If the ontology element was added to the ontology after the ontology has been localized, the ontology element is passed to the “New status, and remains there until the element is localized.
- *Changed*: If the original label of the ontology element has changed, then the element is passed to the “Changed” status, and remains there until the element is checked to be localized again.
- *Unused*: If the ontology element has been deleted, then this element is passed to the “Unused” status, and remains there until the ontology is synchronized (see synchronization component in the previous section).

The localization component controls also the status of the translations. Figure 6.6 shows the workflow to translation level. States are denoted by rectangles and actions by arrows. The actors in the figure specify the actions that an ontology stakeholder can perform depending on its role. In the following we provide a detailed explanation:

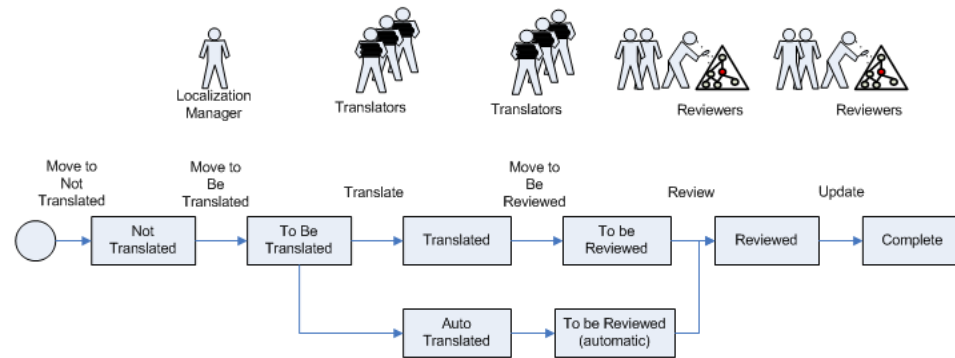


Figure 6.6: Workflow to the Translation level.

- *Not translated*: This is the status assigned to any translation when it first passes into the collaborative workflow or when any change has been performed in the element of the ontology under consideration.

- *To be Translated:* Once the Localization Manager selects the translations with the “Not translated” status, these translations are passed to the “To be Translated” status, and remain there until a “Translator” translates them.
- *Auto translated:* If a “Translator” uses the automatic translation algorithm provided by the system, then the translation is passed to the “Auto-translated” status, and remains there until the own “Translator” sends the translations to the “To Be Reviewed (auto)” status.
- *Translated:* If a “Translator” manually makes a translation, then the translation is passed to the “Translated” status, and remains there until the own “Translator” sends the translations to the “To be Reviewed” status.
- *To be Reviewed:* If a “Reviewer” approves the translations send by the translator, it passes to the “Complete” status. The reviewer knows in advance if translations have been made automatically or manually. For example, the word “automatic” in the message “To Be Reviewed (automatic)” indicates to the reviewer that these translations have been obtained automatically. Additionally, when the translations reach the “Complete” status, they are automatically updated in the Ontology Repository.

Note that during the collaborative workflow, actions are performed either implicitly or explicitly. For instance, when a user updates (i.e. modifies) an ontology label, he does not explicitly perform an update action. In this case the action has to be captured from the user interface and recorded when the ontology is saved. In contrast, when Reviewers for example explicitly approve/reject proposed translations, the action is immediately recorded when performed.

6.6 The Ontology Translation Module

In this section we explain the ontology translation approach that was already briefly mentioned in Chapter 4. Also, in Chapter 5 we already described the *label translation component*, showing some natural ways to combine different translation algorithms to localize an ontology. Now we extend the explanation by introducing the additional steps given by the Ontology Translation to discover the more appropriate translations.

With the help of the ontology translator module, the translators/reviewers can reduce the effort to manually localize an ontology. For each ontological label, the translation module first uses the *Leverage component* to try to discover pre-translated ontology labels. If no results are obtained, then, the *Translator component* performs three pipeline steps to translate a label

described in a source natural language and obtain the most probable translation of this ontological label in a target natural language. Figure 6.7 shows the components presented in Figure 6.3 in more detail.

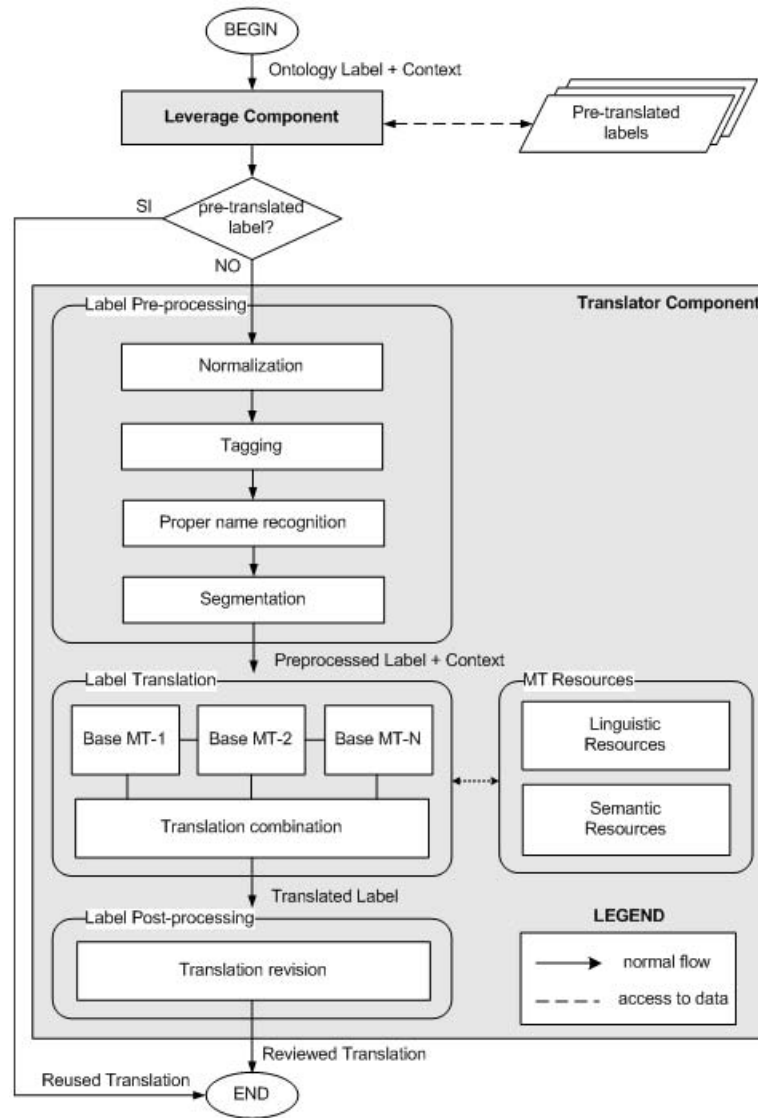


Figure 6.7: Detailed Ontology Translator in a Localization System.

6.6.1 Leverage Component

This component takes as input the ontology label and its context to discover past translations. In its simplest form, this component may rely on a cache to match ontology labels with pre-translated labels from previous ontology lo-

calizations. The software localization industry recommends the use of translation memories as translation reuse technology [Massion, 2005, Lagoudaki, 2006]. For our purposes, a translation memory (TM) system will transform inventories of past ontology translations into a database by automatically extracting and aligning the source labels with the target language labels. This involves the creation of a simple database of aligned words taken out of context. Since this happens without reference to context, TM technology will require a good deal of manual maintenance by a senior linguist to validate and correct misalignments, especially 1:n and n:1 combinations that are readily apparent to human, but not to automated tools [Kuhns, 2007]. A more serious shortcoming of TM systems is the fact that they have no access to the meaning of the translated text and operate on its surface form. As a result, they fail to match words/sentences that have the same meaning, but a different syntactic structure [Kuhns, 2007].

To overcome these shortcomings, a new generation of TM systems has been proposed, which analyse the segments not only in terms of syntax but also in terms of semantics [Gotti et al., 2005, Pekar and Mitkov, 2007, Mitkov and Corpas, 2008]. Some of these works rely on lexical resources such as WordNet to automatically identify synonyms, and therefore, to make a match between synonymous expressions possible. Due to the rather restricted availability of semantic data in relevant subject areas, the relevance of these approaches for commercial implementations is still rather small [Reinke, 2013]. However, we believe that this type of approaches represents a promising way forward to ensure that translators have a wider range of matches in the ontology localization activity. We envision also that the use of context information in both the input term and the translation memory improves matching algorithms.

In our approach, for the time being we only use a cache to avoid translating the same label twice, while we wait until some of the features discussed above can be incorporated into the currently available TM.

In the following section we describe the first step of the translator component and briefly the second and third steps.

6.6.2 Translator Component

The *Translator component* relies on three steps: *label pre-processing*, *label translation* and *label post-processing* to discover appropriate translations according the lexical and semantic context of the original ontology label. The three steps identified above are executed if the *Leverage component* does not return any results. The output of this component is an automatically translated label and manually validated by an expert.

Label Pre-Processing

We consider that ontology label pre-processing is essential in an ontology localization system, in order to simplify the core translation processing and make it both quality and time effective. The ontology labels pose different challenges to MT, which can be attributed to two distinct characteristics:

- Ontology labels differ linguistically and stylistically from written language: phrases are shorter and in some cases poorly structured, also they can contain ungrammaticality expressions (e.g., `Service_Transport` instead of `Transport_Service`)
- The current “standard” for naming the ontology labels is to use a CamelCase⁴ approach. Therefore, we cannot rely on the initial uppercase letter to identify a phrase initial word nor to recognize proper names, since names cannot be identified by an initial capital.

These problematic factors are dealt with in a pre-processing pipeline that prepares the input for processing by a core MT technique. Thus, the task of the ontology label pre-processing pipeline is to make the input amenable to a linguistically-principled, domain independent treatment. This task is accomplished in two ways:

1. By normalizing the input, i.e. removing noise, reducing the input to standard typographical conventions, and also restructuring and simplifying it, whenever this can be done in a reliable, meaning-preserving way.
2. By annotating the input with linguistic information, whenever this can be reliably done with a shallow linguistic analysis, to reduce input ambiguity and make a full linguistic analysis more manageable.

In the following we describe the functionalities of the different tasks in more detail:

Normalization

The label normalization groups three components, which clean up and tokenize the input.

The *text-level normalization* phase performs operations at the string level (ontology term comments by example), such as removing extraneous text

⁴CamelCase (also spelled camel case, camel-case or medial capitals) is the practice of writing compound words or phrases in which the elements are joined without spaces, with each element’s initial letter capitalized within the compound, and the first letter is either upper or lower case as in “LaBelle”, “BackColor”, or “iPod”. The name comes from the uppercase “bumps” in the middle of the compound word, suggestive of the humps of a camel. The practice is known by many other names.

6.6. THE ONTOLOGY TRANSLATION MODULE

and punctuation (e.g., brackets, used to mark synonyms or usage context), or removing periods from abbreviations. E.g.,:

“A publication may have an I.S.B.N.”

↓

“A publication may have an International Standard Book Number”

The *tokenization* phase breaks an ontology label into words. The *token-level normalization* recognizes and annotates tokens belonging to special categories (times, numbers, etc.), expands contractions (e.g. AssistProfessor to AssistantProfessor), recognizes, and normalizes typographic errors (e.g., Profeser by Professor), and identifies compound words.

“British” “System” “Education”

↓

“British” “System Education”

Tagging

In the *tagging* phase a tagger system⁵ assigns parts of speech to tokens. Part of speech information is used by the subsequent pre-processing modules, and also in parsing, to prioritize the most likely lexical assignments of ambiguous items.

Proper name recognition

Proper names are ubiquitous in ontology labels, specially in instance terms. Their recognition is important for deciding what instances should be translated, with an annoying effect if any instance term is systematically mistranslated (e.g., a sport domain ontology where the golfer named Tiger Woods is an instance systematically referred to as “los bosques del tigre”, lit. “the woods of the tiger”).

Name recognition is harder in the ontology domain due to the fact that capitalization information is commonly used for naming all types of ontological terms (concepts, properties and instances), thus making unusable all methods that rely on capitalization as the main way to identify candidates. Of course, this problem is even larger when no capitalization information is given. For instance, an expression like “mark shields”, as a possible instance in the ontology, is problematic in the absence of capitalization, as both ‘mark’ and ‘shields’ are three-way ambiguous (proper name, common noun and verb). Our approach does not support the proper name recognition for the moment.

⁵A tagger system is a tool for annotating text with part-of-speech and lemma information.

Segmentation

Segmentation breaks an ontology label into one or more segments, which are passed separately to subsequent modules. For our purposes, the translation units that we identify are syntactic units, motivated by cross-linguistic considerations. Each unit is a constituent that can be translated independently. Its translation is insensitive to the context in which the unit occurs, and the order of the units is preserved in the translation.

One motivation for segmenting is that processing is faster: syntactic ambiguity is reduced, and backtracking from a module to a previous one does not involve re-processing an entire phrase, but only the segment that failed. A second motivation is robustness: a failure in one segment does not involve a failure in the entire phrase, and error-recovery can be limited only to a segment. Further motivations are provided by the problems of the conventional MT systems. These systems have serious difficulties in dealing with long sentences due to the grammar coverage, memory limitation and computational complexity. Without proper treatment of long phrases, the base MT systems, may fail to produce understandable translations. Although in our proposal we did not treat the translation of phrases (as found in term annotations), however we considered this component of utmost importance for future versions of the system.

In our approach we use a basic segmentation process to divide the tokens of a compound label. However, for the translation of phrases we devised a segmentation component based on machine learning techniques [Kim et al., 2001], syntactic analysis techniques [Kim and Kim, 1997] or support vector machines [Kim and Oh, 2008].

Label Translation

After preparing the ontology element for an effective translation processing, the Ontology Translator invokes the label translation component, which obtains the most probable translation for each ontology label (see section 4.4.2). This component integrates different translation methods, combining the output by means of different translation strategies. Some natural ways to compose different translation algorithms was presented in section 5.8.1. In addition, in section 5.8.2 we summarized some of the well-known combination methods used to integrate the output of different MT approaches. Each translation method relies on different linguistic and semantic resources to obtain candidate translations. The output of this component is a ranked set of translations for each ontology label.

Label Post-Processing

This component shows the translations to the user for review of their translation quality. The quality of the translations is measured by two factors

adequacy and fluency. Adequacy determines the quantity in that the meaning of a correct translation is preserved. Fluency determines how well the corresponding translation in the target language has been done.

The checking of the quality of a translation is the only task of the ontology localization activity in which the user necessarily needs to interact.

6.7 LabelTranslator System

In this section we first describe the general comments and different technical details related to the LabelTranslator system. LabelTranslator is our approach to automatically localize an ontology into different natural languages. Next, in the rest of this section we describe the details of the design and implementation of the main components of our system.

6.7.1 Technical Details of the LabelTranslator System

LabelTranslator [Espinoza et al., 2008a, Espinoza et al., 2008b, Espinoza et al., 2009a] has been designed to support ontology localization by automating the main components described in our generic architecture and with the aim of reducing human intervention. The tool has been implemented in the ontology editor NeOn Toolkit as a plug-in⁶.

A previous version of the system, which didn't include support for a collaborative and distributed process and only had a basic approach to the translation of ontology labels, was implemented as a first prototype. This version has been replaced by the system described in this thesis. Concretely, in the Localization Management module we design and implement support for the workflow. This new feature helps to manage, monitor and control all localization activity. Also, in the Ontology Translator module we incorporate a label pre-processing component in order to simplify the core translation processing and make it both quality and time effective. Finally, in this same module we improve the translation task, combining different translation strategies depending on the type of label to be translated. These implementation changes have significantly improved the quality of our system.

In the following we describe different features and technical information of the main components of the system from the point of view of implementation.

6.7.2 Ontology Management

The Ontology Management is the key of the multilingual information management of LabelTranslator as it allows for the definition, storage and re-

⁶<http://www.neon-toolkit.org/wiki/LabelTranslator>, last access on October 2012.

trieval of both the source ontology(ies), which provides the conceptual information to be localized and the linguistic model(s) adopted for organizing and relating linguistic information with each ontology in the system.

In our approach we have implemented the linguistic support in the NeOn toolkit, a state-of-the-art, open source multi-platform ontology engineering environment, which provides comprehensive support for the ontology engineering life-cycle. For our purposes, we have extended the architecture of the NeOn toolkit incorporating two components: the *localization GUI component* and the *repository component*.

Localization GUI component

This component provides additional extension-points to modify the main components of the NeOn ToolKit, with the aim of controlling the aspects related to the localization activity. For the different localization stakeholders, a whole set of interfaces has been developed, which interact with the different components of the system. In this section we will show some of the interfaces used in our tool. The following functionalities have been implemented:

- Configuration of the parameters used in the localization activity.
- Creation of a new ontology localization project.
- Assignment of the participants, roles, skills and tasks of the workflow.
- Management of multilingual labels.
- Visualization and editing of the linguistic information associated to each ontology.

All these functionalities are provided by different interfaces. The system starts with an ontology (described) in OWL-DL or RDF which is provided by the user. LabelTranslator uses some views⁷ of the Neon ToolKit to load the ontology and store the multilingual results. In Figure 6.8, we show a screenshot of the Ontology Navigator view, which contains all created/imported ontology projects. Each ontology project can contain one or more ontologies to be localized.

Repository component

This component captures all the linguistic information associated with the localized ontology elements. LabelTranslator supports the linguistic information repository model [Montiel-Ponsoda et al., 2011] (LIR) designed for

⁷In the NeOn ToolKit a view is typically used to navigate a hierarchy of information, open an editor, or display properties for the active editor.

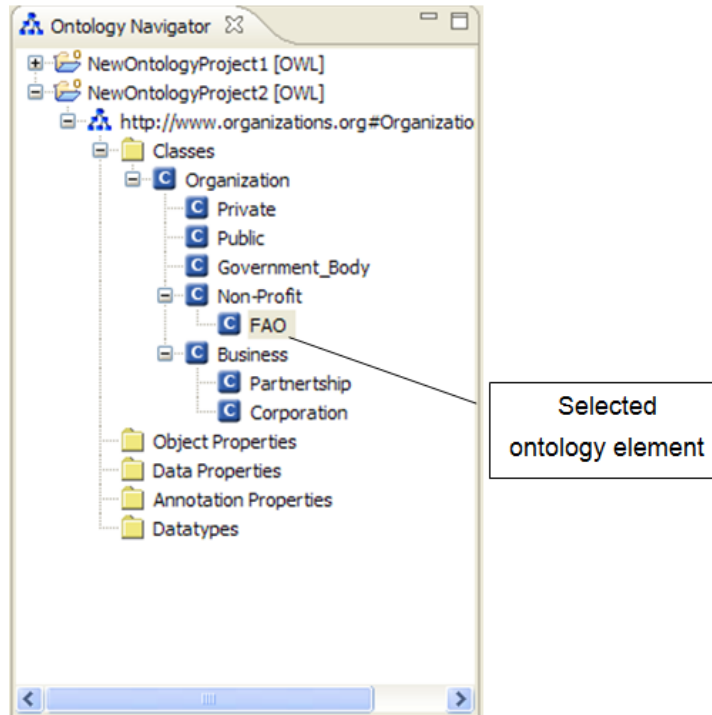


Figure 6.8: Ontology Navigator with a selected ontology element.

the representation of multilingual information in ontologies. The LIR model is a structured, non-exhaustive set of linguistic and terminological data categories, built up on the basis of existing standards. The inclusion of the LIR in the system ensures separation of information that is considered orthogonal in nature; we refer to the ontological and linguistic information.

Figure 6.9 shows the association between the OWL meta-model and the LIR. This association is established by the `hasLexicalEntry` relation between `OntologyElement` and `LexicalEntry`. The latter manages the access to the linguistic and terminological knowledge. The units of description that have been selected for the LIR such as: lexicalization, sense, definition, usage context, and notes form an eclectic set of data categories. These units constitute useful information for ontology engineers when e.g., editing lexicalizations and browsing available linguistic information such as alternative lexicalizations and translations.

In Figure 6.10 we show the *Linguistic Information* page implemented in LabelTranslator to allow ontology users to manage the linguistic information provided by the LIR model. The page shows five sections that correspond to the lexical entries of the selected ontology element (*FAO* in our example). For instance, in this case the concept *FAO* has three lexical entries, two in English and one in Spanish.

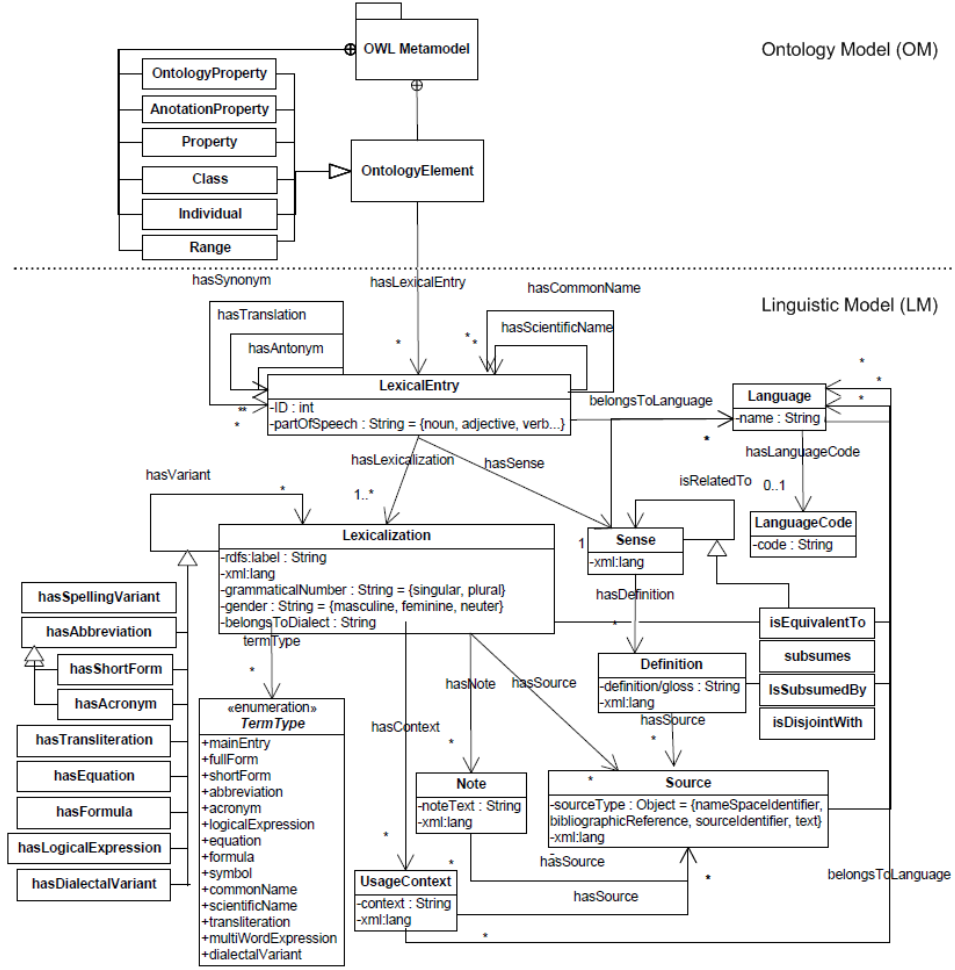


Figure 6.9: Connecting the Ontology Model with the Linguistic Model (taken from [Montiel-Ponsoda, 2011b]).

In our approach, the LIR model is represented as an ontology, with instances representing the lexical knowledge. All information managed by the LIR model is controlled by a specialized unit of the repository component. This unit provides the following features:

- It provides a special API to retrieve linguistic knowledge or to update the linguistic model. Also, it acts as a wrapper around any possible representation of the model.
- It implements both load and save mechanisms which can serialize the lexical entries associated with one ontology.

6.7. LABELTRANSLATOR SYSTEM

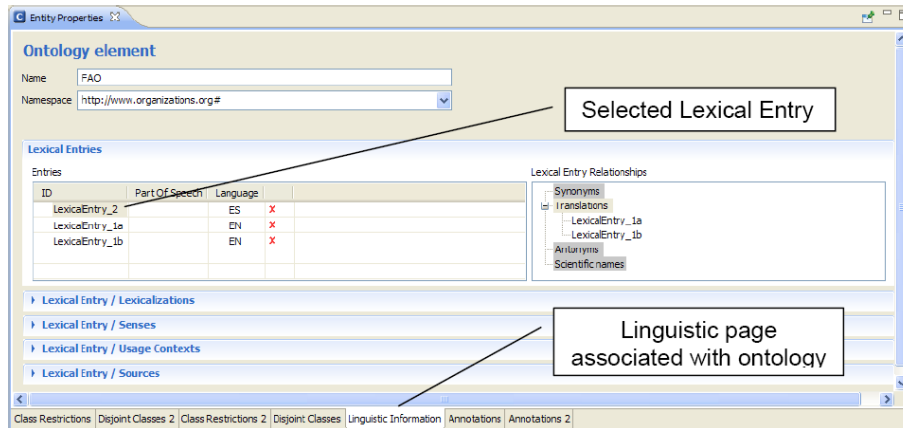


Figure 6.10: Linguistic Information page that support the LIR model.

6.7.3 Localization Management

The Localization Management is the core of the LabelTranslator system. Our solution provides a flexible mechanism to support a collaborative and distributed scenario for localizing an ontology. In our implementation, the workflow localization management implements the strategy described in section 6.5

Each one of the ontologies imported in the NeOn toolkit is associated with a set of initialization parameters (e.g., user roles, assigned tasks, etc.) which define the behavior of the workflow. In order to configure the localization parameters the localization management module extends the NeOn toolkit with a set of wizards⁸. For example the *user wizard* shown in Figure 6.11, allows for the managing of the profile of each participant of the localization activity. The wizard records information about the skills of each participant (source and target languages), and describes the roles, operations and policies that apply to a certain ontology. All this information is used by LabelTranslator for checking the users credentials at login time, and for determining whether a user is allowed to perform a certain operation based on the policies of the ontology to be localized. In our approach a user can play several roles in the localization activity. For example, a user Elena can play the role of Translator and Reviewer.

In the remainder of this section, we first present the synchronization component, which is used to maintain the ontological and linguistic information updated. Then, we describe the main features of the automatic localization workflow, which provides a flexible mechanism to supports a collaborative and distributed scenario.

⁸A software wizard is a user interface element that presents a user with a sequence of dialog boxes that lead the user through a series of well-defined steps. Tasks that are complex, infrequently performed, or unfamiliar may be easier to perform using a wizard.

Localization

Users
Adds users to Ontology Localization Activity

New user

First Name* | Last Name* |

Email* |

Role* | Language skills*

☐ Translator | ☐ English-Southern_Sotho

☐ Reviewer | ☐ English-Spanish

☐ English-Sundanese

☐ English-Swahili

Fields marked with * are mandatory.

Save Cancel

Available Users

Name	Email	Role	Skills
elena montiel	jmem_e...	Translator	English-Spanish
jorge gracia	mespino...	Reviewer	English-Spanish

? < Back Next > Finish Cancel

Figure 6.11: User wizards used by the Workflow Localization Manager.

Synchronization component

In the NeOn ToolKit, an advanced change tracking based on Resource Delta⁹ is able to capture changes even when ontological terms have changed their position within the ontology model. By adopting this feature, the synchronization component can accurately identify the minimal set of changes needed to adjust the structure of the linguistic model.

In a nutshell, the synchronization component is notified about events that consist of ontology changes performed by the user in the ontology editor. For each of these events, the synchronization component stores the change information in the Sqlite database¹⁰.

Concretely, the information stored for each change is a tuple with i) the type of change (e.g., add, delete or rename¹¹), ii) the type of ontology term

⁹A resource delta represents changes in the state of a resource tree between two discrete points in time.

¹⁰SQLite is a software library that implements a self-contained, serverless, zero-configuration, transactional SQL database engine (<http://www.sqlite.org/>).

¹¹This operation is triggered when the label of an ontological term is renamed.

(e.g., concept, attribute, relation or instance), iii) the related ontology (e.g., name or identifier of the ontology), and iv) the label or identifier of the ontological term on which the change was executed. For example, adding a concept in the ontology editor creates the following tuple in the database (“add”, “concept”, “university Ontology”, “academicSupervisor”). Finally, this component is also in charge of synchronizing the changes by using the method presented in section 6.5.1

Localization Management Component

In our implementation, the localization component automatically implements the actions defined in the workflow. Thus, this component takes care of enforcing the constraints imposed by the collaborative workflow. In detail, whenever a new workflow action is performed, the component performs the following tasks:

- It gets the identity and role of the user performing the action.
- It gets the status of the ontology element/translation associated to the action/change
- It verifies that the role associated to the user can perform the requested action when the ontology element/translation is in that particular status.
- If the verification succeeds, it performs the workflow action (e.g., enabling all corresponding fields in the interfaces); if the verification does not succeed, no action is performed.

Additionally, the localization component extends some views in the Neon toolkit which allow ontology localization stakeholders i) to see the appropriate information of the translations in the workflow and ii) to perform the applicable workflow actions (select, translate, review, etc.), depending on their role (as described in the section 6.5.2).

Figure 6.12 shows the perspective¹² used by LabelTranslator in order to support the localization workflow. The Ontology Navigator in the figure is located on the left hand side of the main view. It contains all ontologies that need to be localized. The localization view is located in the middle of the main view. This view is used to add or to update the translations associated with the ontology terms that have been selected in the project tree on the left. Each ontology term is located in its own row. The localization view contains several shortcuts that make work faster, and are enabled according to user profile. Finally, the filter view is located on the right hand side of the figure. It contains several check boxes that allow the user to modify what items are shown in the localization view.

¹²A perspective is a visual container for a set of views and content editors.

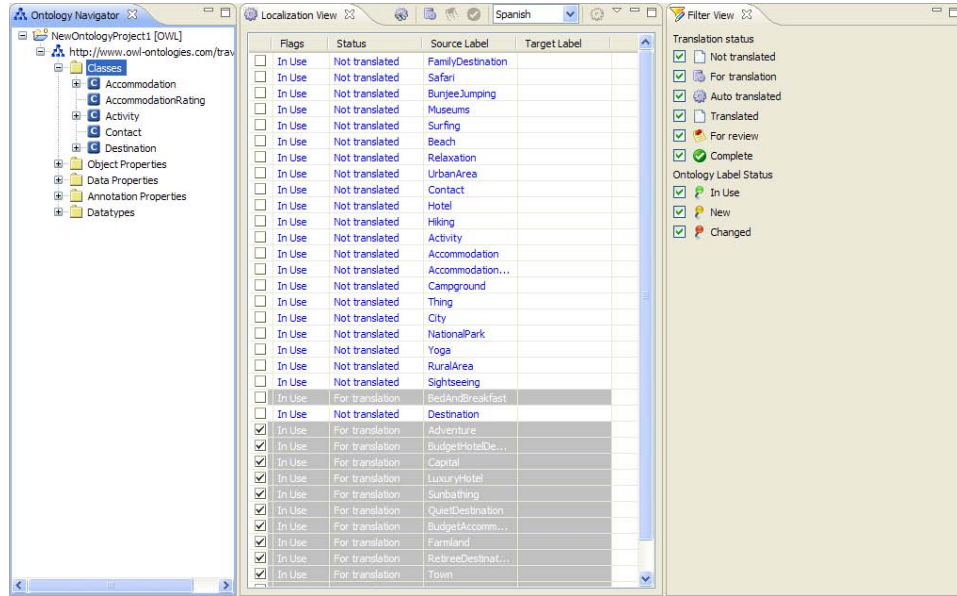


Figure 6.12: A perspective of the Ontology Localization Activity.

6.7.4 Ontology Translator

This component is responsible for obtaining the most probable translation for each ontology label. LabelTranslator allows localizing ontologies in English, German and Spanish. The operations executed by the Ontology Translator module are achieved with the help of different translation methods. In our approach these methods are strategically composed with the aim of improving the quality of the obtained translations.

The current version of LabelTranslator allows adapting (translating) an ontology to Linguistic Level - difficulty 1. This means that our system only allow the translation of ontology concepts, attributes and relations (see section 4.3 for more details). For the translation of these type of ontology elements we have identified two translation strategies: one for *simple labels* and one for *compound labels*.

In the following we describe the main features of the translation strategies identified for the translation of simple and compound labels.

6.8 Translation Strategies Used in LabelTranslator

As a motivating example to illustrate the results obtained by our system, let us consider the extract of the sample university ontology shown in Figure 6.13. Let us suppose that the user wants to translate the term *chair* from English into Spanish. According to the domain of the sample ontology,

6.8. TRANSLATION STRATEGIES USED IN LABELTRANSLATOR

the correct translation of the selected term should be in the sense of a professor, not in the sense of a place where a person can sit down and neither in the sense of an execution instrument by electrocution, etc.

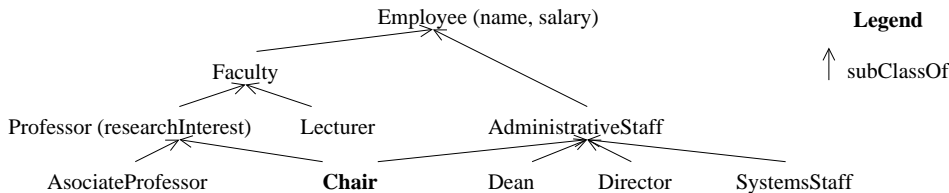


Figure 6.13: Extract of the sample university ontology.

6.8.1 Translating Simple Labels

The strategy used for the translation of simple labels is a hybrid composition of two components, see also figure 6.14. The first component combines three different translation methods (*dictionary-based*, *thesauri-based*, and *online MT systems*) to discover candidate translations. The second component uses a method based on the *comparison of ontology structures* to discover the semantic senses of each translated label.

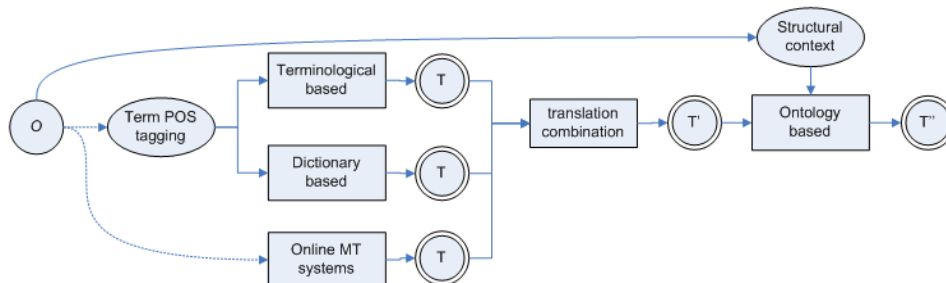


Figure 6.14: LabelTranslator strategy to localize concept, attributes and relation terms represented by simple labels.

First component - obtaining candidate translations.

The first component takes as input an ontology label l described in a source language and returns a set of possible translations $T = \{t_1, t_2, \dots, t_n\}$ in a target language. In order to discover the translations of each ontology label, each translation method accesses different lexical resources. On the one hand, the thesauri-based approach uses IATE¹³. On the other hand, the

¹³<http://iate.europa.eu/iatediff/SearchByQueryLoad.do?method=load>

dictionary-based approach uses the multilingual dictionary Wiktionary¹⁴. The online MT approach uses different multilingual systems such as Google-Translate¹⁵, Babelfish¹⁶, and FreeTranslation¹⁷. A *buffer* stores previous translations to avoid accessing the same data twice.

The algorithm used by the first component is summarized as follows:

1. If the selected ontology label is already available in the target language in our buffer, then LabelTranslator just displays it, with all the relevant available information,
2. If the translation is not stored locally, then each translation method accesses remote repositories to retrieve possible translations. A simple disambiguation process based on term POS tagging is used to avoid an explosion of nuisance candidate translations.
3. If no results are obtained from the two previous steps, then the user can enter his/her own translation (together with the definition).

To combine the output of the different translation methods we use a linear combination of translations [Nie et al., 2001]. In our approach, the translation of an ontology label denoted by t , is a tuple $\langle trs, senses \rangle$, where trs is a translated label in the specific target language, and $senses$ is a list of semantic senses extracted from different knowledge pools. In the following we briefly describe the task of automatically retrieving the possible semantic senses of a translated label (second component in our translation strategy).

Second component - obtaining semantic senses.

In order to discover the senses of each translated label (t_i), we have considered the *ontology structure comparison* approach proposed in a previous work [Trillo et al., 2007]. Our system takes as input a list of words (each t_i), discovers their semantics in run-time and obtains a list of senses extracted from different ontology pools; it deals with the possible semantic overlapping among senses. We summarize here the key characteristic of the sense discovering process:

1. To discover the semantic of the input words, the system relies on a pool of ontologies instead of just a single ontology.
2. The system builds a sense (meaning) with the information retrieved from matching terms in the ontology pool.

¹⁴<http://en.wiktionary.org/wiki/>

¹⁵<http://www.google.com/translate.t>

¹⁶<http://babelfish.altavista.com/>

¹⁷<http://ets.freetranslation.com>

6.8. TRANSLATION STRATEGIES USED IN LABELTRANSLATOR

3. Each sense is represented as a tuple $s_k = \langle s, \text{grph}, \text{descr} \rangle$, where s is the list of synonym names¹⁸ of keyword k , grph describes the sense s_k by means of the hierarchical graph of hypernyms and hyponyms of synonym terms found in one or more ontologies, and descr is a description in natural language of such a sense.
4. Matching terms could be ontology classes, properties or individuals, three lists of possible senses are associated with each keyword k : S_k^{class} , S_k^{prop} and S_k^{indv} .
5. Each keyword sense is enhanced incrementally with the synonym senses (which also searches the ontology pool).
6. A sense alignment process integrates the keyword sense with those synonym senses representing the same semantics, and discards the synonym senses that do not enrich the keyword sense.

A detailed description of this process can be found in [Trillo et al., 2007]. In order to perform cross-language sense translations, the external resources are limited to those resources that have multilingual information like EuroWordNet; however other resources can be used too. For example, a specific domain resource for the FAO (Food and Agricultural Organization) is Agrovoc¹⁹, which could cover the vocabulary missed in EuroWordNet. The multilingual retrieval of a word sense (synset) in EuroWordNet is done by means of the InterlingualIndex (ILI) that serves as a link between the different wordnets. For example, when a synset, e.g., “chair” with the meaning “the position professor”, is retrieved from the English wordnet, its synset ID is mapped through the ILI to the synsets IDs of the same concept in the different languages-dependent wordnets,(German, Spanish, etc.) that describe the same concept, but naturally contain the word description in its specific language. A similar retrieval process is used in the case of multilingual ontologies, but using the references between concepts and labels as offered by the standard owl:comment and rdfs:label properties.

Coming back to the example, in Figure 6.15 we show the translations of the ontology label “chair” from English into Spanish; our prototype finds eight translations, in the figure we only show three. Notice that t_3 has the desired semantics according to the similarity with the lexical and semantic ontology context (see figure 6.13).

Once the semantic senses have been identified, the *ontology structure comparison* method uses a ranking method for sorting the list of translations according to similarity with the *structural context* of the label to be

¹⁸The system extracts the synonym names of a term by consulting the synonym relationships defined in the ontology of such a term.

¹⁹http://www.fao.org/aims/ag_download.htm

some translations of the ontology label “chair” (3/8)	
$t_1 = \langle \text{silla} ; \left[s_1^{\text{class}} = \{ \text{EWN1\#silla, silla, “asiento para un persona”} \} \right] \rangle$	<pre> asiento v mesa barbero trono ... </pre>
$t_2 = \langle \text{presidente} ; \left[\begin{array}{l} s_1^{\text{class}} = \{ \text{EWN1\#presidente, presidente, “jefe de estado de EEUU”} \} \\ \text{Adams Arthur Buchanan ...} \\ \text{presidente de mesa} \\ s_2^{\text{class}} = \{ \text{EWN2\#presidente, presidente, “líder de un encuentro”} \} \\ \text{administrador academico} \\ s_3^{\text{class}} = \{ \text{EWN3\#presidente, presidente, “persona que dirige ...”} \} \\ \dots \end{array} \right] \rangle$	<pre> jefe de estado v Adams Arthur Buchanan ... v presidente de mesa v administrador academico v ... </pre>
$t_3 = \langle \text{cátedra} ; \left[s_1^{\text{class}} = \{ \text{EWN1\#cátedra, cátedra, “la posición de un profesor”} \} \right] \rangle$	<pre> posición v ... </pre>

Figure 6.15: Some translations of the ontology label “chair” into Spanish.

translated. The ranking method relies on the disambiguation algorithm described in [Trillo et al., 2007]. Once all the translations are ranked, the method allows two operation modes:

- Semi-automatic mode: It shows a list with all the possible translations sorted decreasingly. The method proposes the most relevant translation to be selected first although the user can change this default selection.
- Automatic mode: It automatically selects the translation with the highest score.

Next, we first describe how the system obtains the context of each ontology label, and then we describe the disambiguation algorithm used to sort the translations according to similarity with their context.

Determining the Context of an Ontology Term

We defined context as the information/knowledge that can be used additionally to perform some task. In our approach, the context of an ontology term is used to disambiguate the lexical meaning of the term. To determine the context of an ontology term, the system retrieves the labels of the set of terms associated with the term under consideration. The list of context

6.8. TRANSLATION STRATEGIES USED IN LABELTRANSLATOR

labels, denoted by C , comprises a set of names which can be direct label names and/or attribute label names, depending on the type of term that is being translated.

In order to mitigate risks associated with system performance, the *ranking method* limits the number of context labels used to disambiguate the translated label. Every context label $c \in C$ is compared with the ontology label l using a measure based on Normalized Google Distance [Cilibrasi and Vitányi, 2007] (NGD). NGD measures the semantic relatedness between any two terms, considering the relative frequency in which two terms appear in the Web within the same documents. Those labels with the higher values of similarity are chosen (maximum 3). To discover the senses of each context label (denoted by S_c), the system performs the same process used to discover the senses of each translated label (as explained in the previous section).

In Figure 6.16, on the left, the dashed area represents all the context labels found for the ontology label “chair”. Our prototype finds five labels, but only selects three (see the dotted area) to disambiguate the term. In the table on the right, we show for each type of ontology term (concept, attribute, or relation) the context labels that could be extracted. For instance, for the concept “chair” the system retrieves its hypernyms, hyponyms, attributes, and sibling concepts.

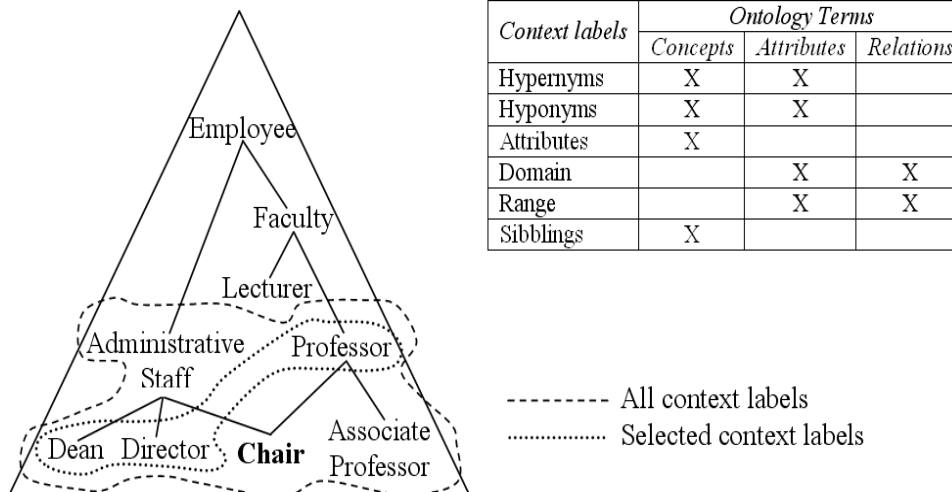


Figure 6.16: Context of the ontology label “chair”.

Disambiguating the Senses of the Translations

In some works [Pazienza and Stellato, 2006, Pedersen et al., 2005] the glosses are considered as a very promising means of measuring relatedness, since they can be used: 1) to make comparisons between concepts semantically different, and 2) to discover relations of which no trace is present in the

resource they come from. For the first version of the prototype, we use a ranking method based on glosses as proposed in [Pedersen et al., 2005] to sort the translations according to their context. However, we recognize that glosses are by necessity short and may not provide sufficient information on their own to make judgments about relatedness.

For the second version of the prototype, we carry out disambiguation in relation to the senses of each translated label and the senses of the context labels. The ranking method we use to compare structures relies on an equivalence probability measure between two candidate structures, as proposed in [Trillo et al., 2007]. We assume that we have a taxonomy or ontology entity o_1 and we wish to deduce if it is similar to another taxonomy or ontology entity o_2 from a reference taxonomy or ontology (i.e., EuroWordNet) in the same language. We shall make a simplifying assumption that each ontology entity is associated with a unique label, e.g., l_{o_1} . As such we wish to deduce if o_1 represents the same concept as o_2 and hence if l_{o_2} is a translation for l_{o_1} . Our model relies on the Vector Space Model [Raghavan and Wong, 1986] to calculate the similarity between different labels, which essentially involves calculating a vector from the bag of words contained within each label and then calculating the cosine similarity between these vectors. We shall denote this as $v(o_1, o_2)$. We then use four main features in the calculation of the similarity [McCrae et al., 2011a]:

- The VSM-similarity between the labels of entities, o_1, o_2 .
- The VSM-similarity between any glosses (descriptions) that may exist in the source or reference taxonomy/ontology
- The hypernym similarity given to a fixed depth d , given that set of hypernyms of an entity o_i is given as a set

$$h^O(o_i) = \{h | (o_i, h) \in H\}$$

Then we calculate the similarity for $d > 1$ recursively as

$$s_h(o_1, o_2, d) = \frac{\sum_{h_1 \in h^O(o_1), h_2 \in h^O(o_2)} \sigma(h_1, h_2, d)}{|h^O(o_1)| |h^O(o_2)|}$$

$$\sigma(h_1, h_2, d) = \alpha v(h_1, h_2) + (1 - \alpha) s_h(h_1, h_2, d - 1)$$

And for $d = 1$ it is given as

$$s_h(o_1, o_2, 1) = \frac{\sum_{h_1 \in h^O(o_1), h_2 \in h^O(o_2)} v(h_1, h_2, 1)}{|h^O(o_1)| |h^O(o_2)|}$$

- The hyponym similarity, calculated as the hypernym similarity but using the hyponym set given by

$$H^O(o_i) = \{h | (h, o_i) \in H\}$$

We then incorporate these factors into a vector x and calculate the similarity of two entities as

$$s(o_1, o_2) = w^T x$$

Where w is a weight vector of non-negative reals and satisfies $\|w\| = 1$, which we set manually.

In our example, “cátedra” (cathedral) in the sense of “the position of professor” is ranked as first translation of the ontology label “chair”. Once the right sense has been selected, the system updates the linguistic information of the corresponding ontological term.

6.8.2 Translating Compound Labels

Compound labels which have an entry in linguistic resources such as lexical databases, dictionaries, etc. (for example “jet lag”, “travel agent” and “bed and breakfast”) are treated as single words in our approach. Others like “railroad transportation”, which have no entry in the previous resources, are translated using a compositional method (see figure 6.17)

This approach uses a hybrid composition of two translation components. The first component is the same as the first component used for translating simple labels. The second component relies on a similar approach to the example-based method identified as part of the corpus-based translation techniques (see section 5.5). There are two main differences. First, instead of extracting the bilingual translation templates from a simple monolingual corpus or from parallel corpora, we derived these templates from different ontologies. The second difference relates to the used method to discover the translations. We do not extract the translations from a corpus, but from different linguistic resources.

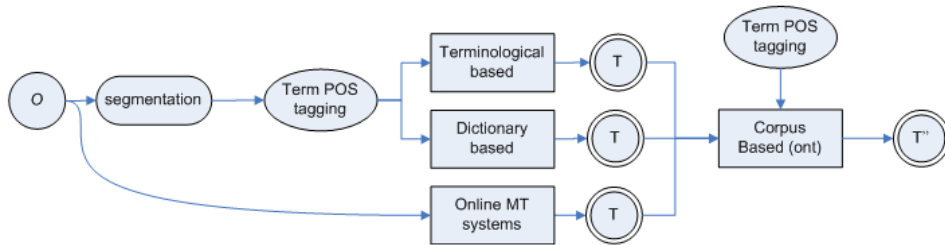


Figure 6.17: LabelTranslator strategy to localize concept, attributes and relations represented by compound labels.

In a nutshell, this method splits the label into tokens (“railroad” and “transportation” in the example); the individual components are translated

and then combined into a compound label in the target language. Care is taken to combine the components respecting the word order of the target language. A set of lexical templates derived from different ontologies are used to control the order of translation. The main steps of the algorithm are:

1. The compound label is normalized, e.g., rewritten in lowercase, hyphens are removed, it is split into tokens (see segmentation task in the figure), etc.
2. A set of possible translations is obtained for each token of the compound label using the different translation paradigms (first component in our translation strategy). The method uses all possible combinations of translation obtained for each token.
3. Since translations between languages do not keep the same word order, the algorithm creates candidate translations in the target language using lexical templates²⁰. Each lexical template contains at least a pair of patterns, namely ‘source’ and ‘target’ patterns. A source pattern is a template to be compared with the *tagged compound label*²¹, described in the source language, while the target pattern is used to generate the label in the target language. If no applicable template is found, the compound label is directly translated by the translation service.
4. All the candidate labels that fulfill the target pattern are returned as candidate translations of the compound label.

In the following we describe the process to learn the lexical templates which are used to control the order of translation of compound labels.

Learning Lexical Templates from Ontological Labels

We believe that lexical templates used to translate compound labels are a necessary component to produce high quality translations because 1) it guarantees grammatical output and, 2) it makes sure that the structural source language meaning is preserved. In our approach, we used a semi-automatic process to obtain the lexical templates. As we explained before, each lexical template is composed of source and target patterns. The ontology labels used to learn the source patterns were extracted from different domain ontologies expressed in English, German, or Spanish. Each label was tokenized and tagged using the language independent part-of-speech tagger proposed in [TreeTagger, 1997]. The labels used to learn the target

²⁰The notion of lexical template proposed in this paper refers to text correlations found between a pair of languages.

²¹We use TreeTagger [TreeTagger, 1997] in order to annotate the compound labels with part-of-speech and lemma information.

6.8. TRANSLATION STRATEGIES USED IN LABELTRANSLATOR

patterns were extracted either from the multilingual information associated with each ontological term or by means of a manual translation process. The same process used to annotate part of speech (POS) in the labels of the source patterns was used to annotate the labels of the target patterns. The empirical results collected during the learning of lexical templates are briefly described below:

- *Existing ontologies share the same lexical patterns.* For instance, approximately 60% of the labels that describe an ontological concept make use of an adjective followed by a noun (e.g., spatial region, industrial product, natural hazard, etc.). Other labels use as lexical pattern ($\approx 30\%$) a noun followed by another noun (e.g., transport vehicle, knowledge domain, etc.).
- *Ontology labels usually have less than four tokens.* Approximately 85% of labels fulfill this. Thus, for the current prototype we only focus on the definition of lexical templates for compound labels of two or three tokens.

A repository is used to store all the lexical templates obtained for each pair of languages. Table 6.1 shows a sample list of the lexical templates learned to translate compound labels from English into Spanish.

Table 6.1: Some lexical templates to translate a compound label from English into Spanish.

<i>Templates (4/25)</i>	<i>Samples of source and target patterns</i>	
	<i>English</i>	<i>Spanish</i>
$[J_1 N_2]_{en} \rightarrow [N_2 J_1]_{es}$	spatial region→ industrial product→ natural hazard→	región espacial producto industrial peligro natural
$[N_1 N_2]_{en} \rightarrow [N_2 \langle pre \rangle N_1]_{es}$	transport vehicle→ knowledge domain→ research exploration→	vehículo de transporte dominio del conocimiento exploración de la investigación
$[J_1 VB_2]_{en} \rightarrow [VB_2 \langle pre \rangle J_1]_{es}$	remote sensing→	detección remota; detección a distancia
$[J_1 N_2 N_3]_{en} \rightarrow [N_2 \langle pre \rangle N_3 J_1]_{es}$	associated knowledge domain→	dominio de conocimiento asociado

J: adjective; N: noun; VB: verb

As an illustrating example of the compositional method, we show in Figure 6.18 the steps of the algorithm when collecting Spanish translations for

the English compound label “AssociateProfessor”, which was introduced in our motivating example. Our system finds ten translations for the token “associate” and one for “professor” (normalized in the first step). In the next step, our tool searches a lexical template (in our repository) to create candidate translations. In the template found, $[J_1 N_2]_{en}$ represents the source pattern in English whilst $[N_2 J_1]_{es}$ represents the target pattern in Spanish. In both cases, numbers represent the position of each token of the compound label. Notice that, in the last step the candidate translations “profesor socio” (professor member) and “profesor compañero” (accompanying professor) are discarded because they do not fulfill the target pattern.

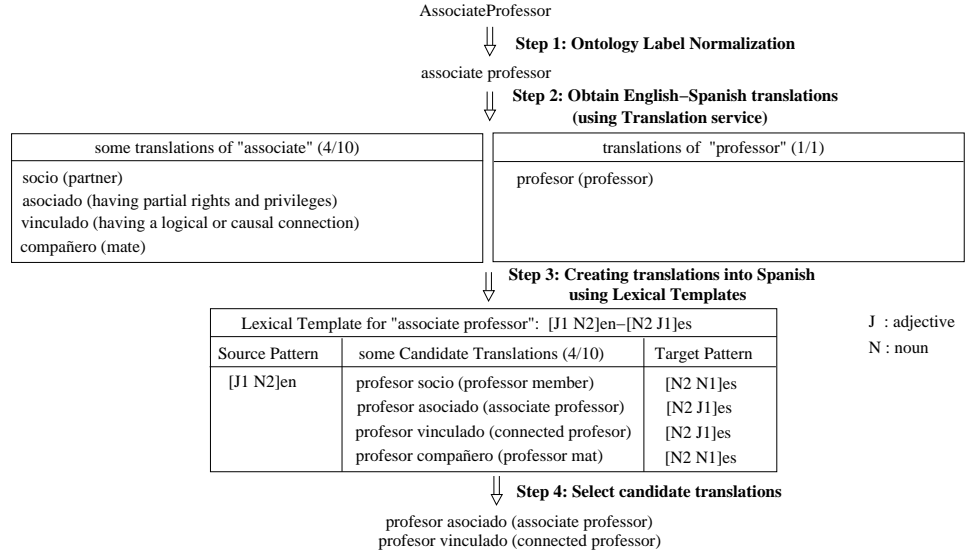


Figure 6.18: Algorithm to translate the compound label “AssociateProfessor” into Spanish.

6.9 Summary of the Chapter

In this chapter we have discussed two important issues related to the ontology localization activity: life-cycle model and system architecture. We have first described the phases of life-cycle model of the localization activity. A description of the key concepts and elements needed to build a ontology localization system have been included.

Second, we have explained the different modules in the generic architecture for an automated ontology localization in collaborative and distributed environments. We have presented the global architecture motivated by the phases identified in the life-cycle model. Also, in order to define the infrastructure requirements we took different key factors from different software

6.9. SUMMARY OF THE CHAPTER

localization approaches and then we compared them with our own observations in the field. Concretely we described three groups of requirements: collaboration and distribution of the tasks, automated translation, and extensibility.

The Ontology Management Module has been introduced as the module that enables ontology editors to automatically manage the multilingual content for localization. We have also explained that the control and management of the localization activity is performed by the Localization Management module. The functionality of the Ontology Translator module has been presented later.

Finally, we have presented the LabelTranslator system, which is our approach to automatically localize ontologies among English, Spanish and German. We have included the description of technical details of the main components in the architecture. Furthermore, we have discussed the different aspects related to the implementation of the Ontology Repository component included in the Ontology Management module. A description of the capabilities of the implemented interfaces has also been included. With regard to the Localization Management module we have described its main functionalities, which are: synchronize the changes of the ontology and linguistic model and implement the actions described in the collaborative workflow. We have finished with the description of the technical details of the main components in the architecture by commenting on some implementation details of the Ontology Translator module. Here, we have described the translation strategies used to localize simple and compound labels.

Chapter 7

Methodological Guidelines

In this chapter we present efficient, prescriptive and detailed methodological guidelines for the ontology localization activity, which are inspired on Software Engineering methodologies. We first present the scope of the methodological guidelines together with a brief description of the NeOn Methodology, which proposes as one of its methodological scenarios the localization of ontologies to support the adaptation of an ontology to different languages and cultures. Second, we describe the process followed to define the activities and tasks of the methodological guidelines. Finally, we detail the guidelines for the ontology localization activity (including the filling card and the activity workflow).

7.1 Neon Methodology as Framework for the Localization of Ontological Resources

Contrary to traditional methodologies [Fernández-López et al., 1999, Staab et al., 2001, Pinto et al., 2004] that provide methodological guidance for ontology engineering, the NeOn Methodology [Suarez-Figueroa, 2013] identifies a set of flexible scenarios that supports different aspects of the ontology development process, as well as the reuse and dynamic evolution of networked ontologies in distributed environments, where knowledge is introduced by different people (domain experts, ontology practitioners) at different stages of the ontology development process. The nine scenarios proposed in the methodology cover commonly occurring situations in the ontology development process that can be combined according to the ontology requirements and the existing resources in the domain. Figure 7.1 presents the nine identified scenarios for building ontologies and ontology networks. The directed arrows with numbered circles associated represent the different scenarios. Each scenario covers a specific process or activity (represented with coloured circles or with rounded boxes) that has to be followed to develop an ontology whenever certain requirements or premises

are given.

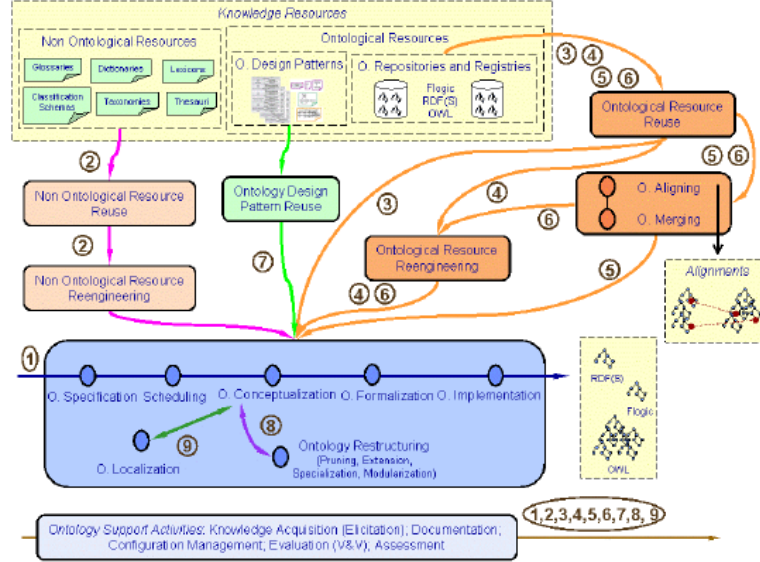


Figure 7.1: NeOn Methodology scenarios for building ontology networks).

The methodological guidelines that we propose in this thesis are intended to assist Scenario 9: Localizing Ontological Resources. The aim is to carry out the localization process of ontologies already conceptualized. Usually, these ontologies are designed without taking into account the multilingual and localization aspects. Therefore, these guidelines aim to reduce costs, improve its quality, and increase the consistency of the localization activity. The proposed guidelines [Espinoza et al., 2012] are particularly intended for ontology stake-holders such as localization managers, translators and reviewers, who are concerned with the ontology localization activity. In addition, they are intended for communities interested in localizing ontologies and those international firms that may promote multilingualism in their working environments for a variety of reasons.

To the best of our knowledge, no guidelines exist for supporting the ontology localization activity. However, software localization methodologies could be adapted for ontology localization, as these methodologies are very general. In the following section we will define the process followed to develop this methodology.

7.2 Research Methodology

This section describes the process followed to develop the ontology localization methodology. The NeOn Methodology [Suarez-Figueroa, 2013] considers the localization of an ontology as one activity. Thus, for describing the

7.2. RESEARCH METHODOLOGY

methodology, the definitions of activity and task set out by the IEEE [IEEE, 2000] have been strictly followed:

- *Activity*. A constituent task of a process. An activity is a defined body of work that is to be performed, including its required input and output information.
- *Task*. The smallest unit of work subject to management accountability. A task is a well-defined work assignment for one or more project members.

For the development of the ontology localization methodology, we adopt the process followed by Garcia Castro [García-Castro, 2008] in the development of a benchmarking methodology. The steps followed are described in the following:

1. To analyze different relevant methodologies from the Software Localization area.
2. To identify the main tasks of selected works by choosing the tasks that are considered in most of the works.
3. To complete these tasks in order to cover the ontology localization requirements.
4. To analyze task dependencies in order to define task order.

The next sections will examine each step in more detail.

7.2.1 Analysis of Relevant Methodologies

In the first step we analyze different relevant methodologies related to the ontology localization activity. In particular, we focus on the Software Localization approaches, which provide an overview of different tasks that deal with issues relevant to this activity, such as the analysis of source material, translation process, or product quality assurance.

In what follows, we will describe different software localization methodologies or in some cases best practices extracted from academia and industry. These methodologies view the localization as a mechanism to satisfy the needs and requirements of international markets or cultural nuances and they also consider the localization as a continuous process. The list of selected methodologies and models are neither meant to be exhaustive nor complete but rather only informative.

Software Localization Methodologies

Unlike software development projects, in which well-established and precise practices and methodologies exist the localization projects do not explain the localization process with the same style and granularity. Possibly, one of the reasons of the lack of a coherent methodology is that the localization does not consist of a discrete process or a defined set of tasks, but rather represents a focal point in the corporate matrix at which various business units, objectives, and processes intersect [Dunne, 2006].

However, in literature there are various works that follow different methodological steps, which are not applicable to all projects, but are to be seen as general goals in the localization chain. For example, Esselink [Esselink, 2000] is one of the first in identifying and describing some typical phases for a localization project sequence:

- *Analysis of source material.* Its goal is to analyze all aspects of the material to be localized to create the foundation for an effective localization. The essential steps of this phase are:
 - To identify the localization problem areas.
 - To select the localization tools.
 - To analyze all aspects of the new localization project.
- *Scheduling and budgeting.* The goal of the schedule and budget task is to ensure a timely shipping of localized software. Some steps are recommendable:
 - To identify all tasks and activities.
 - To define the dependencies between activities
 - To establish the sequences of activities.
- *Identification and setup of both source and target language terminology.* Its goal is to define the basic terminology list, also called project glossary, which would typically contain terms that are commonly used in the product user interface or support documentation. Some tasks of this phase are:
 - To select methods for collecting terminology information.
 - To extract the terminology information.
 - To identify the terms that should not be localized such as proper names.
- *Preparation of source material.* Once the material has been analyzed a translation kit is created for the translators. The preparation of the source material includes:

7.2. RESEARCH METHODOLOGY

- To investigate leveraging possibilities for the software, i.e., checking whether existing translations can be automatically re-used.
- *Translation of software.* The goal of this task is to translate the software resources such as dialog boxes, menus and strings and to validate the translations in context (in the running applications).
- *Translation of online help and documentation.* As soon as a software glossary or a preliminary build of the localized software is available, translation of online help and documentation can start.
- *Engineering and testing of software and online help.* The engineering task involves resizing the user interface, assigning unique hot keys, and compiling the localized resource files into a running application. The main step of this phase is:
 - To test the functionality of localized versions.
- *Processing updates.* The goals of this task are to process updates using files compares, copying and pasting, or translation memory tools and to prevent unnecessary software engineering, testing or desktop publishing work.
- *Product quality assurance and delivery.* Its goal is to check the quality of all localized material. The pre-delivery QA check includes:
 - To review the quality translations.
 - To finalize bug or problem reports.
 - To ensure that the instructions given in the initial hand-off or statement of work from the publisher were covered.
- *Project closure.* The goal of the project closure task is to organize a post-mortem or project audit with the localization vendor after a project has been completed. Issues involved in this step include:
 - To process the evaluation of the completed project
 - To evaluate the technical and linguistic quality of deliverables.
 - To identify the areas for improvement.
 - To suggest process modifications for future projects.

In a publication on MultiLingual Computing¹, Muller [Müuller, 2009] identifies nine phases for software localization:

¹Multilingual Computing is one of leading industry magazine for Web site globalization, international software development and language technology (<http://www.multilingual.com>)

- *Project setup* phase. The phase is devoted to establish the project plan with milestones, time buffers and constraints. The essential tasks of this phase are:
 - To analyze the characteristics of the material to be translated.
 - To identify the time for translation preparation and revision.
 - To identify the constraints such as project end, money and resources of each task.
- *Translator training* phase. The goal of this phase is to train the translators in the software to be localized. The idea of this phase is to avoid that the first contact of the translator with the software is only a list of words without context. The tasks of which this phase comprises are:
 - To organize a training course of the software to be localized.
 - To provide support tools for solving the translator's problems.
 - To introduce the localization kit so that the translators become familiar with the style guide and the workflow they should follow.
- *Terminology definition* phase. The goal of this phase is to define the basic terminology to be used for translation. The main step of this phase is:
 - To extract the terminology from the elements to be translated.
- *User interface translation*. The main purpose of this phase is to start the localization of the software. The main tasks that follow this phase are:
 - To select the tools for supporting the translation.
 - To enable the interpretation of context-less strings, providing additional information.
- *Test of user interface translation*. The goal of this phase is to control the quality of the localized versions of the software. The tasks recommended are:
 - To design a list of tests that cover a wide range of topics (e.g., check messages for consistent wording, check that the text is not truncated due to its length, etc.)
 - To implement all designed tests.

7.2. RESEARCH METHODOLOGY

- *Documentation translation* phase. Its objective is to translate the documentation of the software product.
- *Review of documentation translation*. In this phase, the translated documentation is checked for possible errors. The steps recommended are:
 - To design and implement both usability and quality documentation tests.
 - To ensure the consistency between the user interface and the documentation.
- *Finalize documentation translation* The main goal of this phase is to update the localized documentation.
- *Lessons learned*. Its objective is to find problems during the localization process between all internal and external team members, the root causes of the problems and the enablers that particularly contribute to the difficulties. This phase involves the following tasks:
 - To identify problems in the performance level.
 - To review the causes of the problems based on the lessons learned log maintained during the project.

In her doctoral work about Internet Software Localization, Jevsikova-[Jevsikova, 2009] identifies five phases that should be carried out to localize a software product. For each phase she describes the goals and the processes used:

- The *Preparational* phase of the localization process aims to evaluate the number of potential users of the localized product, the software and the potential of localization. The preparational phase also helps to choose suitable software to localize.
- *Software adaptation* is the second phase of the localization process, which can be run in parallel with the *translation and adaptation of the dialogs phase* (explained latter on). The difficulty of this phase depends on the level of internationalization of the software. Basically, this phase involves the adaptation of all cultural elements of the software to the target locale. Locale definitions are used to prepare the software for such adaptation.
- The *Translation and adaptation of dialogs* involve as first task the preparation of contextual information for translating the user interface strings. Then, the experimentation of the software is performed

by running the program and looking for interface strings. After the correction of the translation, the testing and correction cycle is repeated, because the correction of one string can cause errors in another interface part.

- *Translation and adaptation of help documents* The main goal of this phase is to ensure the consistency between the user interface and the help information.
- *Overall localization* testing includes internal and external testing when all the previous stages of localization are completed. This stage includes testing of consistency of user interface elements, consistency in token functionality, aesthetics, inter-product communication, scripting considerations, error messages, cross-platform, hardware platform and other [O'Sullivan, 2001].

The SDL Language Technologies², incorporate nine basic steps in their software localization tool (SDL Passolo³). The standard localization process includes the following steps:

- Analysis of the material received and evaluation of the tools and resources required for localization.
- Cultural, technical and linguistic assessment.
- Creation and maintenance of terminology glossaries.
- Translation to the target language.
- Adaptation of the user interface, including resizing of forms and dialogs, as required.
- Localization of graphics, scripts or other media containing visible text, symbols, etc.
- Compilation and build of the localized files for testing.
- Linguistic and functional quality assurance.
- Project delivery.

As can be observed, the previous works contain some similar tasks. In the next step of the development of the methodology we identify the set of common task which these works treat.

²SDL Language Technologies is a division of SDL International, the world leader in Global Information Management (GIM).

³<http://www.translationzone.com/en/products/software-localization/sdl-passolo.asp>

7.2.2 Identification of Main Task

From the works previously described we selected the common tasks, as can be seen in Table 7.1. These common tasks have been grouped into phases according to the phases used in the selected works.

Table 7.1: Common tasks in software localization methodologies

<i>Phase</i>	<i>Task</i>
Preparation	<ul style="list-style-type: none"> • Choose suitable tools for localizing • Select methods and tools for collecting terminology information • Create a translation kit for translators
Translation	<ul style="list-style-type: none"> • Provide contextual information for the translation. • Translate software strings. • Translate documentation and help.
Revision	<ul style="list-style-type: none"> • Implement functionality tests • Implement linguistic and quality tests.
Delivery	<ul style="list-style-type: none"> • Update and compile the localized software

7.2.3 Task Completion

The third step in the development of the methodological guidelines is to complete the ontology localization requirements with additional tasks. In our case, only one task was included to complete the coverage of the selected task. This task has as goal to choose the element that must be localized. This step may choose to discover the translations of certain candidate ontology elements and ignore others (e.g., only localize ontology concepts and not ontology relations). Therefore, a task related to this involvement was added.

7.2.4 Analysis of Task Dependencies

The last step, once a set of candidate tasks was selected was to identify the logical order in which these tasks are to be performed. To arrange the order of the tasks, a task A was considered to be previous to a task B if the output of task A is needed as an input in task B.

An ordered ontology localization process with sequential tasks was obtained. To simplify the methodological guidelines, some tasks were merged into one. For example, the *choose suitable tools for localizing*, *select methods and tools for collecting terminology information*, and *create translation kit for translators* were merged because the tasks and their outputs were highly coupled.

The resulting tasks were re-labeled for our purposes. Therefore, the final list of tasks include the following:

- Select the most appropriate linguistic assets.
- Select ontology label(s) to be localized.
- Obtain ontology label translation(s)
- Evaluate label translation(s)
- Ontology update

In the following sections we will define the actors involved in the different tasks of the ontology localization activity. Then, we will describe in detail the tasks for carrying out this activity.

7.3 Methodological Guidelines for Ontology Localization

In this section, our purpose is to explain the guidelines set out to help ontology developers in the ontology localization activity. The principles that guide the construction of such guidelines are the following:

- The guidelines should be general enough in the sense that they should help software developers and ontology practitioners to localize ontologies in different natural languages and domains.
- The guidelines should define each activity or task precisely; they should clearly state its purpose, its inputs and outputs, the actors involved, when its execution is more convenient, and the set of methods, techniques and tools to be used for executing them.
- To facilitate a prompt assimilation of the ontology localization by software developers and ontology practitioners, we present the guidelines in a prescriptive way, not specifically oriented to researchers

First, we present the different kind of actors involved in the ontology localization activity. Then, we describe the guidelines for localizing ontologies to different natural languages.

7.3.1 Ontology Localization Actors

The different tasks involved in the ontology localization activity are carried out by different actors according to the kind of roles that must be performed in each task. In the following, we briefly describe the main actors involved in the localization activity

7.3. METHODOLOGICAL GUIDELINES FOR ONTOLOGY LOCALIZATION

- *Domain Experts and Ontology Development Team.* The domain expert or experts and the Ontology Development Team (ODT) are responsible for performing one of the first tasks in the ontology localization activity. Their work consists of selecting the right resources and tools to perform the ontology localization activity.
- *Localization Manager.* The localization manager plays a key role in the localization activity, as s(he) must prepare all technical aspects of the localization activity, including the localization material (e.g., identifying the ontology elements to be localized) and distributing it to the localization team, setting up the localization team, as well as assigning and monitoring the tasks. Another task to be performed by the localization manager is the updating and final quality revision of the translated ontology
- *Linguists.* These specialists can either be:
 - *Translator (Localization Specialist).* Once the localization manager assigns the localization tasks to each member, the translator or localization specialist takes care of discovering the most appropriate translations for each ontology element.
 - *Reviewer (QA Specialist).* The reviewer or Quality Assurance (QA) specialist reviews the translated ontology elements. A reviewer does not necessarily focus on the quality of the translations, but on the linguistic and stylistic quality of the translated ontology elements. The revision is a final language check for spelling errors, grammar mistakes and consistency

The current industry trend is to use external localization service providers for the translation task to avoid the high fixed cost of using in-house translators, and use translators focused on the target markets and who know the up-to-date usage of particular languages. We conceived a similar situation for ontology localization, in which translators and reviewers can be internal or external to the organization that develops the ontology, and who work in a distributed environment. Figure 7.2 shows the high-level overview of the people who are directly involved in the ontology localization activity, both on the localization service and on the ontology publisher side. The localization manager and the ontology expert are responsible for the communication between both groups. In Fig. 7.2, the Quality Assurance Department (QA Department) performs a final quality check on all localized ontology elements received from the localization service provider to find out possible problems in the translations.

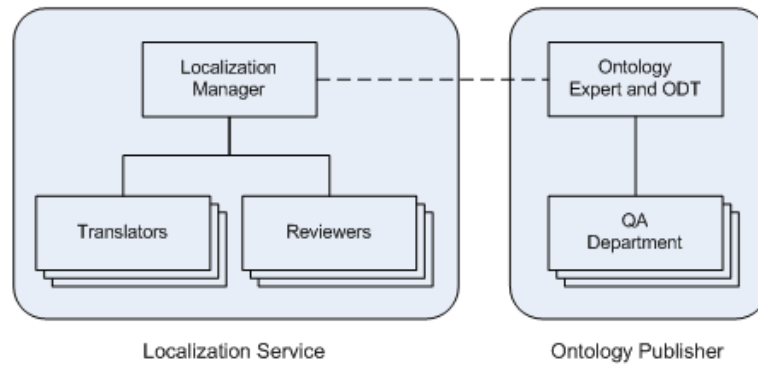


Figure 7.2: Actors involved in the Ontology Localization Activity.

7.3.2 Ontology Localization Guidelines

The ontology localization guidelines have been created in the context of the methodological work included in [Suarez-Figueroa, 2013] to build ontology networks. Thus, taking into account the aforementioned methodological work, we provide the filling card for the ontology localization shown in Figure 7.3. Such kind of filling card allows us to explain the information of the ontology localization activity in a practical and easy way.

The methodological guidelines for carrying out the ontology localization activity can be seen in Fig. 7.4. The workflow shows the main tasks involved, their in-puts, outputs and actors. The result of this activity is an enriched ontology (multi-lingual) with linguistic information (into target language) associated to each localized term. The tasks for carrying out the ontology localization activity are explained in detail in the following:

Task 1. Select the most appropriate linguistic assets

The goal of this activity is to select the most appropriate linguistic assets that help in the localization activity. Domain experts and ODT carry out this activity, taking as input the ontology to be localized. The activity output is a set of linguistic assets that can help to reduce the cost, improve the quality and increase the consistency of the localization activity. The choice of a specific resource is performed manually, taking into account that the linguistic assets comply with the following characteristics:

- *Consensus.* Resources used should contain multilingual terminology consensually accepted by the community (authoritative resources), thus the effort and time spent in finding out adequate translation labels for ontology terms would decrease considerably. In this sense, internal resources, such as terminology databases, glossaries, etc., maintained by the organization or individual itself are good representatives of consensual resources.

7.3. METHODOLOGICAL GUIDELINES FOR ONTOLOGY LOCALIZATION

Ontology Localization	
<i>Definition</i> Ontology localization refers to the adaptation of an ontology to particular language and culture	
<i>Goal</i> To translate an ontology expressed in a source natural language into a target natural language.	
<i>Input</i> An ontology whose ontology labels are expressed in one or several natural languages, from which one is selected as source natural language.	<i>Output</i> An ontology whose ontology labels have been translated to the target natural language. The resulting translations are added to available labels of the original ontology already in one or several languages.
<i>Who</i> Software developers and ontology practitioners, who form part of the ontology development team, in collaboration with domain and linguistic experts.	
<i>When</i> Once the conceptual model of the ontology is stable, with the aim of avoiding spending time and resources in a model that is not definitive.	

Figure 7.3: Ontology Localization Filling Card.

- *Broad coverage.* Resources should cover translation information from general to specific domain labels. It is advisable to use domain specific resources (e.g., a glossary of financial terms or a legal dictionary) when translating domain ontologies, since they will contain the appropriate terminology. Also, since each resource supports different features and language sets, the selected resources should cover all target languages for current and possible future ontology localization projects.
- *High precision.* Resources used for ontology localization should be able to identify the morphological and lexicographical differences that exist between different natural languages.

To select the appropriate translation tool for performing the ontology localization activity, the preliminary guidelines presented in Table 7.2 are recommended.

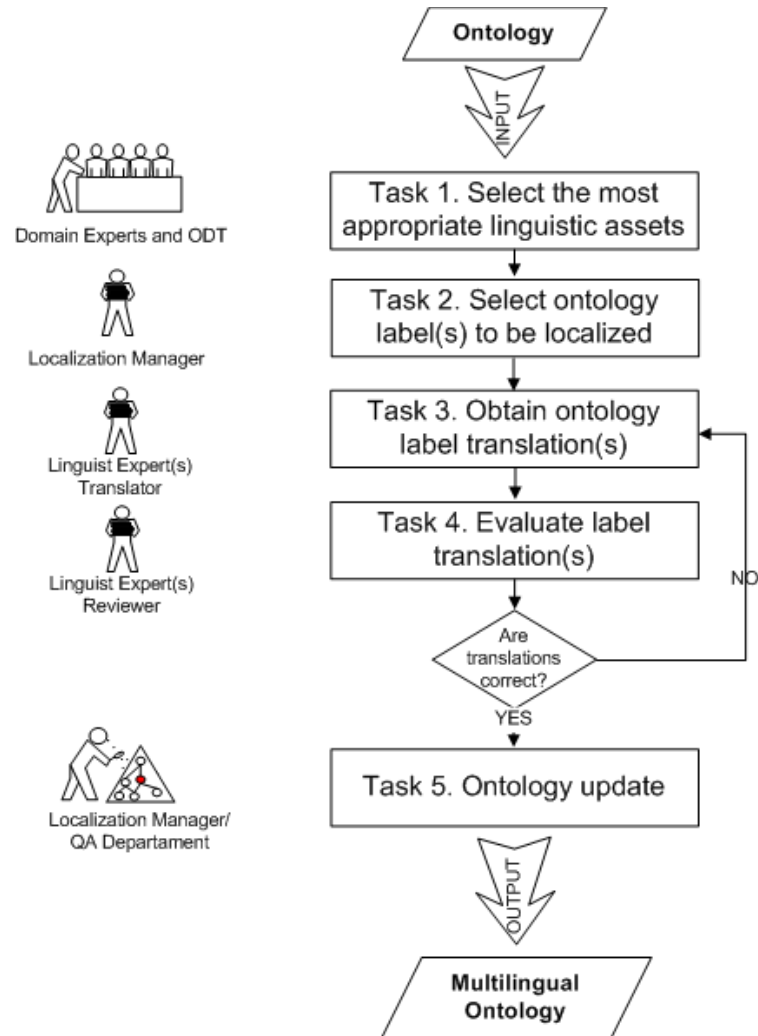


Figure 7.4: Tasks for Ontology Localization.

Task 2. Select the ontology label(s) to be localized

The goal of this task is to select the ontology label(s) to be localized. The Localization Manager carries out this task, taking as input an ontology whose labels are expressed in a source natural language and need to be localized to a target language. By default, all labels of concepts, attributes, relations and instances will be selected to be translated. However, it may happen that the ontology has been partially localized, and only the remaining labels need to be translated, or re-translated if, according to the Localization Manager, they have not been properly translated. The task output is a set

7.3. METHODOLOGICAL GUIDELINES FOR ONTOLOGY LOCALIZATION

Table 7.2: Ontology Localization task and its corresponding tool.

<i>Type of Ontology Term</i>	<i>Translation Tool</i>	<i>Comments</i>
Ontology concepts, attributes and relations	Ontology Localization Tool	<p>The main difficulty at this level is related to the fact that labels for concepts, attributes and relations are usually short (isolated) labels, not inserted in a sentence or text. Therefore, a tool designed for the purpose of translating ontologies is required at this stage to extract the label specification and its context correctly. The label context is required for discovering the label sense (for disambiguation purposes).</p> <p>Consider, for example, the word ‘plant’, which depending on the context can be translated into Spanish as ‘planta’ in the sense of “living organism” or ‘fábrica’ in the sense of “industrial plant”.</p>
Ontology instances	Computer aided translation tool or Ontology localization tool	<p>The main complication at this level is to decide which instances should be translated and which ones should not. A big part of the instances are represented by proper names, and therefore should not be translated (e.g., a label containing “Michael Schumacher” should not be translated). However, other instances such as “South America” should be translated to other natural languages, as they have traditionally well-established and accepted translations.</p>
Ontology term annotations	Translation memory tool	<p>The major cost involved at this level is the difficulty in correctly translating long pieces of text. To provide a human-readable description of a term, the RDF(S) and OWL ontology languages use for example the <code>rdfs:comment</code> statement, where a textual comment can be added. Thus, this level involves the difficulty to translate a whole sentence (not isolated labels or terms) which are part of the annotation of concepts, attributes or instances in ontologies.</p>

of ontology labels and their context⁴. The context describes the meaning of a specific label in the ontology and consists of a small excerpt of ontology labels around the ontology label itself (e.g., direct hypernym labels, hyponym labels, etc.).

Task 3. Obtain ontology label translation(s)

For each ontology label, the goal of this task is to obtain the most appropriate translation in the target language. Translators carry out this task taking as input the ontology label(s) to be localized. Different machine translation techniques can be used to perform this task in an automatic manner (see section 5.2 for more details). Basically, the MT techniques proposed in this thesis base their operation on some lexical or semantic resources for discovering the more appropriate translations. Thus, we can identify translation techniques based on: dictionaries, terminologies, thesaurus, online-services, corpora, ontologies, etc. The identification and combination of techniques will depend on two factors:

- *The type of knowledge domain represented in the ontology.* We mainly consider here two types of domains: internationalized domains, i.e., domains whose categorization usually finds consensus among different cultures, and culturally-dependent domains, i.e., domains whose categorization is normally influenced by a certain culture. On the one hand, ontologies categorized within the first domain type will require translation techniques that allow identifying direct correspondences between words. Techniques based on linguistic resources such as dictionaries, terminologies, etc., can be used in this case. On the other hand, ontologies representing a culturally-dependent domain (e.g., the judiciary), in which categorizations tend to reflect the particularities of a certain culture, will require translation techniques that allow identifying semantic correspondences.
- *The type of ontology element to be localized.* A second factor to be considered is the type of ontology elements to be localized. Depending on the ontology elements considered for localization, the algorithms of localization can be more or less complex. For example, the localization of ontology concepts and relations has a higher level of complexity than the localization of ontology instances, because a big part of the instances are represented by proper names, and have previously agreed translations or should not be translated.

The task output is a ranked set of labels in the target language for each ontology label(s).

⁴In NLP, context refers to the environment in which a word is used, and provides the information needed for figuring out the meaning of homonyms or polysemic words.

Task 4. Evaluate label translation(s)

Translation quality measurements must accomplish two basic criteria

- *Repeatable.* Two assessments of the same sample must yield similar results.
- *Reproducible and objective.* Different evaluators should arrive at a similar assessment for the same piece of translation.

The goal of this task is to evaluate label translations in the target language. At this stage, translators and/or reviewers carry out this activity taking as input the labels in the target language. The output of this task is a set of labels with its corresponding evaluation. Different linguistic criteria can be used for the evaluation of label translations. We propose two levels of evaluation criteria and for each level a set of tests, which should be automated as far as possible.

- *Semantic fidelity evaluation.* The aim of this evaluation criterion is to control that the label translation is conceptually equivalent to the ontology label in the source language. A way of evaluating the semantic fidelity is to perform a backward translation test, which provides a quality-control step demonstrating that the quality of the translation is such that the same meaning is derived when the translation is moved back into the source language.
- *Stylistic evaluation.* The aim is to control the clarity and syntax of the target language, which depends on the style of the source language and on the features of the individual idiolect. Special attention should be paid to certain stylistic aspects (e.g., “transport service” instead of “service of transport”), misspellings and typos (e.g., “women” instead of “woman”, “ig” instead of “big”, etc.).

Task 5. Ontology update

The goal of this task is to update the ontology with the label translations obtained for each localized label. The Localization Manager/QA Department carries out this task taking as input the selected label translations. The activity output is an ontology enriched with labels in the target language associated with each localized term. The ontology enrichment can follow two different modelling options. If only labels in different languages are to be included in the ontology, we can make use of the `rdfs:label` and `rdfs:comment` properties of the OWL language (Model 1). If, on the other hand, the final application demands further linguistic data than just labels, an external model capturing linguistic descriptions can be associated to the ontology (Model 2). These models were introduced in section 4.2. The

choice of the modelling option for the linguistic information will be mainly determined by two factors:

- The type of domain of knowledge represented by the ontology, and,
- The amount of linguistic information required by the final application

Taking these variables into account, we envision the two following scenarios [Espinoza et al., 2009b]:

- If the conceptualization represents a consensual domain, we can opt for the inclusion of multilingual information in the ontology (Model 1), or for the association of an external model with the ontology (Model 2). The decision between these two options will depend on the linguistic needs of the final application [Montiel-Ponsoda, 2011a]. If morphosyntactic data are needed for the purpose of information retrieval or information extraction, for example, the most suitable option will be the association of an external model. In the state of the art we find some suitable models in this sense, such as LingInfo, which enriches the ontology with morphosyntactic information, or LexInfo, which additionally accounts for the syntactic realization of ontology terms in a certain linguistic structure.
- If the conceptualization represents a culturally-dependent domain, and conceptualization mismatches among different cultures exist, we will opt for the association of an external model that permits to account for those cultural divergences at the terminological layer (Model 2). In this sense, we refer to the LIR (Linguistic Information Repository) a model that accounts for term variants within one language, and cultural divergences across languages at the terminological layer.

In both scenarios, the ontology is updated with the multilingual information resulting from the Localization Activity.

7.4 Summary of the Chapter

In this chapter we have presented the methodological guidelines that we propose to help ontology practitioners with the localization activity. These guidelines assume that users have some knowledge of ontology localization. However, the guidelines are presented so that non-experts can understand them. To the best of our knowledge, the study presented here is the first attempt to offer guidelines for the localization of ontologies.

First, we have described the objective and scope of the guidelines. We have showed that the methodological guidelines for ontology localization intended to assist the Scenario 9: Localizing Ontological Resources of the

7.4. SUMMARY OF THE CHAPTER

NeOn Methodology. Second, we have described the process used to develop the guidelines. As first step of the process, we have analyzed five relevant methodologies from the Software Localization area. Then, we have identified the main tasks of the selected methodologies. In the third step of the process we included additional tasks for covering the ontology localization requirements. As final step, we analyzed the task dependencies in order to define task order. Once the final task had been identified, we have described the inputs, outputs, and the actors involved in each one of the activities of the methodological guidelines.

Chapter 8

Experimentation

In this chapter we describe a set of experiments that were carried out with the objective of evaluating the methodological and technological aspects of the localization activity. First, we describe in section 8.1 the experiments used to evaluate some aspects related to the quality of the translation algorithm. Section 8.2 describes the study used to assess the usability of the LabelTranslator system for carrying out the ontology localization activity in distributed and collaborative environments. Finally in section 8.3 we describe two case studies to measure the understanding and usability of the methodological guidelines.

8.1 Translation Algorithm Evaluation

The localization activity (when discussed generally, and specifically for ontology engineering) is commonly suggested to give two kinds of benefits: to guarantee high productivity and outstanding quality. In this work, high productivity is understood as reducing the human effort to manually localize an ontology. Outstanding quality is here concerned with the quality of the obtained translations. The experiments described below intend to address all of these issues.

In the following sections we describe the experiments designed to measure the quality of the translation algorithm used to perform the localization activity. To select the best candidate translation, our algorithm relies on an automatic ranking method, which uses the context of an term to discern among the different meanings that an ontological label may have.

The experiments were divided into four phases, each of which is reported here as a separate study (mainly for the purpose of clarity). The first two tests were executed following a manual evaluation and using a similar group of source ontologies. These tests were performed on two different instances of the translation algorithm. Thus, the first experiment was performed with the initial version of the algorithm, whilst the second experiment was per-

formed with the current version of the algorithm implemented in this thesis. The two manual experiments evaluated different aspects of the algorithm, using different human assessment measures such as: translation quality, precision, fluency, adequacy, and correctness. The last two tests were executed using automatic machine translation metrics. This second group of tests differs from the first on the ontology corpus used for evaluation, since it uses a set of prominent ontologies with terms in different languages, which are used as reference translations. The aim of these experiments was to evaluate additional aspects of the translation algorithm as the individual quality of the translation techniques used in this work and the contribution of the resources used to obtain candidate translations.

The experiments can be summarized as follows:

1. Evaluation of quality of ranking method, completeness of the translations, and quality of compound labels.
2. Evaluation of translation quality and identification of error types.
3. Evaluation of the individual quality of translation techniques.
4. Evaluation of the contribution of translation resources.

These experiments allowed us to assess hypothesis H1, H2, H3 and H4 of this thesis (see section 3.4).

8.1.1 Quality Evaluation of Ranking Method

To evaluate the quality of translation of the LabelTranslator system we conducted in March 2008 a preliminary experiment [Espinoza et al., 2008a] involving 7 PhD students. Most of them were computer science students whose background included databases, software engineering, AI, and some experience in ontology engineering. Although none of the students had English as its mother tongue, they had a very good level of written and spoken English. One of the evaluators had a fluid level of German.

The tool that we used to perform this experiment was relying on different linguistic resources 1) remote lexical databases as WordNet, 2) multilingual dictionaries as GoogleTranslate, Babelfish, and FreeTranslation, and 3) other lexical resources as IATE. In this version, the NeOn ToolKit was used for storing the multilingual information related to a specific ontology label. Also, to disambiguate the candidate translations, we used a relatedness measure based on glosses, which compares the senses associated to each possible translation and their context.

The main goal of the experiment was to evaluate the translation ranking algorithm used by the system to select the most appropriate translation for each ontology label. This was done on the basis of the comparison of the

8.1. TRANSLATION ALGORITHM EVALUATION

translations provided by an expert (gold standard) with the translations provided by the ranking algorithm used in LabelTranslator.

The ontology corpus used for the evaluation was selected from the set of KnowledgeWeb [Corcho et al., 2006] ontologies used to manage EU projects. All selected ontologies are monolingual and their labels are defined in English. The corpus statistics are given in Table 8.1.

Table 8.1: Ontologies corpus statistics.

<i>Ontology</i>	<i>Number of Ontological Terms</i>			<i>% Compound labels</i>	
	<i>concepts</i>	<i>attributes</i>	<i>relations</i>	<i>≤3tokens</i>	<i>>3tokens</i>
Documentation &Meeting	42	61	22	44%	25.6%
Person&Project	25	18	12	47.2%	10.9%
Organization	10	7	11	46.4%	7.1%
Office	20	12	8	12.5%	0%
University	30	10	12	17.3%	0%

Analysis and Discussion

We evaluated in particular three aspects of the algorithm: the quality of the output when the algorithm automatically suggests a translation, the quality of all the set of translations, and the quality of translation of the compound labels. Based on these aspects we define some metrics to see whether the algorithm used by the LabelTranslator system facilitates the automatic localization of ontologies between different natural languages.

In all the experiments a reference translation (gold standard) provided by the evaluators was used. The “gold standard” allows users to compare the quality of the translations provided by an expert with the translations provided by the algorithm.

Next, we give an overview of each experiment and show the obtained results.

Experiment 1: Quality of the Ranking Method.

In order to evaluate the quality of the output of the ranking method in automatic operation mode we proposed a measure of accuracy. The accuracy measures the capacity of the algorithm of translation to get, in an automatic way, a correct translation according to its context. To measure the accuracy of the algorithm, we counted the number of times the first translation was correct.

$$accuracy = \frac{\text{number of times where the first translation is correct}}{\text{number of labels of the ontology}}$$

Experiment 2: Completeness of the Translations.

The previous evaluation does not allow for the checking of the completeness of the translations, since it does not observe the behavior of all the translated labels. Thus, we have measured *precision* as the number of correct translations of all the translations provided by the system and divided by the total number of translations provided by the system. To measure the *recall*, we divided the number of correct translations of all the translations provided by the system into the number of correct translations (provided by the gold standard). To calculate both measures each evaluator identifies for each ontology label which one is a correct translation.

$$precision = \frac{\text{number of correct translations of all the provided by the system}}{\text{total number of translations of all the provided by the system}}$$

$$recall = \frac{\text{number of correct translations provided by the system}}{\text{number of correct translations}}$$

Experiment 3: Translation Quality of Compound Labels.

In order to measure the quality of the translation of compound labels we proposed a subjective 1-5 score for adequacy and fluency. Adequacy is used to evaluate the quantity of the information existent in the original text that a translation contains. Commonly fluency refers to the degree to which the translation is well-formed according to the grammar of the target language. In this experiment, each evaluator assigned fluency and adequacy ratings for each translated label. The adequacy and fluency scores of two evaluators for each sentence were averaged, and an overall average adequacy and average fluency score was calculated for each evaluated ontology.

$$adequacy = \text{percentage of obtained values in each category of adequacy}$$

$$fluency = \text{percentage of obtained values in each category of fluency}$$

The used criteria to interpret the measures above described were: for accuracy, precision, and recall measures we expect to obtain values near one. For adequacy and fluency the values ranges from one to five (with one being the lowest grade and five the highest). The scales in each case are given in Table 8.2:

In order to collect the measurements, a set of interfaces were implemented. These interfaces allowed the gathering of the subjective scores and parameters of each ontology label's translations. All data were stored in memory and the results exported to a file. Each elementary experiment was performed only one time per participant. Also, in order to run the experiment it was necessary for one person to supervise.

8.1. TRANSLATION ALGORITHM EVALUATION

Table 8.2: Five point scale for fluency and adequacy measures.

<i>Scale</i>	<i>Adequacy values</i>	<i>Fluency values</i>
5	All	Flawless Spanish/German
4	Most	Good Spanish/German
3	Much	Non native Spanish/German
2	Little	Disfluent Spanish/German
1	None	Incomprehensible

Identified Strengths and Weaknesses

Table 8.3 and Table 8.4 show the results achieved by the first prototype in each experiment for translating an ontology in English to respectively Spanish and German. It is worth mentioning that for ontologies *Documentation* and *Person&Project*, it was not possible to define reference translations in German, provided by an expert; so these ontologies were not considered in the experiment with this language.

All the percentages of adequacy and fluency shown in this table correspond to those translations punctuated with a value greater than 4. The experimental results showed that our system suggested the correct translation 72% of the times. Also, the values of recall obtained suggested that a high percentage of correct translations were part of the final translations shown to the user.

Table 8.3: Results obtained in the three experiments for Spanish.

<i>Ontology</i>	<i>English to Spanish translation</i>				
	<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>	<i>Adequacy</i>	<i>Fluency</i>
Documentation	0.51	0.47	0.39	68%	75%
Person&Project	0.73	0.35	0.81	89%	93%
Organization	0.81	0.41	0.78	87%	95%
Office	0.79	0.49	0.77	93%	95%
University	0.80	0.36	0.87	96%	93%

Table 8.4: Results obtained in the three experiments for German.

<i>Ontology</i>	<i>English to German translation</i>				
	<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>	<i>Adequacy</i>	<i>Fluency</i>
Organization	0.73	0.33	0.64	73%	69%
Office	0.78	0.34	0.74	67%	76%
University	0.71	0.23	0.71	69%	73%

Moreover, the obtained results in each metric helped us to analyze which components need improvement. The main limitations discovered were:

- *The translation service is highly dependent on the types of resources used and their domain coverage.* The worst values of precision and recall were obtained by the documentation ontology, because the domain of this ontology is only partially covered by the resources used for the translation.
- *The lack of learning of new lexical patterns limits the scalability of our tool.* The percentages of adequacy and fluency obtained for English-German compound label translations are in general lower than the percentages of the English-Spanish ones. Our explanation is that a major effort was put (in the first version of our tool) to learn templates between English-Spanish languages. However, this situation can be improved by allowing users to provide, in runtime, new lexical templates when these do not yet exist in any repository.

8.1.2 Translation Quality Evaluation

Once the results of the first experiment were analyzed, we implemented some improvements to the algorithm to solve the main limitations.

For the second version of the translation ranking method, we decided to carry out the disambiguation using the senses of each translated label and the senses of the labels, instead of using glosses. The senses in both cases were built with the information retrieved from matching terms in ontological resources (see section 6.8 for more details).

To improve the translation of compound labels, we implemented a new algorithm to support the translation of these labels, independent of the number of tokens. Also, for this version we incorporated new resources to extract candidate translations and the LIR model for storing the multilingual information associated to each ontology element. In particular, we incorporated as translation resources, the multilingual dictionary Wiktionary, the multilingual lexical database EuroWordNet and semantic resources indexed by Watson or available on the Web.

An additional difference is that the previous implementation of Label-Translator used a single working scenario, where only one person performs all the major steps required for localizing an ontology to (and from) different natural languages. This scenario is feasible only in some cases. However, it is very difficult for a person to update all the linguistic information associated with a particular concept. To address these limitations, we decided to add a workflow-based model for the collaborative localization of ontologies in distributed environments and describe the components required to support it. Although this version of our tool already had this functionality, to perform this test we used the non-collaborative mode.

In order to test the implemented improvements a new experiment [Dzbor et al., 2009] was carried out in the “Artificial Intelligence (AI)” master

course at the Facultad de Informática (Universidad Politécnica de Madrid) with 17 master students, having background in databases, software engineering, and artificial intelligence, but no extensive practical experience in ontology engineering. The language used in the course was English, although none of the students had English as its mother tongue. At the time of the experiment, students had received a broad introduction to the Semantic Web, and had been taught on theoretical and practical aspects of ontologies and ontology languages (RDF and RDF Schema), methodologies for the development of ontologies (specifically the NeOn Methodology) and some aspects of computational linguistics (terminology and multilingualism in ontologies).

We decided to use a questionnaire that allows to measure the translation algorithm's capacity to provide correct translations according to the context.

Analysis and Discussion

For this experiment we selected two ontologies from the set of Knowledge Web¹ ontologies used in our first attempt to measure the quality of translation. The selected ontologies *Documentation* and *Person&Project*, registered the worst values in the quality of the output of the ranking method. Therefore, our goal is to re-evaluate the quality of translations on these ontologies, but using the new algorithm implemented in the second version of our ontology localization system.

The two selected ontologies are in English and our aim was to localize them into Spanish. In this experiment we decided to use a questionnaire that allows to assess how well the translation ranking works, selecting correct translations according to the context. In addition, the questionnaire included questions to classify the types of errors found in the translation of simple and compound labels. The questions used to evaluate the quality of the translations deal with the weaknesses found in our first evaluation (see section 8.1.1).

The rationales for discarding the evaluation metric scores used in the previous experiment and the decision to use a simple questionnaire to assess to quality of translations were:

1. Separate scales for fluency and adequacy were used under the assumption that a translation might be disfluent but contain all the information from the source. However, as pointed out by the authors in [Koehn and Monz, 2006], it seems that people have a hard time separating these two aspects of translation. In fact, the correlation found in our tests, between evaluators' fluency and adequacy scores indicate that the distinction might be false.

¹<http://knowledgeweb.semanticweb.org/semanticportal/sewView/frames.html>

2. The manual computation of precision and recall proved to be too costly in time and effort. For example, the determination of the number of correct translations provided by the system required several seconds on average for each label. As a consequence, the test duration would have taken several hours, caused some discomfort among reviewers.

The questionnaire included the following questions:

1. Are the translations in the target language correct? If not, can you mark the level of correctness? 30%, 50%, 70%, 90%, other
2. If they are not correct, what are the types of errors, in your opinion?
 - Lack of the correct equivalent
 - Errors in lexis/terminology
 - Errors in syntax/style
3. Are the compound labels translated correctly? If not, what are the main problems encountered?

After a short introduction about the Ontology Localization activity, we gave some instructions about the questionnaire. Since the notion of correctness is not intuitively clear in the context of MT then, to clarify question one we proposed to the reviewers a definition of a correct translation. For our purposes a correct translation is a label that effectively transfers the information from a source language to a target language (preserving the meaning of the input term) and without errors. We explained that when the two conditions (mentioned in the definition) are satisfied, then, the translation can be labeled as correct.

To properly categorize the errors (question two), we offered a translation error classification taking as core the work presented in [Flanagan, 1994], which proposes more than twenty categories based on the observation of the most frequent error types in MT outputs. We grouped these categories in the three groups included in question two. The first category concerns the preservation of the meaning of the label translated. The second category concerns the use of the vocabulary, and the third the grammatical and the linguistic accuracy of the machine translated texts, including the format and style of the produced translations.

Students were provided with illustrative examples of possible errors in each category. Finally, students were informed that the compound labels needed to be analyzed separately (question three).

The eight groups of master students (conformed by 1 or 2 persons) performed the ontology localization following the new algorithm used to rank the translations. For each ontology term, students compared the quality of

8.1. TRANSLATION ALGORITHM EVALUATION

the translations provided by the algorithm to the expected translations (according to the ontology domain). Once students had completed the exercise, they were asked to fill out the questionnaire.

Findings and Observations.

In this section, we provide some findings extracted from the analysis of the experiment results.

From the second experiment, in which students evaluated the quality of the translation obtained from the new translation-ranking algorithm, we can mention the following observations:

- As shown by Figure 8.1 the percentage of labels considered translated correctly for the *Documentation* ontology was high, between 70% and 85%, with an average of 81%. For the *Person&Project* ontology, 85% of the labels (on average), were marked as correct translations. If we distinguish the quality of translations between simple and compound labels, we can comment that in the case of the first ontology, 31 out of a total of 38 (81%) simple labels were considered translated correctly. The average translation quality of compound labels, in the same ontology was 80%. In the case of the *Person&Project* ontology, the translation qualities were on average 75% and 91% for simple and compound labels, respectively.

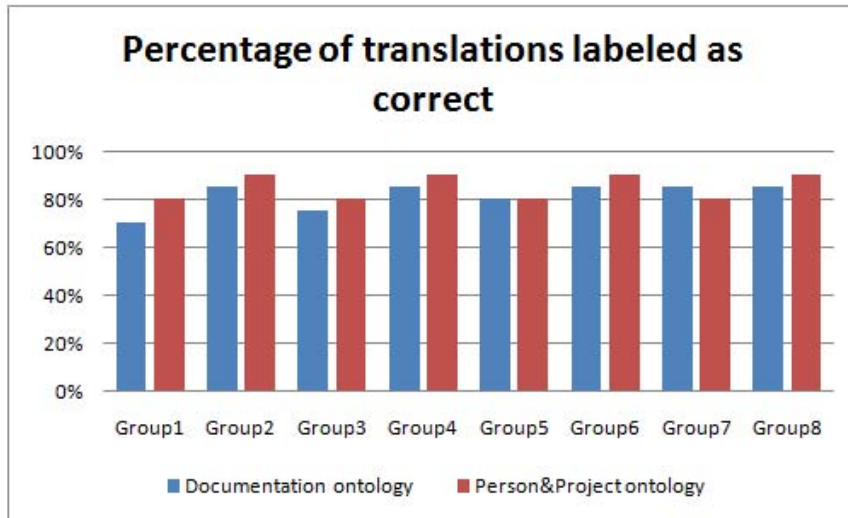


Figure 8.1: Level of correctness in label translations.

- Regarding the results of question two, it should be noted that of the 180 labels automatically translated by the algorithm and belonging to the two ontologies, only 32 were marked as incorrect translations. In

general, there was consensus among the different groups of evaluators with regard to the labels considered mistranslated. Of these 32 translation errors, 12 correspond to simple labels and 20 to compound labels. When trying to find common problems in the translation of simple labels, two main problems were identified, as illustrated in Figure 8.2: 83% of errors found in the translation of these labels correspond to errors in the terminology used in the translation and 17% of the errors correspond to problems in the lack of a correct equivalent in the target language.

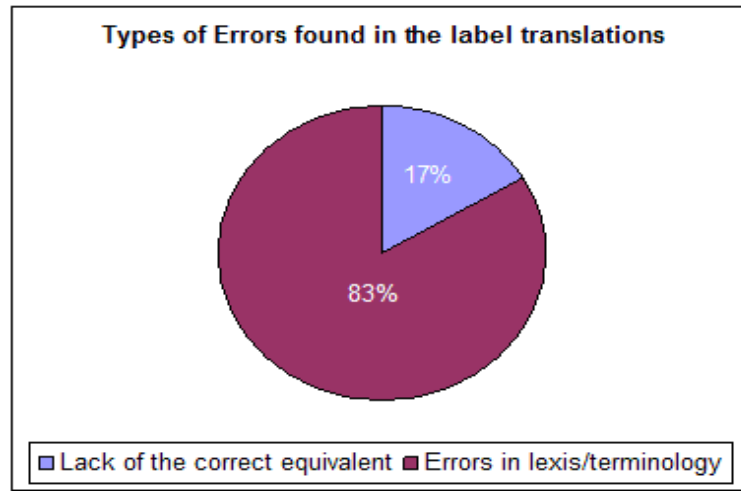


Figure 8.2: Type of errors found in the translation of ontology labels.

- Finally, to the question “Are the compound labels translated correctly?” the majority of the students believed that the quality of translation of the compound labels was correct. However, they reported errors in those labels that contained tokens with acronyms, for example, “Workshop_URL”, “EPMB_Meeting_Minutes” or “EC_Templates”. Also, the evaluators reported few problems of syntax, such as the misuse or omission of the definite article, wrong prepositions, or wrong personal pronouns.

Concluding Remarks

Comparing these results to the previous experiment we may note an increase in the accuracy of the algorithm to select correct translations. The accuracy of the algorithm in the case of the *Documentation* ontology was 20% higher than the results obtained in the first experiment, and approximately 8% better in the case of compound labels. Similar values were obtained for the *Person&Project* ontology, with the accuracy increasing from 21% to

8.1. TRANSLATION ALGORITHM EVALUATION

25% in the translation of the ontology, and for compound labels the same values of accuracy as those obtained in the previous experiment were maintained (91%).

Basically, there was a significant improvement in the translation of the compound labels; this is not surprising, because in the initial translation algorithm we only focused on the definition of lexical templates for compound labels of two or three tokens. The goal of the lexical templates is to produce high quality translations; however, for those compound labels with more than three tokens the algorithm relied directly on the output of the different resources of the used translations. In the current version of the algorithm we implement two improvements:

- a recursive function that attempts to match the bi/tri-tokens of a compound label with the lexical templates² stored in the database, or
- a method that learns new lexical templates from the translations supplied by the user.

Additionally to the previous comments, it is important to mention that high precision values obtained in this test were closely related with an improvement in the algorithm, but also with the inclusion of new translation resources. In fact, resources as Wikitionary, EuroWordNet, and other ontologies available on the Web reduced the lack of appropriate translations to specific-domain terms.

8.1.3 Translation Techniques Evaluation

The main motivation of this experiment was to evaluate the quality of a sample group of translation techniques proposed in this thesis for the automatic localization of an ontology. Based on our quantitative evaluation, we will provide substantial evidence to assist in the proper selection and combination of these translation algorithms, which is the first step towards the proper use of them.

As described in Chapter 5, there are a large variety of translation techniques that can be used to localize an ontology to a linguistic level. Their translation quality varies from language to language and from text to text. Our interest here is to examine their quality in the domain of ontological terms translations. Also, it will be interesting to know whether the strategies proposed in this work are doing its job: which is selecting the best translations.

²The notion of lexical template refers to text correlations found between a pair of languages.

Analysis and Discussion

First, we seek to evaluate the translation strategies used in this work (see section 6.8), separately from its implementation of specific translation techniques. However, since translation strategies perform no translations on its own, but only combine and choose among the translations provided by the translation algorithms, it is necessary to include these algorithms in order to perform the evaluation. For this reason, we will simultaneously evaluate the translation strategies as a whole as well as each translation technique on its own. By comparing the results, we hope to shed light on the contribution of the strategies and techniques proposed in the whole translation process.

To evaluate the quality of some of the techniques proposed in this thesis, we analyze the contribution of each one of the components that are part of the translation strategy for simple labels. For this case, the translations to be evaluated will come from three sources. For each source ontology label to be translated, we will first translate it using the strategy used for the translation of simple labels as it is currently designed (see section 6.8.1). Then, we translate the ontology label again using each of the separate translation methods that are part of our translation strategy. The strategy used for this experiment employs two translation components:

- *First component.* This component uses different translation methods (*dictionary-based*, *thesauri-based*, and *online MT systems*) to obtain candidate translations.
- *Second component.* In this case the translation process is performed through the comparison of ontology or taxonomy structures (*ontology structure comparison*) containing source and target labels.

In section 6.8.1, we analyzed these translation components in more detail, focusing on those aspects that could contribute to the localization of ontologies.

In the strategy used for translating simple labels, each translation method is permitted to hypothesize multiple translations for each ontology label. The multiple overlapping translations are sorted out using the context of the source term and a linear combination approach. These methods perform a search over the set of candidate translations to find the subset that exactly covers the input and yields the best translation. When testing the translation methods separately from the translation strategy, other means must be employed for sorting out the candidate translations. For each of the three translation sources, the following describes the technique used to obtain a final translation among the candidate translations offered by each method.

- *Translation strategy for simple labels.* In this case we use the optimal translation as proposed by the strategy implemented in this work.

8.1. TRANSLATION ALGORITHM EVALUATION

- *First component.* As this component is based on different methods, we evaluated each one independently. Final translations were obtained taking into consideration the ranking proposed by each of these methods, without any additional process.
- *Second component.* This approach needs as input a list of words in order to discover the senses of each translated label, for which we use all candidate translations provided by the first component. In this case, no method of disambiguation was used to avoid an explosion of nuisance candidate translations. The final translations were obtained using the ranking method to sort the list of translations according to similarity with the structural context of the label to be translated.

Next we describe the metrics selected to evaluate a sample group of translation techniques used in this work.

Experiment Design

Unlike the approach of evaluating MT quality by human judgment used in the previous tests, this evaluation adopts the state-of-the-art automatic quantitative MT evaluation technology. Although we recognize that manual evaluation has different advantages compared to automatic evaluation, automatic evaluation has some peculiarities we cannot ignore. One of the most important peculiarities is that a good evaluation should provide the same evaluation values for two perfectly equal texts (even if they are the output of different translation systems): yet two human evaluators that have to judge the same text could give two different evaluations, as might the same evaluator at different moments in time [Fordyce, 2007]. For this reason, we decided to use the following evaluation metrics: BLUE, NIST, WER, Meteor, and NGram to perform this test. These metrics were classified and explained in Chapter 4, section 4.4.2.

Baseline scores were generated using the Phrasal system [Cer et al., 2010] trained with the EuroParl dataset [Koehn, 2005]. Phrasal is a state-of-the-art phrase-based machine translation system, which provides an easy to use API to implement new decoding model features. The Europarl dataset is extracted from the proceedings of the European Parliament. It includes versions in 21 European languages: Romanic (French, Italian, Spanish, Portuguese, Romanian), Germanic (English, Dutch, German, Danish, Swedish), Slavik (Bulgarian, Czech, Polish, Slovak, Slovene), Finni-Ugric (Finnish, Hungarian, Estonian), Baltic (Latvian, Lithuanian), and Greek.

We choose to make the evaluation on the following public domain and specific multilingual ontologies³ (see Table 8.5) used in the Monnet project⁴

³This multilingual ontologies can be found at <https://subversion.deri.ie/monnet-wp5/Benchmark/MultilingualOntologies/>

⁴<http://www.monnet-project.eu/Monnet/Monnet/English?init=true>

for the investigation of hybrid methods used in domain training for term translation.

Table 8.5: Multilingual Ontologies used in the evaluation.

<i>Ontologies</i>	<i>Number of Labels by Language</i>	
	<i>English</i>	<i>Spanish</i>
FOAF	89	84
DOAC	47	37
GEOSKILLS	90	67

To perform this experiment we used the evaluation framework implemented in the Monnet project to evaluate the quality of the designed translation components. This tool is available as an OSGi bundle and can be executed as a Web application. A test cycle can be summarized as follows:

- A translation technique is chosen for evaluation
- Then, the ontology that needs to be localized can be selected
- Finally, the source and target languages are chosen.

In the evaluation framework, each generated translation is compared to its corresponding reference translation (extracted from the own multilingual ontology) and then translation metrics are computed for each label. For the evaluation, we explored the localization of the ontologies listed above from English to Spanish.

Results and Discussion

In this section we present statistics comparing each one of the individual components used for translating simple labels with the translation strategy for simple labels, which combines these components. Table 8.6 shows the scores of the automatic evaluation metrics of the individual components that are part of the translation strategy for simple labels. In the case of the first component, we also present the automatic metric scores of each individual translation technique that forms part of this component. Each score shows, in a comparative way, how well an ontology is translated from a particular language by a particular translation technique.

In general, all translation techniques' performance on translating domain ontologies are relatively similar. As from the results shown in table 8.6, we see that no technique individually overcomes the performance of the proposed translation strategy for simple labels (see the last line in the table for each ontology). The previous analysis provides positive evidence for hypothesis H4, which shows that an appropriate combination of translation methods leads to better localization results than only using one at a time.

8.1. TRANSLATION ALGORITHM EVALUATION

After analyzing the scores of each component, we concluded that, although the quality of the first component of translation is very high, the second component contributes by means of the disambiguation of senses to improve the results. It suggests that the exploration of the context for each ontology term might be beneficial for the selection of the best translations.

Table 8.6: Translation Techniques Comparison.

Translation sources	Ontologies	Simple Labels	Automatic Machine Translation Measures				
			BLEU	BLEU-2	METEOR	NIST	WER
<i>First component</i>	Multilingual Friend Of A Friend (FOAF) multilingual_foaf.owl	47	0,00000	0,17578	0,19935	2,44925	0,85532
dictionary-based			0,00000	0,09845	0,11165	1,36345	0,47905
thesauri-based			0,00000	0,11635	0,13195	1,61135	0,56615
online MT systems			0,00000	0,17184	0,19488	2,39471	0,83616
<i>Second component</i>			0,00000	0,17723	0,20099	2,45619	0,86238
<i>Translation strategy</i>			0,00000	0,17900	0,20300	2,47900	0,87100
<i>First component</i>	Description of a Project (DOAP) doap.owl	22	0,00000	0,14494	0,20607	1,95425	0,84007
dictionary-based			0,00000	0,08741	0,12427	1,17849	0,50659
thesauri-based			0,00000	0,10211	0,14517	1,37669	0,59179
online MT systems			0,00000	0,13671	0,19437	1,84326	0,79236
<i>Second component</i>			0,00000	0,14668	0,20854	1,97764	0,85013
<i>Translation strategy</i>			0,00000	0,14700	0,20900	1,98200	0,85200
<i>First component</i>	GeoSkills (GEO) geoskills.owl	32	0,00000	0,17551	0,17058	3,10590	0,97121
dictionary-based			0,00000	0,10584	0,10286	1,87297	0,58568
thesauri-based			0,00000	0,12364	0,12016	2,18797	0,68418
online MT systems			0,00000	0,16554	0,16089	2,92950	0,91605
<i>Second component</i>			0,00000	0,17620	0,17125	3,11819	0,97505
<i>Translation strategy</i>			0,00000	0,17800	0,17300	3,15000	0,98500

8.1.4 Translation Resources Evaluation

This section is devoted to the description of an experiment that we performed with the purpose of evaluating the contribution of translation resources on the quality of the obtained translations.

Taking into account that each translation method relies on different resources to provide candidate translations that preserve the original meaning of each ontological term into the target language, we are also interested in evaluating the contribution of these resources.

Analysis and Discussion

To evaluate the contribution of the different resources used to obtain candidate translations, we analyzed the strategy provided in this work for translating compound labels (see section 6.8.2). Let us remember that this approach uses a hybrid composition of two translation components. The first component is the same as the first component used for translating simple

labels. The second component relies on a set of lexical templates derived from different ontologies to control the order of translation.

Using the first component of our translation strategy, a set of possible translations is obtained for each token of the compound label. In the original approach, the translation strategy for compound labels uses all possible combinations of obtained translations for each token without discarding any. Therefore, we need to evaluate the contribution of each individual resource in the generation of high quality translations.

Let us remember that the first component of the translation strategy used for compound labels relies on three translation methods: *dictionary-based*, *thesauri-based* and *online MT-systems*. The first method uses Wiktionary as translation resource. The thesauri-based method relies on IATE and three different resources are the core of the online MT-system method: Altavista, Bing, and Google. For the translation method based on online-MT systems each resource was evaluated independently.

Experiment Design

For the execution of this experiment, the same metrics, ontologies and even the same evaluation framework introduced in the previous section were used. For each of the methods and its corresponding resources, the evaluation framework calculated the values of each metric, by comparing the obtained translations to the reference translations.

Results and Discussion

In Table 8.7, we show statistics about the contribution of different resources to the translation of compound labels. The first lines of the table (for each ontology) contain the scores of the evaluation metrics of the translation strategy for compound labels (TS4CL). It is important to emphasize that TS4CL (in this case) uses all the candidate translations retrieved from all resources available to build final translations in the target language using lexical templates. The other lines of the table represent the obtained evaluation results by each individual resource (see TS4CL+Google by example). In this case the translation strategy is the same, but with the difference that the inputs to the lexical templates come from a single resource.

The obtained results suggest that the individual contribution of each resource is less than the contribution of all the resources together. This analysis provides positive evidence for hypothesis H3, in the sense that the use of more than one resource into the translation process gives a wider range of translation candidates to choose from, and the correct translation is more likely to appear in multiple translation resources than a in single translation resource.

8.2. COLLABORATIVE LOCALIZATION EVALUATION

Table 8.7: Translation Resources Comparison.

Resources used	Ontologies	Compound Labels	Automatic Machine Translation Measures				
			BLEU	BLEU-2	METEOR	NIST	WER
Baseline (TS4CL)	Multilingual Friend Of A Friend (FOAF) multilingual_foaf.owl	37	0,00000	0,17542	0,19894	2,42942	0,85358
TS4CL + Altavista			0,00000	0,08771	0,09947	1,21471	0,42679
TS4CL + Bing			0,00000	0,12455	0,14125	1,72489	0,60604
TS4CL + Google			0,00000	0,12104	0,13727	1,67630	0,58897
TS4CL + IATE			0,00000	0,12104	0,13727	1,67630	0,58897
TS4CL + Wiktionary			0,00000	0,07017	0,07958	0,97177	0,34143
Baseline (TS4CL)	Description of a Project (DOAP) doap.owl	15	0,00000	0,14349	0,20401	1,93471	0,83167
TS4CL + Altavista			0,00000	0,01291	0,01836	0,17412	0,07485
TS4CL + Bing			0,00000	0,05740	0,08161	0,77388	0,33267
TS4CL + Google			0,00000	0,10044	0,14281	1,35430	0,58217
TS4CL + IATE			0,00000	0,09471	0,13465	1,27691	0,54890
TS4CL + Wiktionary			0,00000	0,05453	0,07753	0,73519	0,31604
Baseline (TS4CL)	GeoSkills (GEO) geoskills.owl	35	0,00000	0,17375	0,16887	3,07484	0,96150
TS4CL + Altavista			0,00000	0,05213	0,05066	0,92245	0,28845
TS4CL + Bing			0,00000	0,09904	0,09626	1,75266	0,54805
TS4CL + Google			0,00000	0,13900	0,13510	2,45987	0,76920
TS4CL + IATE			0,00000	0,09730	0,09457	1,72191	0,53844
TS4CL + Wiktionary			0,00000	0,07819	0,07599	1,38368	0,43267

8.2 Collaborative Localization Evaluation

In this section we focus on the evaluation of the usability of our ontology localization system.

8.2.1 Overview and Objectives

The main motivation of this experiment was to evaluate the usability of the LabelTranslator system proposed in this thesis for supporting an automated localization in distributed and collaborative environments. This experiment allowed us to asses hypotheses H7 and H8 of this thesis (see section 3.4).

In particular we evaluated the following items: i) the adequacy of the workflow model with respect to the users' actions, and ii) the overall usability of the system.

To measure the adequacy of the workflow model we analyzed if ontology stake-holders were able to perform all activities of the ontology localization life-cycle. For each possible action, we verified that it could be represented with our model and that it captured all the information required by the ontology stake-holders.

To assess the usability of the LabelTranslator system we conducted an experiment following the Software Usability Measurement Inventory (SUMI) method [Kirakowski and Corbett, 1993]. The SUMI questionnaire includes 50 items for which the user selects one of three responses: "agree", "don't know", "disagree" (see Appendix A). The following sample shows the kind

of questions that were asked:

- This software responds too slowly to inputs.
- I would recommend this software to a my colleagues.
- The instructions and prompts are helpful.
- I sometimes wonder if I am using the right command.
- Working with this software is satisfactory.
- I think that this software is consistent.

The questionnaire is designed to measure the affect, efficiency, learnability, helpfulness and control of a software product [Dumas and Redish, 1993]. SUMI is also mentioned in the ISO 9241 standard as a recognized method of testing user satisfaction [ISO, 1992].

8.2.2 Experiment Setting

The experiment involved 10 participants, most of whom were PhD students with a good command of ontology engineering. We believe that this set of participants could be considered a good representative of potential users of our method for the localization of ontologies. Before starting the experiment, we configured the collaborative infrastructure as the configuration illustrated in Figure 8.3

As we introduced in section 6.7, our approach has been implemented within the NeOn Toolkit, an extensible ontology engineering environment based on Eclipse, by means of a set of plugins and extensions. In the scenario shown in the Figure 8.3, the team of ontology users (localization manager, translators, and reviewers) will be collaboratively localizing an ontology following a well defined process (i.e. workflow). The following characteristics describe the environment:

- There is only one copy of the ontology (ontology model) which is stored in a central server. Additionally, each ontology user has its own copy located at his/her PC.
- Ontology users are working in a distributed manner (i.e. they are physically located at different PCs).
- Ontology users are working concurrently.
- Each ontology user uses his/her own NeOn Toolkit installation and connects to the central server.

8.2. COLLABORATIVE LOCALIZATION EVALUATION

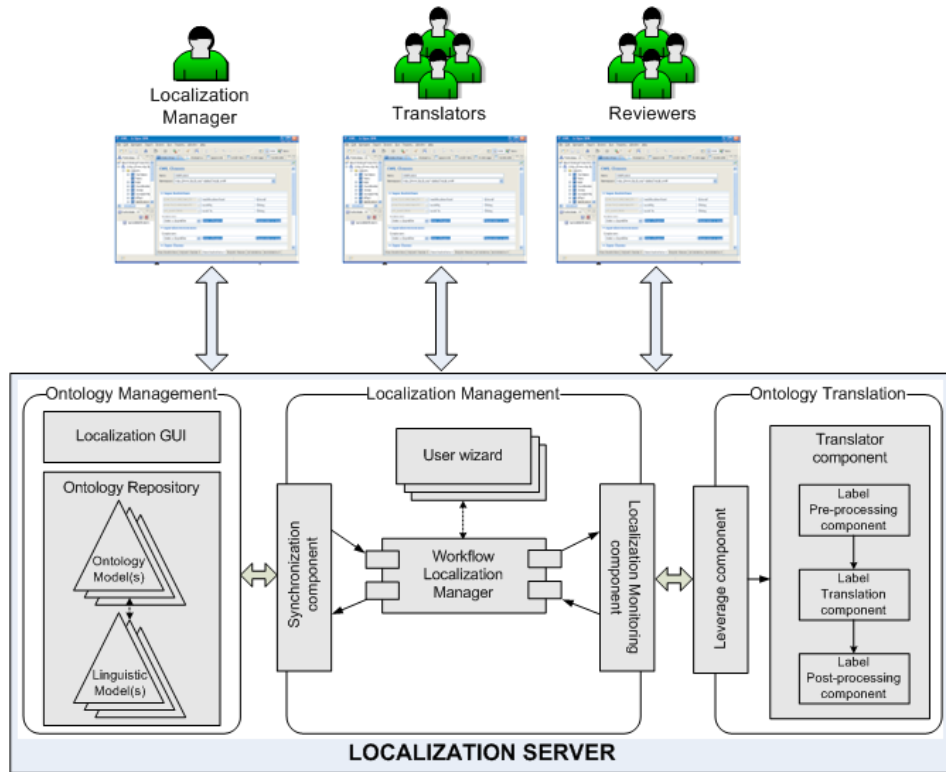


Figure 8.3: LabelTranslator Configuration for Collaborative Ontology Localization.

- Each ontology user specifies his/her credentials (e.g. name and role) in the NeOn Toolkit.
- The translations of an ontology label from all translators are stored at one specific place.
- Each NeOn Toolkit has configured that label translations are stored at a specific location (i.e. a specific localization server) and connects to it.

Regarding the configuration of machines:

- One personal computer (PC) is configured as the server. This PC has to be running the Localization Server.
- The PCs used by the ontology users are only running a NeOn Toolkit installation configured as described above

The ontology used for the localization was Documentation&Meeting, which is the documentation schema used in the Knowledge Web's Semantic

Web Portal⁵. The experimenters met with all participants for 30 minutes to explain the purpose of the evaluation session, to give a brief introduction to the system and presented the methodology of the SUMI evaluation. The 10 participants were divided in two groups composed of one Localization Manager, two Translators, and two Reviewers. All participants was provided with an user guide of the tasks s(he) had to perform according her(his) role. A detailed description of the three user guides are available in Appendix B.

To measure the adequacy of the workflow model, we requested that the 10 participants of our case study, perform every possible action according to their role, and asked them to verify that their actions were correctly executed. Each of the Localization Managers were in charge of identifying the ontology elements to be localized and distributing it to the localization team, setting up the localization team, as well as assigning and monitoring the tasks. The Translators was requested to download the selected labels to be localized and (s)he to perform the translations. The same translators upload the translated ontology labels and send them for review. The Reviewers were then requested to download the translated labels and to approve/reject them.

It is important mention here that that all the participants who had the role of Translator used the automatic translation algorithm provided by the system. All participants had 20 minutes to test the LabelTranslator system, and 10 minutes to fill out the SUMI questionnaire for user-interaction satisfaction. During the execution of the experiment, the evaluators were taking note of the behavior of the participants, their questions and problems.

8.2.3 Findings and Observations

During the experiment, we verified that none of the 10 participants had problems performing the possible actions according to his role and consequently, we can affirm that all workflow actions could be represented by our collaborative localization model. This statement can be verified through the high values of efficiency obtained in the SUMI questionnaire. The impression of the users was around 58% positive, around 34% was undecided, and only around 8% was negative. After analyzing the feedback received from the users on the negative answers and undecided answers when available, we found that in general they refer only to desired improvements in the GUI to facilitate or make some tasks more intuitive. Nevertheless, feedback also showed that evaluators were in general highly satisfied with the LabelTranslator system and that they agreed on its usefulness and correctness.

In the rest of this section, we analyze the individual goals of the SUMI method. Figure 8.4 shows the percentage values for three grades (positive, negative or undecided) of user perception with respect to the goals of each

⁵<http://knowledgeweb.semanticweb.org/>

8.2. COLLABORATIVE LOCALIZATION EVALUATION

SUMI dimension. In the following we describe the results obtained for each dimension of SUMI questionnaire:

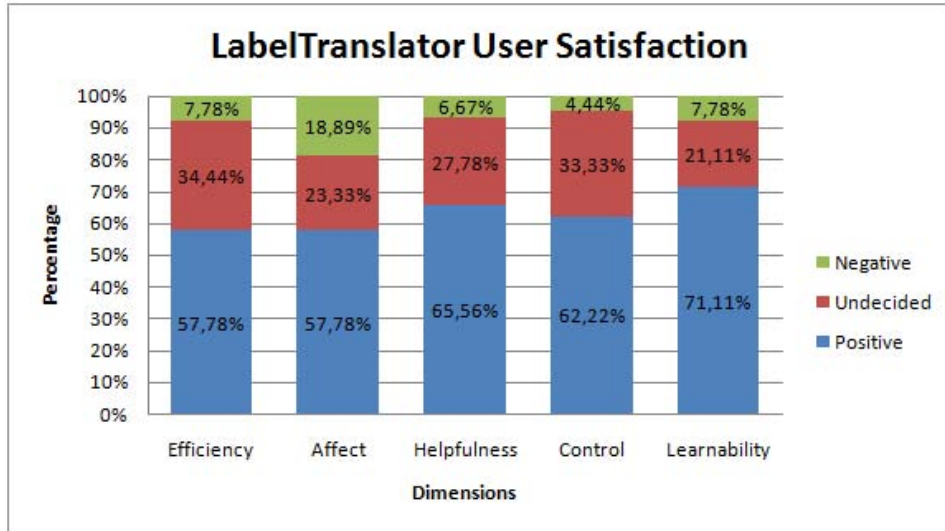


Figure 8.4: Results of SUMI Questionnaire for LabelTranslator.

- Efficiency.** After analyzing each of the 10 questions for measuring the degree to which users feel that the software assists them in their work, we found that only one question contributed in particular to the 7.78% of disagreement: “I sometimes don’t know what to do next with this software”. This means that the majority of the users did not have problems using the tool. Moreover, we found that the two questions that most contributed to 34.44% of indecision were: “If this software stops, it is not easy to restart” and “I find that the help information given by this software is not very useful”. However, this situation can be taken positively because it means that users did not have the opportunity to check these events.
- Affect.** The affect dimension measures the user’s general emotional reaction to the software - it may be glossed as Likeability. For this dimension we found that the question that most contributed to 18.89% of disagreement in the user’s general reaction to the software was: “I feel safer if I only use a few familiar commands or operations”. We believe that we must improve this aspect of the system, so that all functionalities can be perceived with the same degree of positivity by its users.
- Helpfulness.** 65.56% of the users believe that the software is self-explanatory (helpful). Moreover, we found that the question that

mostly contributed to 27.78% of indecision was: “This software is awkward when I want to do something which is not standard”. This means that the majority of the users did not have the need to find alternative options to perform the available actions in the system.

- *Control.* We consider that the evaluation of the degree to which the user feels that (s)he, and not the product, is setting the pace, is satisfactory, because we only obtained 4.44% disagreement. In the same sense, 33.33% of indecision, corresponds to aspects that did not appear in the software such as “Error prevention messages are not adequate”, which is positive.
- *Learnability.* measures the speed and facility with which the user feels that (s)he has been able to master the system, or to learn how to use new features when necessary. 71.11% of the users coincided in that the software i) it has a very attractive presentation, ii) it is relatively easy to move from one part of a task to another, iii) it is not necessary to look for assistance to use the software.

8.2.4 Identified Strengths and Weaknesses

Taking into consideration that this is the first implementation (at the best of our knowledge) of a complete infrastructure that addresses the collaborative ontology localization of ontologies, we claim that the results are highly encouraging and motivational. In particular, the results provide an indication of the real value and practical usability of the collaborative workflow proposed in this work. Nevertheless, we need additional experiments and more users to draw full conclusions.

The most important findings of the experiment are related with the high level of learnability shown by LabelTranslator, especially in the case of a novice user. There was only one evidence about the need of making minor modifications in the LabelTranslator user interface to improve affect and efficiency with better navigation and informative functions. These aspects have been taken into consideration for the version of the system described in this thesis.

8.3 Methodological Evaluation

In Chapter 7 we described in detail the methodological guidelines for supporting the Ontology Localization activity. Our concern here is to describe two cases that can show the effects and benefits of using ontology localization guidelines. There are several different aspects of guidelines that need to be studied and several types of effects of guidelines usage that need to be defined and measured. So far no indisputable evidence has been put forward

8.3. METHODOLOGICAL EVALUATION

to support the benefits of using ontology localization to build a multilingual ontology. Only in the software engineering field, where the localization is used to adapt a software product to a specific region or language, can we find some evidence of the benefits of localization.

8.3.1 Usability of the Methodological Guidelines

In this section, we propose an experiment to learn about the understandability and usability of the methodological guidelines for carrying out the ontology localization activity. The main goal of the case study is to test the benefits of using the proposed methodological guidelines and additional material included in Chapter 7 for obtaining a multilingual ontology as output of the ontology localization activity. This experiment allowed us to assess the hypothesis H5 (see section 3.4).

Assumptions and user study setup

In this experiment we proposed a questionnaire about the methodological guidelines for the ontology localization activity, to be answered by people carrying out the experiment. The experiment was carried out in the “Artificial Intelligence (AI)” master course at the Facultad de Informática (Universidad Politécnica de Madrid) with master students, having backgrounds in databases, software engineering, and artificial intelligence, but no extensive practical experience in ontology engineering. We proposed a questionnaire about the use of methodological guidelines for ontology localization activity. Figure 7.4 shows the workflow corresponding to such guidelines. To interpret the results, we analyzed the answered questionnaires and extracted some statistics.

The questionnaire includes the following questions:

1. Are the proposed guidelines well explained?
2. Is more detail needed in the guidelines? If so, please explain in detail in which sense and in which tasks
3. Are these guidelines complete? If not, what is missing?
4. Do you think more techniques and tools should be provided?
5. How can we improve the proposed guidelines?
6. Did you find these localization guidelines useful?

The experiment was divided in the following phases:

- Lecture will provide to students the proposed guidelines.

- Student groups followed the methodological guidelines to carry out the ontology localization activity. Students had two weeks for carrying out the experiment using the provided material.
- Students documented in detail each task proposed in the methodological guidelines and performed during the ontology localization activity.
- Students filled out a questionnaire about the proposed guidelines.

Analysis and Discussion

The experiment included six questions about localization guidelines solved by 15 students, and as a general conclusion we can say that students did not have problems with the use and understanding of each one of the tasks identified in the methodological guidelines. In the following, we provide some observations extracted from the analysis of the experiment results:

- 95% of the comments provided by the students to question 1 indicated that the guidelines were well explained.
- For the comments obtained to question 2: “Is more detail needed in the guidelines?”, we can say that 85% of the students consider that more detail is not necessary in the guidelines, however 15% think there is an opportunity to improve the explanations of i) how to select the most appropriate linguistic assets (step 1 in the guidelines), and ii) how to obtain the ontology term translations (step 3 in the guidelines).
- In question 3: “Are the guidelines complete?”, 95% of the evaluators believe that the guidelines to perform the localization activity are complete. However 5% consider it necessary to enhance the guidelines to support the evaluation of the obtained translations.
- For the comments obtained to question 4: “Do you think more techniques and tools should be provided?”, we can say that all evaluators believe that the techniques and tools to execute each activity of the guidelines are sufficient.
- The generalized comment to question 5: “How can we improve the proposed guidelines?” is to include more examples of how to use the proposed guidelines for the ontology localization activity and what results are to be expected.
- Finally, with respect to question 6: “Did you find these localization guidelines useful?”, all students believed that the guidelines were useful, but also necessary.

8.3. METHODOLOGICAL EVALUATION

Identified Strengths and Weaknesses

Based on the comments obtained in the experiment we can say that the majority of the students found that the methodological guidelines were useful and understandable. The main weaknesses included a more complete description of some tasks of the methodology. Some examples are:

- A more detailed description of the criteria to choose a technique to help in the localization activity.
- The lack of basic guidelines to select a localization tool depending on the type of ontology to be localized, or,
- An exhaustive description of the different levels of difficulty that can be found in the translation of ontology labels

Based on the analysis carried out with the data extracted from the questionnaires, we included more detail in the two first tasks of the methodological guidelines and we added a new task to support the evaluation of the translations of each ontology term. These changes are reflected in the guidelines proposed in this work.

8.3.2 Methodological Guidelines Evaluation Through Use Cases

In this section we include two different examples of how to use the proposed guidelines for the ontology localization activity and the obtained results. These use cases allowed us to assess hypotheses H6 of this thesis.

The first example describes the usability evaluation of the methodological guidelines using a manual translation with independence from the utilized software. The goal is to localize the FAO Pest control ontology, by means of using the guidelines proposed in this thesis. Basically, the ontology localization activity was carried out by FAO Information Management specialist with the contribution of domain and linguistic experts.

The second example refers to the automatic localization of the “EconomyActivity” ontology, an ontology developed within the SEEMP⁶ project, using the LabelTranslator system described in Chapter 6. It is important to mention that the work done within the LabelTranslator system has been one of the inputs to get preliminary guidelines for this activity. Such preliminary guidelines have been extended, improved, and proposed in this thesis.

Manual Localization - Pest control ontology

The objective of this example is to localize the Pest control ontology from English to French and Italian. The input ontology is a module of the

⁶<http://droz.dia.fi.upm.es/hrmontology/>

AGROVOC Concept Server⁷ containing English terms identifying one or more concepts.

Task 1. Select the most appropriate linguistic assets. In general, FAO experts make use of several computer aided translation tools to perform the localization of their ontologies. Basically, FAO experts mainly use FAOTERM, the institutional multilingual terminological system⁸. However it only covers the six official languages of the FAO: English, Spanish, French, Arabic, Chinese, and Russian. In addition to FAOTERM, another important asset used in the localization activity was the Google define functionality⁹. Finally, for this case FAO experts used some cataloguing systems such as AGRIS¹⁰ or FAODOC¹¹.

Task 2. Select ontology label(s) to be localized. From the Pest control ontology, they manually extracted the ontology labels to be localized. For example, they extracted the ontology label “pest control” and their related terms to be localized into French and Italian. The related terms are: “postharvest sparring”, “product protection”, “postharvest control”, and “postharvest treatment”.

Task 3. Obtain ontology label translation(s). For each ontology term, they used a manual process for discovering translation equivalents, discovering the possible senses or definitions of the translations, and to disambiguate the translation senses.

Candidate translations extraction. Translations in French were obtained using FAOTERM, but for Italian they resorted to specialized dictionaries, online or printed. For example for the term “pest control”, FAOTERM returns 11 entries (see Figure 8.5). Most of them are titles of conferences or journals. However, two entries refer to terminology in the area of plant production (“control (of a pest)”) and in the area of pest control (“pest control”).

The multilingual information related to the label “pest control” in the sense of “regulation or management of a species defined as a pest” is shown in Table 8.8. The multilingual information related to the result “control (of a pest)” in the sense of “plant production” is shown in Table 8.9.

As we can see, this technique may be only useful for a limited number of languages, mostly the official FAO languages. Domain experts at this point referred to specialized dictionaries, online or printed to discover translations for Italian. Thus, for example some candidate translations for

⁷<http://naist.cpe.ku.ac.th/agrovoc/>

⁸<http://www.fao.org/faoterm/>

⁹<http://www.google.com/help/features.html>

¹⁰<http://www.fao.org/agris/search/search.do>

¹¹<http://www.fao.org/Documents/>

8.3. METHODOLOGICAL EVALUATION

Found 11 item(s)

- Committee of Experts on **Pest Control** [TITLES]
- **control** (of a **pest**) [PLANT PRODUCTION]
- FAO/UNEP Panel of Experts on Integrated **Pest Control** and Resistance Breeding [TITLES]
- Informal Consultation on Medium and Long-term Policies and Actions for Improved Migratory **Pest Control** in Africa [TITLES]
- Insect and **Pest Control** [TITLES]
- Insect **Pest Control** Section [TITLES]
- Inter-Country Programme for Integrated **Pest Control** in Rice in South and Southeast Asia [TITLES]
- **pest control** [PEST CONTROL]
- Senior Officer (Insect and **Pest Control**) [TITLES]
- Senior Officer (Insects and **Pest Control**) [TITLES]
- Senior Officer (Migratory **Pest Control**) [TITLES]

Figure 8.5: Related items for the term “pest control” extracted from FAOTERM.

the term “pest control” were: “lutte contre les ravageurs”, “lutte phytosanitaire”,... (French), and “nebulizzazione postraccolta”, “difesa dei prodotti immagazzinati”, ... (Italian).

Sense Discovery. In order to discover the definitions, domain experts used the Google define functionality. For example, the search [define:pest control] will show you a list of definitions for “pest control” gathered from various online sources. In Figure 8.6 we show a sample of the definitions obtained for the term “pest control”.

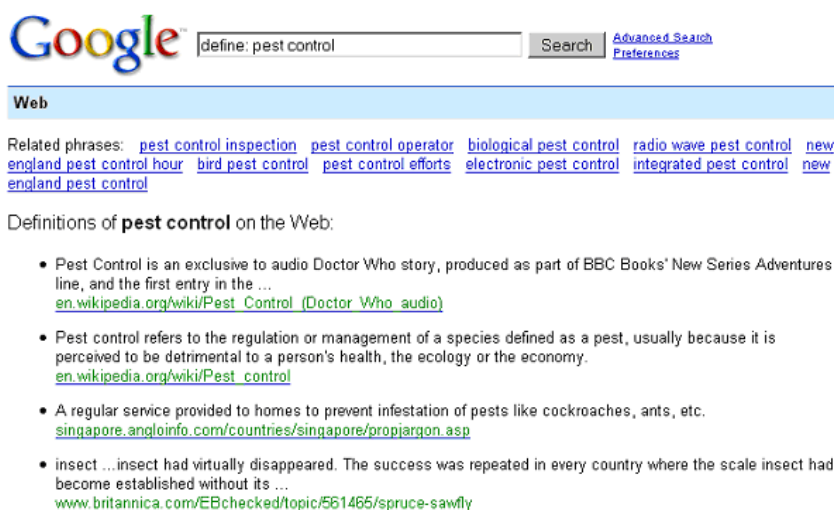


Figure 8.6: Google definitions of the ontology label “pest control”.

Table 8.8: Linguistic information related to the area of “pest control”.

Arabic	
	آفة _ مُضَلِّم ي
	Source: ATG/KCCM, FAO, 2007.
English	
	pest control
Spanish	
	lucha contra las plagas
	Remarks: FAOTERM-Subjects; TRG(A-Sys1)/KCCM, FAO, 2007.
	lucha antiparasitaria
	control de plagas
French	
	lutte contre les ravageurs
	Remarks: FAOTERM-Subjects; TRG(A-Sys1)/KCCM, FAO, 2007.
	lutte phytosanitaire
Russian	
	сельскохозяйственный вредитель
Chinese	
	有害生物防治
	Source: CTG/KCCM, FAO, 2007.
Record info	
Entry Number:75416	
Category	Terminology
Status	Validated
Reliability	2
Source Language	en
Source	FAOTERM-Subjects; TRG(A-Sys1)/KCCM, FAO, 2007.
Subject	PEST CONTROL

Additionally, they checked the use of the term “pest control” and possibly translated documents that make use of its translations in the desired languages using the AGRIS/CARIS resource¹². Figure 8.7 shows a screenshot of the document related to the sample label “pest control”.

Translation Ranking. With the information obtained in the previous steps they used a manual disambiguation process to rank the translations of each ontology term. For example in Table 8.10 we show the ranked translations obtained for the sample term “pest control”.

¹²<http://www.fao.org/agris/search/search.do>

8.3. METHODOLOGICAL EVALUATION

Table 8.9: Linguistic information related to the area of “control (of a pest)”

Arabic	
	مُحكّم
	Definition مُحكّم - امحكّم د ، ءاوتحكّم وأ ابك تح تحريشج
English	
	control (of a pest)
	Definition Suppression, containment or eradication of a pest population.
Spanish	
	control (de una plaga)
	Definition La supresión, contención o erradicación de una población de plagas.
French	
	lutte (contre un organisme nuisible)
	Definition Suppression, enrayement ou éradication de la population d'un organisme nuisible.
Chinese	
	病虫害控制
	病虫害防治
	Source Chinese Academy of Agricultural Sciences, CAAS, Beijing, FAO Language Resources Project 2005. Ref.1: English-Chinese Dictionary of Agriculture, the First Edition, 1998, China Agricultural Press, Beijing. (www.ccap.com.cn).
Record info	
Entry Number:17532	
Category	Terminology
Status	Validated
Reliability	5
Source Language	en
Source	IPPC GLOSSARY
Subject	PLANT PRODUCTION
Glossint	Phytosanitary Terms

Task 4. Evaluate label translation(s). Based on the proposed guidelines they identified the following situation:

- *Semantic fidelity evaluation.* In order to evaluate the semantic fidelity of the translation they used the “Backward Translation” method. In many cases the translation did not exactly match the original meaning, but in a deeper analysis, taking into consideration the context and the topics (agriculture), they identified that the semantic fidelity was covered 100% while the syntactic fidelity was not ensured.
- *Stylistic evaluation.* In this case, they checked elements such as acronyms,

the use of multiple words, capitalizations, etc. For the mentioned case, no particular problems arised but the use of the parenthesis: for example, the English term “Product protection (stored)” appears to be translated in Italian as “Difesa dei prodotti immagazzinati”, and they proposed the label “Difesa dei prodotti (immagazzinati)”. In other cases instead, the proposed translations were consistent.



Figure 8.7: Uses and “possibly” translated documents of the ontology label “pest control”.

Task 5. Ontology update. In this task domain experts stored the translated labels in a external module linked to the AGROVOC Concept Server. This model has been implemented through the AGROVOC Concept Server Workbench tool, which allows users to easily update the ontology. The final ontology will contain at least the terminology shown in Figure 8.8.

Automatic Localization (with LabelTranslator) - EconomyActivity Ontology

The objective of this example is to localize some terms of the EconomyActivity ontology from English to Spanish using the LabelTranslator system. As we introduced in section 6.7, LabelTranslator has been designed with the aim of automating ontology localization, and has been implemented in the ontology editor NeOn Toolkit as a plug-in. In its current version, it can localize ontologies in English, German and Spanish. In its design, the

8.3. METHODOLOGICAL EVALUATION

Table 8.10: Ranked translations of the term “pest control” for French and Italian

English	French	Italian
pest control	lutte contre les ravageurs	Nebulizzazione postraccolta
pest control	lutte phytosanitaire	Difesa dei prodotti immagazzinati
control (of a pest)	lutte (contre un organisme nuisible)	Difesa postraccolta
		Trattamento postraccolta
		Controllo dei parassiti (postraccolta)

methodological guidelines proposed in this thesis have been followed, and some of the techniques described in section 5.2 have been used.

In order to illustrate the results obtained by our system, we will consider the extract of the sample EconomyActivity ontology shown in Figure 8.9. Let us suppose that the user wants to translate the term “Bars” from English into Spanish. According to the domain of the sample ontology, the correct translation of the selected term should refer to a room or establishment where alcoholic drinks are served over a counter, not to a horizontal rod that serves as a support for gymnasts as they perform exercises neither to a rigid piece of metal or wood, etc.

In the following part, we briefly describe how the tasks are performed by our system, and which techniques and tools are used for each task

Task 1. Select the most appropriate linguistic assets. The linguistic assets used by the current version of the LabelTranslator plug-in are multi-lingual linguistic resources, (Wiktionary, or IATE), translation Web services (GoogleTranslate, BabelFish, etc.), semantic Web resources (EuroWordNet and third-party resources retrieved through Watson, a search engine which indexes many ontologies available on the Web), and remote lexical resources. The addition of further domain specific resources is foreseen for domain ontologies

Task 2. Select ontology term(s) to be localized. Once an ontology has been created or imported in the NeOn Toolkit, LabelTranslator allows users and domain experts to manually/automatically sort out the ontology elements that should undergo localization. By right clicking on a frame (concept, attribute, or relation), the Translate action performs the translation of an ontology label (see Figure 8.10).

For each ontology element, LabelTranslator retrieves its local context, its neighboring terms, which are interpreted by the system using a structure-

Italiano	  Nebulizzazione postraccolta   Difesa dei prodotti immagazzinati   Difesa postraccolta (Preferred) published   Trattamento postraccolta   Controllo dei parassiti (postraccolta)
Français	  Lutte après récolte (Preferred)   Pulvérisation après récolte   Lutte antiparasite (après récolte)   Traitement postrécolte   Protection des denrées (stockées)
English	  Postharvest spraying   Pest control (postharvest)   Product protection (stored)   Postharvest treatment   Postharvest control (Preferred)

Figure 8.8: Final Ontology using an external module.

level approach. In our approach, the context of an ontology term is used to disambiguate the lexical meaning of an ontology term. To determine the context of an ontology term, the system retrieves the labels of the set of terms associated with the term under consideration. The list of context labels comprises a set of names which can be direct label names and/or attributes label names, depending on the type of term that is being translated. More details of this process can be consulted in section 6.7.4.

Task 3. Obtain ontology term translation(s). In order to obtain the most appropriate translation for each ontology element in the target language, LabelTranslator uses the following techniques in the indicated order:

8.3. METHODOLOGICAL EVALUATION

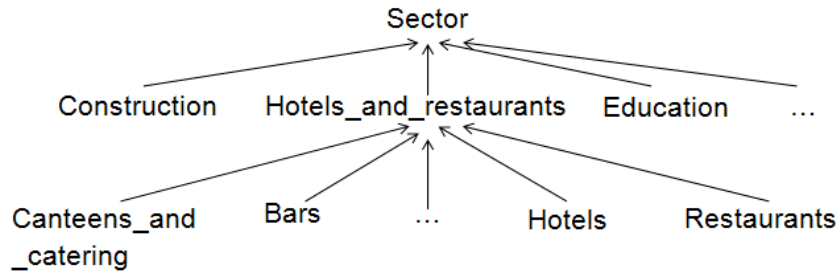


Figure 8.9: Extract of the sample economy activity ontology.

- In step 1 the system obtains equivalent translations for all selected labels by accessing the linguistic assets listed in Task 1.

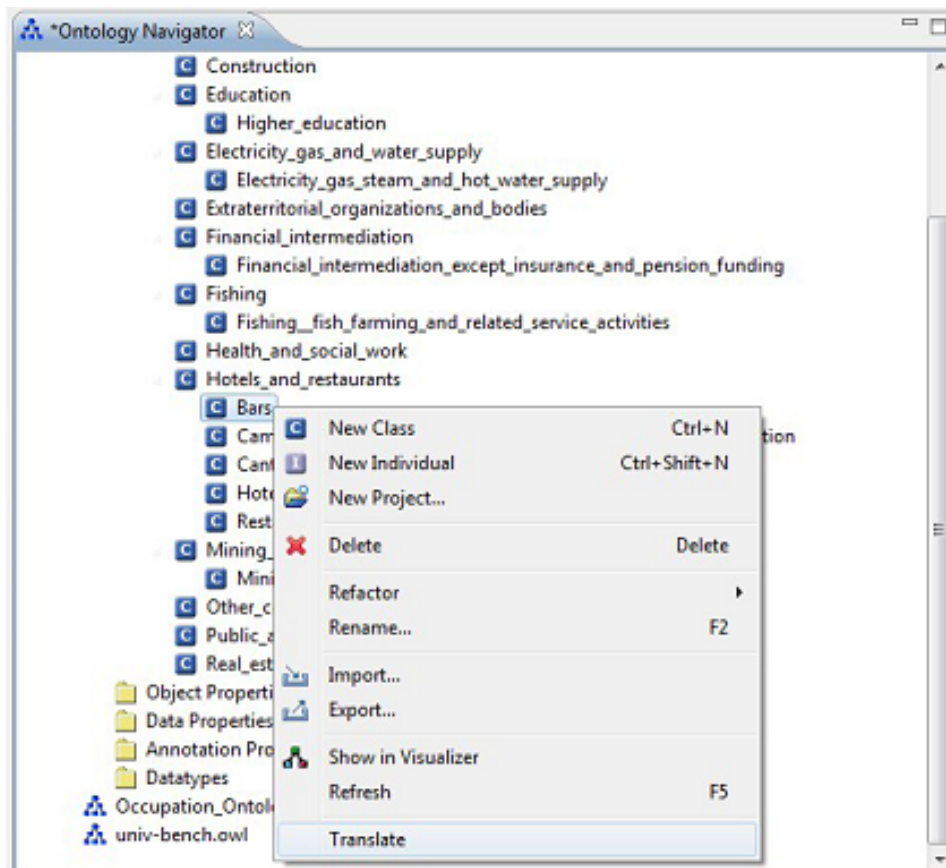


Figure 8.10: Screenshot of the Ontology Navigator view with the Translate action used by the LabelTranslator plug-in.

- In step 2 the system retrieves a list of semantic senses for each translated label, querying Watson and EuroWordnet.

Coming back to our example, in Figure 8.11 we show the translations of the ontology label “Bars” from English into Spanish; our prototype finds eight translations, but we only show three. Notice that t_1 has the desired semantics according to the similarity with the lexical and semantic ontology context (see Figure 8.9).

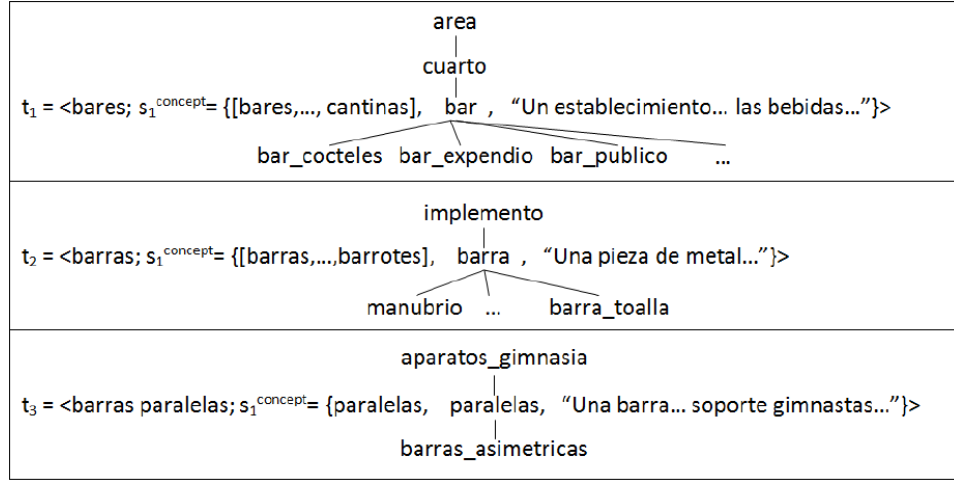


Figure 8.11: Some translations of the Ontology label “Bars” into Spanish.

- In step 3 the system uses a disambiguation method to sort the translations according to their context. LabelTranslator carry out this task in relation to the senses of each translated label and the senses of the context labels. At this stage, domain experts and translators may decide to choose the most appropriate translation among the ranked ones. By default, the system will consider the one in the highest position

In Figure 8.12, we show a sample of the equivalent translations obtained for the term “Bars”. Notice that the obtained translations are ranked according to the ontology context.

Task 4. Evaluate label translation(s). The current version of LabelTranslator does not provide a method for semi-automatically evaluating the translations obtained in the previous step. Therefore, we used a manual evaluation to perform this task. Based on the NeOn methodological guidelines we would identify the following situation

- *Semantic fidelity evaluation* In order to evaluate the semantic fidelity of the translation we would implement the “Backward Translation” criteria. Table 8.11 shows the semantic fidelity evaluation results (only a

8.3. METHODOLOGICAL EVALUATION



Figure 8.12: Equivalent Translations for the Term “Bars”

few cases have been analyzed) for some terms translated into Spanish. The middle column shows the translations obtained by LabelTranslator in Spanish

Table 8.11: Semantic fidelity evaluation results.

<i>Original Term (EN)</i>	<i>Translation (ES)</i>	<i>Backward Translation (EN)</i>
bars	bares	bars
		drinks cabinet
	barras paralelas	parallel bars
	barras	bar
		rod
		stick
		loaf

In many cases the backward translation did not exactly match the original meaning. Thanks to a deeper analysis, which took into consideration the context (hotels and restaurants), we identified that the translation “barras”, for example, did not match the original meaning

- *Stylistic evaluation* The current version of LabelTranslator does not support an automated stylistic evaluation. This task was manually

carried out by an expert in the domain. The translations proposed were consistent in all cases, according to the context of the ontology.

Task 5. Ontology update. The ontology is updated with the resulting linguistic data, which are stored in the LIR model, a separate module adopted by the LabelTranslator NeOn plug-in for organizing and relating linguistic information within the same language and across languages to domain ontologies. Figure 8.13 shows the Linguistic Information page of the sample term “Bars”. The linguistic page uses a model based on a modular approach to store the linguistic information associated with each ontology term. So, one can see that the translation proposed, “bares”, is the full form of a term, is masculine and is considered the main entry in this domain.

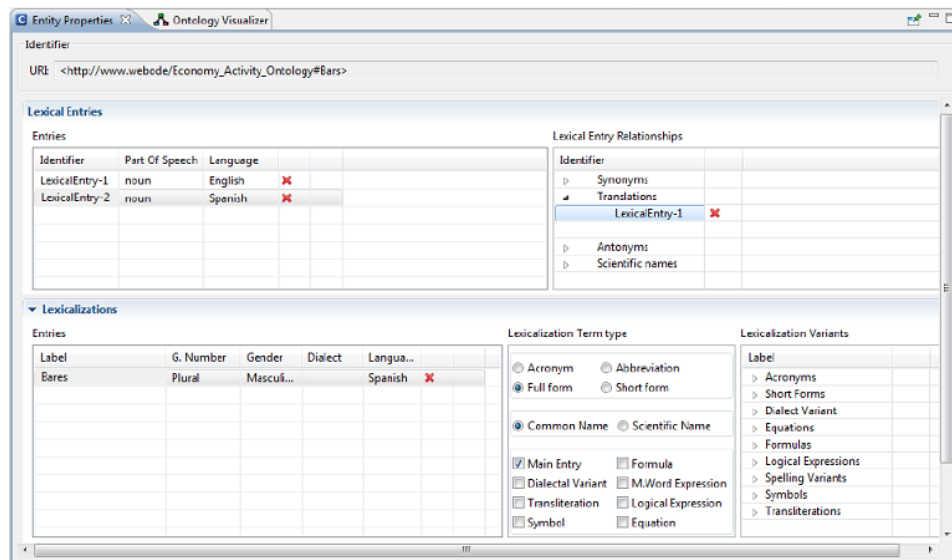


Figure 8.13: Linguistic Information associated to the Ontology Term “Bars”

8.4 Summary of the Chapter

The objective of this chapter was to show how the methods, techniques, and tools proposed in this thesis satisfy the requirements laid down for the ontology localization activity by means of three validation tests.

The first one involved the evaluation of the aspects related to the translation ranking techniques used to select the most appropriate translation for ontology elements. By means of some specific measures, we demonstrate the quality of the output when the algorithm automatically suggests a translation, the quality of all the set of translations, and the quality of the compound labels’ translations.

8.4. SUMMARY OF THE CHAPTER

The second test described the study used to assess the usability of the LabelTranslator system for carrying out the ontology localization activity. We conducted an experiment following the Software Usability Measurement Inventory, which includes around 50 items to measure the affect, efficiency, learnability, helpfulness, and control of our tool.

The last test described two case studies to measure the understanding and usability of the methodological guidelines. These guidelines have not been formally evaluated. Nevertheless, as shown above, we have validated their applicability by using them in two concrete scenarios: using a manual translation with independence of the utilized software and using an automatic localization tool. We believe that the guidelines proposed are effective because following the different tasks and obtaining the expected results in each task they ensure that progress is being achieved and that the goals of the localization activity are being met in the end.

Chapter 9

Conclusions

In this last chapter, we present different conclusions about this work, focusing on the main advances made by the author to support an automatic ontology localization. Basically, in this thesis we have attempted to cover ontology localization in its diversity. In particular, we have identified that the main goal of the ontology localization activity is to enrich an ontology with multilingual information. Also, we have shown that there are different applications that may use the localization output to perform their tasks and that the pressure of applications on ontology localization is tangible. Ontology localization can take advantage of innumerable basic techniques composed and supervised in diverse ways. We analyzed the implications of localizing an ontology, and the different strategies that can be followed to solve translation problems, and to store the multilingual information resulting from the Localization Activity.

We have provided a systematic view of the resources to help users, and developers in performing the localization of an ontology. This has been substantiated by identifying application needs and classifying localization techniques. Finally, we have presented generic methodological guidelines for the localization of ontologies whenever a conceptualization is available. The proposed guidelines have been used in the design of a tool for automatically localizing ontologies.

In the remainder of this chapter we first enumerate our main contributions to the state of the art of the automatic building of multilingual ontologies; second, we present a description of the results achieved; and third, we present some promising research directions which we believe to be worth while and which need further investigations.

9.1 Main Contributions

As described in Chapter 3, the objectives of the work here presented were the following:

- The identification and implementation of the methods, techniques and tools for the management of ontology localization in distributed and collaborative environments.
- The development of an ontology localization methodology for guiding users in the development of multilingual ontologies, based on existing localization methodologies and practices in other areas, as general and open as possible to cover the scenarios to perform the localization activity.

The next sections present the conclusions related to the main contributions made to the state of the art by this thesis.

9.1.1 Identification and Implementation of the Technological Support for Ontology Localization

The future of the Web, the Semantic Web will allow the integration of data-oriented applications as well as document oriented applications. In the process of achieving this goal, ontologies and more precisely multilingual ontologies have become a core technology for representing structured knowledge as well as an instrument to enhance the quality of information retrieval and machine translation.

Before starting this thesis, we have identified two current trends for the building of a multilingual ontology, i.e. the establishment of a new multilingual ontology from scratch and the reconciliation and merging of existing ontologies. However, neither of these methods reduces the cost and effort that means enriching an ontology with multilingual information. In this thesis, we have presented a novel approach which attempts to automatize the enrichment of an ontology with multilingual information, using an ontology localization method. The main contributions of our approach to the state of the art are:

- The definition and explanation of the main processes related to the ontology localization activity such as internationalization and translation.
- The analysis of the different levels of localization used in ontologies, depending on the type of the ontology elements to be localized and the level of adaptation required to make the ontology accessible to speakers of different natural languages.
- The identification and formalization of a generic process for automating the task of the translation of ontology labels.
- The identification and classification of different translation techniques based on the way of modeling the context used for disambiguate the

candidate translations and the type of resources used to support the automated translation of the lexical information associated with the ontology.

- The description of effective translation strategies, which carries many of the advantages of the individual translation models and suffers from few of their disadvantages.
- The description of a comprehensive localization life cycle model, which shows i) what information and knowledge should be specified or defined at different phases, and ii) how the results of ontology element translations provide feedback to other phases of the life cycle.
- The design of a generic architecture for an automated ontology localization in collaborative and distributed environments.

In this thesis, we discuss how an existing ontology whose labels described in a source natural language can be localized into different natural languages. We also identify the translation strategies most suitable for the different ontology elements. Based on this, we use one method for the translation of simple labels and another for the translation of compound labels. In both cases, the methods rely on different linguistic and semantic resources for discovering the more appropriate translations. All translations are ranked based on similarity with their context in the ontology, and the ranked list is used to either present to the user the best candidates, or to use the highest-scoring candidate to automatically translate the label.

We propose an architecture that supports two work scenarios: single and collaborative scenario. In the first scenario, there is only a person impersonating the different roles of the localization activity (e.g., project manager, translator, and reviewer) all at the same time. As such, (s)he will have to perform all organizational and translation tasks. While in the collaborative scenario, there is a team sharing the translation work. This requires a higher volume of organizational work. In this case, one of the participants should be appointed to assume the responsibility of project management.

For the collaborative scenario we also advocate the use of a workflow model that addresses the organizational setting typically followed by organizations in the development and localization of ontologies. The above ideas have been implemented as part of the LabelTranslator system. The system is being used in the NeOn Toolkit, a state-of-the-art, open source multi-platform ontology engineering environment, which provides comprehensive support for the ontology engineering life-cycle.

LabelTranslator System

Concerning to contributions of each module in the architecture of LabelTranslator system, we can emphasize the following issues:

The Ontology Repository

The Ontology Repository is the critical component, which supports the association of the ontological model(s) (source ontologies to be localized) with the linguistic model (multilingual information). The independency of the information stored is one of the most valuable features of the Ontology Repository. The definition, storage and management of multilingual information are completely separated from ontologies. Thus, changes related to improving the linguistic information associated with ontology terms can be performed by different linguistic experts. The following are some other main reasons for its relevance in the system:

- It allows for the inclusion of as much linguistic information as wished, as well as the possibility of establishing links among the linguistic elements within one language or across languages.
- It stores the linguistic information in an ontology format. The benefit of having a linguistic ontology is that it allows for the formal and explicit representation of the linguistic knowledge in a machine-understandable format, which can be easily integrated with other models.

The Localization Manager

This is the module which interacts with the localization stakeholder and guides the process that begins with the selection of the ontology elements to be localized and ends with the updating of the linguistic information into the source ontology. The following are its key features:

- It centralizes the access to information of the localized ontologies. This enables users to view and access localization information from a unique point.
- It tracks and maintains participant information for faster and more efficient selection and turnaround.
- It detects changes in the source ontology, enabling content managers to determine what action needs to be taken without having to manually track content updates.

9.1. MAIN CONTRIBUTIONS

- It enforces and automatically executes critical localization tasks such as ontology submission, change detection, e-mail notification of localization management tasks and events, and real-time tracking and reporting of individual localization tasks.
- It takes into account the different role permissions and the status of the ontology elements to be localized to control and distribute the localization tasks.
- It keeps the conceptual and linguistic information associated to each ontology element updated. It listens the changes in the ontology model and then automatically propagates those changes to the linguistic model using synchronization techniques.

The Ontology Translator

The Ontology Translator itself is the key of most of the main contributions of our system. It allows for the automated translation of the ontology labels using different translation strategies. We enumerate in the following its main features:

- It decreases the localization volume by leveraging previous translations.
- It prepares content for the translation to increase translation quality and decrease translation time.
- It integrates different MT approaches, combining the output by means of different translation combination strategies.
- It invokes different linguistic and semantic resources to discover the translations.
- It correlates the different translations coming from different algorithms and presents a ranked list of translations to the user to review their quality.

9.1.2 Development and Use of the Localization Methodology

At the moment of starting this thesis, the identification of the main problems of the ontology localization activity was unknown. Also, the current software development methodologies are difficult to use in ontology localization, because they are not defined in detail.

The main contribution to solving the first problem is the identification of the main dimensions that must be taken into account in the localization of ontologies. In this thesis we explain these dimensions from three different points of view. First, we describe the translation problems that can be found

at the moment of localizing the ontology elements. Second, we describe the management problems, which make reference to the maintenance and updating of translated ontology labels throughout the ontology life cycle. Finally, we present the multilinguality representation issues, which must be determined by the shareability of the conceptualization and the amount of linguistic information required for the final ontology.

The ontology localization methodology proposed in this thesis uses as a starting point some well-known localization models and development methodologies from other areas and reuse the common tasks of these methodologies. To the best of our knowledge, the study presented here is the first attempt to offer guidelines for the localization of ontologies.

These guidelines have not been formally evaluated. Nevertheless, as we show below, they have been validated through a process of checking that they satisfy the necessary and sufficient conditions of a methodology. Specifically, the ontology localization guidelines are

- Based on existing practices because they have been defined by combining tasks of existing methodological guidelines.
- Collaborative, because they contemplate the participation and consensus of different actors who are distributed geographically.
- Open, because they do not limit the types of ontologies or the specific ontology terms (classes, object or datatype properties) to be considered in localization, nor the resources that should be employed in the actual translation.
- Usable, because they are clearly documented and their use does not involve a great amount of effort.

Furthermore, as presented at the end of this section, the use of the ontology localization guidelines has provided us with ideas on how to improve the guidelines with recommendations for the localization of ontologies of different domains.

The **applicability of the ontology localization methodology** has been proven in both NeOn and Seem Projects where this methodology has been used to localize the FAO Pest control ontology and the Occupation ontology, by means of the guidelines proposed in this thesis. We cannot ensure that the guidelines will be valid in all localization scenarios, but further validation of the guidelines will be possible in future ontology localization projects with different settings. Finally, we have proven that it is feasible to perform a manual localization by using basic guidelines instead of a tool-focused approach.

9.2 Evaluation of Results

The work developed in this thesis has been the goal of several publications in international conferences, workshops and journals. Here we enumerate them in a chronological order.

The method used to *discover the set of candidate meanings for a given word* (or words) from a pool of ontologies available on the Web was presented in the conference paper published in [Espinoza et al., 2006a]. This approach is the core of our proposal to automatically discover the translations of an ontology element. A first approach to extend the discovering of semantic words using ontology matching techniques was briefly presented in the workshop paper and poster published in [Espinoza et al., 2006b, Espinoza et al., 2007]. A joined solution to discover and extract the implicit semantics of a set of words, obtaining their most suitable senses according to their context was the goal of the conference paper [Gracia et al., 2006]. A long article which summarizes our work for discovering the semantic of a set of words from available ontology pools was selected to submit to a special issue of the journal of Universal Computer Science [Trillo et al., 2007]. Although the main goal of the previous works was to discover the different semantic meanings of a word to try to discover the more appropriate translations of an ontology element, we present in the conference paper published in [Espinoza and Mena, 2007] another use of the semantic keywords.

The description of the *main components of our approach to automatically localize an ontology to different natural languages* was first presented in proceedings of 5th Semantic Web Conference [Espinoza et al., 2008a]. The paper also included the experiments performed to evaluate the quality of translations obtained with our approach. The description of the main functionalities of the LabelTranslator system was presented in the demo paper published in [Espinoza et al., 2008b]. An approach to extend the ontology localization proposal to incorporate a modular approach to store the linguistic information associated to ontology terms and to manage the conceptual knowledge and the linguistic knowledge by means of synchronization techniques was presented in the demo paper [Espinoza et al., 2009a].

In the journal article [Cimiano et al., 2010] we aim to clarify the notion of ontology localization as well as the different layers of an ontology that are affected in this process. To use statistical machine translation (SMT) techniques for obtaining the most appropriate translations of ontology elements was the goal of the workshop paper published in [McCrae et al., 2011a]. And *the description of a generic Ontology Localization Activity and a methodology for guiding in the localization of ontologies* was presented in the KCAP-09 conference [Espinoza et al., 2009b]. This paper describes a set of experiments used to evaluate the methodological and technological aspects of the Ontology Localization Activity. The previous work is part of the extended version [Espinoza et al., 2012] of a chapter of the book “On-

tology Engineering in a Networked World”, which provides the necessary methodological and technological support for the development and use of ontology networks.

9.3 Future Challenges

Here we present some directions in which, in our opinion, research on ontology localization should or is likely to evolve. In particular, in this section, we point out current needs that are not addressed in this thesis and that will have to be addressed for the field to be considered mature. We detail these improvements in the following:

- *Applications.* In this work we have conceived the use of ontology localization in different applications; however at the moment of writing this thesis neither of these applications has incorporated the output of the ontology localization activity as part of their process. Thus, we can expect that there definitely will be applications which will use ontology localization. They will start in niche places with a specific setting rather than presenting a general solution to a global problem. Then, gradually, the proven solutions will start spreading to other applications.
- *Foundations.* Foundations of ontology localization, and particularly the identification of the necessary methods to localize ontologies with culturally-dependant domains (e.g., the judicature) in which categorizations tend to reflect the particularities of a certain culture, deserve additional investigations.
- *Translation techniques.* As shown in Chapter 5, there is a wealth of basic techniques that can be used to find the translations of the ontology elements. In this thesis we have used only some of these techniques. However, further investigation is necessary in order to incorporate both corpus-based and Web-based techniques in the localization activity.
- *Translation strategies.* In this work we have incorporated two basic translation strategies for discovering the translations of simple and compound labels. One of the most important issues to deal with is the proper combination and integration of various categories of translators. In particular, the integration of corpus-based (statistical) and ontology-based (semantic) techniques is of high interest.
- *Evaluation of localization systems.* It is necessary to design extensive evaluation mechanism of ontology localization systems. Besides evaluating systems, it is necessary to be able to help users in choosing the

appropriate translation technique or to combine the most appropriate techniques for their tasks. We have tried in this thesis to identify some localization strategies, but a lot remains to be investigated, for example the localization of instances and descriptions of ontological terms.

- *Ontology Localization Systems.* Developments in Natural Language Processing technologies promise a variety of benefits to the ontology localization activity, both in its current form in performing bulk community-based localization and in the future in supporting personalized Web-based localization on increasingly user-generated content.

As an increasing variety of natural language processing services becomes available, it is vital that the ontology localization activity employs the flexible software integration techniques that will enable it to make the best use of these technologies. We expect that the ontology localization systems make good use of the benefits of modern integration technologies such as Web service integration and orchestration.

- *Processing translations.* Processing translations according to application needs is the ultimate goal of localization. In this thesis we do not have considered a possible application of the translations obtained for each ontology element. However we believe that the processing of the obtained translations may vary depending on the final application of the multilingual ontology. Therefore, the storing of useful translations in an independent format such as those presented in section 7.3.2 is very important. It would allow for the sharing and processing of these translations in different ways and independently of the applications.

Relevant Publications Related to the Thesis

(in reverse chronological order)

- [Espinoza,12] Mauricio Espinoza, Elena Montiel-Ponsoda, Guadalupe Aguado de Cea, Asunción Gómez-Pérez, chapter “Ontology Localization” of the book “Ontology Engineering in a Networked World, ISBN 978-3-642-24793-4”, Suárez-Figueroa, M.C., Gómez-Pérez, A., Motta, E., Gangemi, A. (ed.), Springer, pp. 490, 2012. February 2012.
- [McCrae,11] John McCrae, Mauricio Espinoza, Elena Montiel-Ponsoda, Guadalupe Aguado-de-Cea and Philipp Cimiano, “Combining statistical and semantic approaches to the translation of ontologies and taxonomies”, Proc. of Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-5), Portland, Oregon, USA, ISBN 978-1-932432-99-2, Association for Computational Linguistics (ACL), June 2011.
- [Cimiano,10] Philipp Cimiano, Elena Montiel-Ponsoda, Paul Buitelaar, Mauricio Espinoza and Asunción Gómez-Pérez, “A Note on Ontology Localization”, Applied Ontology, ISSN 1570-5838, volume 5, number 2, pp. 127-137, 2010.
- [Espinoza,09b] Mauricio Espinoza, Elena Montiel-Ponsoda and Asunción Gómez-Pérez, “Ontology Localization”, Proc. of Fifth International Conference on Knowledge Capture (K-CAP’09), Redondo Beach, California, USA, ACM, ISBN:978-1-60558-658-8, pp. 33-40, September 2009.
- [Espinoza,09a] Mauricio Espinoza, Asunción Gómez-Pérez, and Elena Montiel-Ponsoda, “Multilingual and Localization Support for Ontologies”, Proc. of 5th European Semantic Web Conference (ESWC’09), Heraklion, Crete, Greece, Springer Verlag LNCS, ISBN 978-3-642-02120-6, pp. 821-825, May 2009. Demo paper.
- [Espinoza,08b] Mauricio Espinoza, Asunción Gómez-Pérez and Eduardo Mena, “Enriching an Ontology with Multilingual Information”, Proc. of 5th

- European Semantic Web Conference (ESWC'08), Tenerife (Spain), Springer Verlag LNCS, ISBN 978-3-540-68233-2, ISSN-0302-9743, pp. 333-347, June 2008.
- [Espinoza,08a] Mauricio Espinoza, Asunción Gómez-Pérez and Eduardo Mena, "LabelTranslator - A Tool to Automatically Localize an Ontology", Proc. of 5th European Semantic Web Conference (ESWC'08), Tenerife (Spain), Springer Verlag LNCS, ISBN 978-3-540-68233-2, ISSN-0302-9743, pp. 792-796, June 2008. Demo paper.
- [Trillo *et al.*,07] Raquel Trillo, Jorge Gracia, Mauricio Espinoza and Eduardo Mena, "Discovering the Semantics of User Keywords", Journal on Universal Computer Science (JUCS). Special Issue: Ontologies and their Applications, ISSN 0948-695X, 13(12):1908-1935, Springer Verlag, December 2007.
- [Espinoza,07b] M. Espinoza and E. Mena, "Discovering Web Services Using Semantic Keywords", 5th IEEE International Conference on Industrial Informatics (INDIN-2007), Vienna (Austria), IEEE, ISBN 1-4244-0864-4, pp. 725-730, July 2007.
- [Espinoza,07a] M. Espinoza, J. Gracia, R. Trillo and E. Mena, "Discovering the Semantics of User Keywords", 4th European Semantic Web Conference (ESWC-2007), Innsbruck, Austria, June 2007. Poster.
- [Espinoza,06b] M. Espinoza, R. Trillo, J. Gracia and E. Mena, "Discovering and Merging Keyword Senses using Ontology Matching", 1st International Workshop on Ontology Matching (OM-2006) @ 5th International Semantic Web Conference ISWC-2006, Athens, Georgia (USA), CEUR-WS, ISSN 1613-0073, volume 225, pp. 1-5, November 2006.
- [Gracia *et al.*,06] J. Gracia, R. Trillo, M. Espinoza and E. Mena, "Querying the Web: A Multiontology Disambiguation Method", Sixth International Conference on Web Engineering (ICWE'06), Palo Alto, California (USA), ACM, ISBN 1-59593-352-2, pp. 241-248, July 2006.
- [Espinoza,06a] M. Espinoza, J. Gracia, R. Trillo and E. Mena, "Discovering the Semantics of Keywords: An Ontology-based Approach", The 2006 International Conference on Semantic Web and Web Services (SWWS'06), Las Vegas, Nevada (USA), CSREA Press, ISBN 1-60132-016-7, June 2006.

Bibliography

- [Agirre and Stevenson, 2006] Agirre, E. and Stevenson, M. (2006). Knowledge sources for wsd. In *Word Sense Disambiguation: Algorithms and Applications*, volume 33 of *Text, Speech and Language Technology*, pages 217–252. Springer, Dordrecht, The Netherlands.
- [AGROVOC, 2005] AGROVOC, M. A. T. (2005). Food and Agricultural Organization the United Nations, <http://www.fao.org/agrovoc/>.
- [Akiba et al., 2004] Akiba, Y., Sumita, E., Nakaiwa, H., Yamamoto, S., and Okuno, H. G. (2004). Using a mixture of n-best lists from multiple mt systems in rank-sum-based confidence measure for mt outputs. In *Proceedings of the 20th international conference on Computational Linguistics, COLING '04*, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Allemang and Polikoff, 2004] Allemang, D. and Polikoff, I. (2004). Topbraid, a multi-user environment for distributed authoring of ontologies. In *Proceedings of the Third International Semantic Web Conference (ISWC 2004)*. Springer Verlag.
- [Alonso et al., 2005] Alonso, L. S., Bas, L. J., Bellido, S., Contreras, J., Benjamins, R., and Gómez, J. (2005). D10.7 financial ontology, data, information and process integration with semantic web services, fp6–507483. In *WP10: Case Study eBanking*.
- [Apidianaki, 2009] Apidianaki, M. (2009). Data-driven semantic analysis for multilingual wsd and lexical selection in translation. In *EACL '09: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 77–85, Morristown, NJ, USA. Association for Computational Linguistics.
- [Asanoma, 2001] Asanoma, N. (2001). Alignment of ontologies: Wordnet and goi-taikei. In *WordNet and Other Lexical Resources Workshop Program, NAACL2001*, pages 89–94.
- [Baker, 1995] Baker, M. (1995). Corpora in translation studies: an overview and some suggestions for future research. In *Target 7(2):223–243*.

- [Balkan et al., 2002] Balkan, L., Ken, M., Birgit, A., Anne, E., Myriam, G. B., and Pam, M. (2002). Elsst: a broad-based multilingual thesaurus for the social sciences. In *Third International Conference on language Resources and Evaluation LREC'02*.
- [Ballesteros and Croft, 1998] Ballesteros, L. and Croft, B. W. (1998). Resolving ambiguity for cross-language retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 64–71.
- [Barrasa, 2007] Barrasa, J. (2007). *Modelo para la definición automática de correspondencias semánticas entre ontologías y modelos relacionales*. PhD thesis, UPM, Madrid, Spain.
- [Bechhofer et al., 2001] Bechhofer, S., Horrocks, I., Goble, C., and Stevens, R. (2001). Oiled: a reason-able ontology editor for the semantic web. In *Proceedings of KI2001, Joint German/Austrian conference on Artificial Intelligence, September, Vienna. Springer-Verlag LNAI Vol. 2174, pp. 396–408*.
- [Bechhofer et al., 2004] Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D., Patel-Schneider, P., and Stein, L. A. (2004). OWL Web Ontology Language Reference. Technical report, W3C.
- [Benjamins et al., 2002] Benjamins, R. V., Contreras, J., Corcho, O., and Gómez-Pérez, A. (2002). Six challenges for the semantic web. In *Proceedings of the Semantic Web workshop held at KR-2002*.
- [Bentivogli et al., 2000] Bentivogli, L., Pianta, E., and Pianesi, F. (2000). Coping with lexical gaps when building aligned multilingual wordnets. In *Second International Conference on Language Resources and Evaluation (LREC 2000)*, pages 993–997.
- [Berners-Lee et al., 1992] Berners-Lee, T., Cailliau, R., Groff, J.-F., and Pollermann, B. (1992). World-wide web: The information universe. *Electronic Networking: Research, Applications and Policy*, 1(2):74–82.
- [Berners-Lee et al., 2001] Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The semantic web. *Scientific American*.
- [Bizer et al., 2009] Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3):1–22.
- [Blatz et al., 2004] Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., Sanchis, A., and Ueffing, N. (2004). Confidence estimation for machine translation. In *Proceedings of COLING 04*:

BIBLIOGRAPHY

- The 20th International Conference on Computational Linguistics, pages 315-321, Geneva, Switzerland.*
- [Boiko, 2005] Boiko, B. (2005). *Content Management Bible, 2nd Edition*. Wiley Publishing Inc., Indianapolis.
- [Bonino et al., 2004] Bonino, D., Corno, F., Farinetti, L., and Ferrato, A. (2004). Multilingual semantic elaboration in the dose platform. In *SAC 2004, ACM Symposium on Applied Computing*.
- [Borst, 1997] Borst, W. N. (1997). *Construction of Engineering Ontologies for Knowledge Sharing and Reuse*. Enschede, The Netherlands, Doctoral Thesis of the University of Twente.
- [Brickley and Guha, 2000] Brickley, D. and Guha, R. V. (2000). Resource Description Framework (RDF) Schema Specification. W3C Recommendation, World Wide Web Consortium.
- [Brill, 1995] Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, 21:543–565.
- [Briscoe, 1991] Briscoe, T. (1991). Lexical issues in natural language processing. In *Natural Language and Speech*, pages 39–68. Springer-Verlag.
- [Brown et al., 1990] Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S. (1990). A statistical approach to machine translation. *Comput. Linguist.*, 16(2):79–85.
- [Brown et al., 1993] Brown, P. F., Pietra, V. J., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19:263–311.
- [Budanitsky, 2001] Budanitsky, A. (2001). Semantic Distance in WordNet: An Experimental, Application-oriented Evaluation of Five Measures.
- [Buitelaar et al., 2009] Buitelaar, P., Cimiano, P., Haase, P., and Sintek, M. (2009). Towards linguistically grounded ontologies. In *Proceedings of the 6th European Semantic Web Conference on The Semantic Web: Research and Applications*, ESWC 2009 Heraklion, pages 111–125, Berlin, Heidelberg. Springer-Verlag.
- [Buitelaar et al., 2006] Buitelaar, P., Sintek, M., and Kiesel, M. (2006). A multi-lingual/multimedia lexicon model for ontologies. In *ESWC'06, Budva, Montenegro*.

- [Cadieux and Esselink, 2004] Cadieux, P. and Esselink, B. (2004). Gilt: Globalization, internationalization, localization, translation. *Globalization Insider XI (1.5)*.
- [Callan et al., 2003] Callan, J., Crestani, F., Nottelmann, H., Pala, P., and Shou, X. M. (2003). Resource selection and data fusion in multimedia distributed digital libraries. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 363–364, New York, NY, USA. ACM.
- [Callison-Burch and Flounoy, 2001] Callison-Burch, C. and Flounoy, R. S. (2001). A program for automatically selecting the best output from multiple machine translation engines. In *Proceedings of the Machine Translation Summit VIII*, pages 63–66.
- [Cancedda and Yamada, 2005] Cancedda, N. and Yamada, K. (2005). Method and apparatus for evaluating machine translation quality.
- [Caracciolo et al., 2007] Caracciolo, C., Sini, M., and Keizer, J. (2007). Requirements for the treatment of multilinguality in ontologies within fao. In *OWLED*.
- [Carmel and Agarwal, 2001] Carmel, E. and Agarwal, R. (2001). Tactical approaches for alleviating distance in global software development. *IEEE Software*, 18(2).
- [Carpuat et al., 2002] Carpuat, M., Ngai, G., Fung, P., and Church, K. (2002). Creating a bilingual ontology: A corpus-based approach for aligning wordnet and hownet. In *Proceedings of the 1st Global WordNet Conference, Mysore*, pages 284–292.
- [Carpuat and Wu, 2007] Carpuat, M. and Wu, D. (2007). Context-dependent phrasal translation lexicons for statistical machine translation. In *Proceedings of MT Summit XI, Copenhagen, Denmark*.
- [Cer et al., 2010] Cer, D., Galley, M., Jurafsky, D., and Manning, C. D. (2010). Phrasal: A statistical machine translation toolkit for exploring new model features. In *Proceedings of the NAACL HLT 2010 Demonstration Session*, pages 9–12, Los Angeles, California. Association for Computational Linguistics.
- [Chalupsky, 2000] Chalupsky, H. (2000). Ontomorph: A translation system for symbolic knowledge. In *Principles of Knowledge Representation and Reasoning*, pages 471–482. Morgan Kaufmann.
- [Chang et al., 2008] Chang, P.-C., Galley, M., and Manning, C. D. (2008). Optimizing chinese word segmentation for machine translation performance. In *StatMT '08: Proceedings of the Third Workshop on Statistical*

BIBLIOGRAPHY

- Machine Translation*, pages 224–232, Morristown, NJ, USA. Association for Computational Linguistics.
- [Chen and Fung, 2004] Chen, B. and Fung, P. (2004). Automatic construction of an english-chinese bilingual framenet. In *HLT-NAACL '04: Proceedings of HLT-NAACL 2004: Short Papers on XX*, pages 29–32, Morristown, NJ, USA. Association for Computational Linguistics.
- [Chen et al., 2012] Chen, J., Ding, R., Jiang, S., and Knudson, R. (2012). A preliminary evaluation of metadata records machine translation. *The Electronic Library*.
- [Cheng et al., 2004] Cheng, P.-J., Teng, J.-W., Chen, R.-C., Wang, J.-H., Lu, W.-H., and Chien, L.-F. (2004). Translating unknown queries with web corpora for cross-language information retrieval. In *Proc. of the 27th annual international conference on Research and development in information retrieval*.
- [Chien, 1997] Chien, L. F. (1997). Pat-tree-based keyword extraction for chinese information retrieval. In *Proceedings of ACM-SIGIR97*, 50–59.
- [Choueka, 1988] Choueka, Y. (1988). Looking for needles in a haystack, or locating interesting collocational expressions in large textual databases. In *International Conference on User-Oriented Content-Based Text and Image Handling*, pages 609–623, Cambridge, MA.
- [Chun and Wenlin, 2002] Chun, C. and Wenlin, L. (2002). The translation of agricultural multilingual thesaurus. In *AFITA2002, Asian Agricultural Information Technology & Management. Proceedings of the Third Asian Conference for Information Technology in Agriculture*.
- [Church, 1988] Church, K. W. (1988). A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the second conference on Applied natural language processing*, ANLC '88, pages 136–143, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Cimiano et al., 2007] Cimiano, P., Haase, P., Herold, M., Mantel, M., and Buitelaar, P. (2007). Lexonto: A model for ontology lexicons for ontology-based nlp. In *OntoLex'07, Busan, South Korea*.
- [Cilibrasi and Vitányi, 2007] Cilibrasi, R. L. and Vitányi, P. M. (2007). The Google Similarity Distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370–383.
- [Cimiano et al., 2010] Cimiano, P., Montiel-Ponsoda, E., Buitelaar, P., Espinoza, M., and Gómez-Pérez, A. (2010). A note on ontology localization. *Appl. Ontol.*, 5(2):127–137.

- [Collins, 2001] Collins, R. W. (2001). Software localization: Issues and methods. In Smithson, S., Gricar, J., Podlogar, M., and Avgerinou, S., editors, *ECIS*, pages 36–44.
- [Corcho, 2011] Corcho, O. (2011). *A layered approach to ontology translation with knowledge representation*. PhD thesis, Universidad Politécnica de Madrid, Madrid, España.
- [Corcho et al., 2006] Corcho, O., López-Cima, A., and Gómez-Pérez, A. (2006). The odesew 2.0 semantic web application framework. In *Proceedings of the 15th International Conference on World Wide Web (Edinburgh, Scotland, May 23 - 26, 2006)*. WWW '06, pages 1049–1050. ACM Press, New York, NY.
- [Costa-Juss, 2008] Costa-Juss, M. R. (2008). *New reordering and modeling approaches for statistical machine translation*. PhD thesis, UPC, Barcelona, Spain.
- [Cui et al., 2004] Cui, G., Chen, F., Chen, H., and Li, S. (2004). Ontoedu a case study of ontology-based education grid system for e-learning. In *The Global Chinese Conference on Computers in Education conference*.
- [Decadt et al., 2004] Decadt, B., Hoste, V., Daelemans, W., and van den Bosch, A. (2004). Gambl, genetic algorithm optimization of memory-based wsd. In *Proceedings of the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3)*, pages 108–112, Barcelona, Spain. ACL.
- [Decker et al., 1999] Decker, S., Erdmann, M., Fensel, D., and Studer, R. (1999). Ontobroker: Ontology Based Access to Distributed and Semi-Structured Information. In Meersman, R., editor, *Database Semantics, Semantic Issues in Multimedia Systems*, pages 351–369. Kluwer Academic Publisher, Boston.
- [Declerck et al., 2006] Declerck, T., Gómez-Pérez, A., Vela, O., Gantner, Z., and Manzano-Macho, D. (2006). Multilingual lexical semantic resources for ontology translation. In *Proceedings of LREC 2006*.
- [DeRose, 1988] DeRose, S. J. (1988). Grammatical category disambiguation by statistical optimization. *Comput. Linguist.*, 14(1):31–39.
- [Doddington, 2002] Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

BIBLIOGRAPHY

- [Dorr et al., 2002] Dorr, B., , Dorr, B., and Habash, N. (2002). Interlingua approximation: A generation-heavy approach. In *Proceedings of AMTA-2002*. University of Chicago Press.
- [Dorr et al., 2000] Dorr, B., Levow, G., Lin, D., and Thomas, S. (2000). Large scale construction of chinese-english semantic hierarchy. Technical report, University of Maryland, College Park, MD.
- [Duhl, 2008] Duhl, J. (2008). How to Overcome 3 Common Localization Challenges.
- [Dumas and Redish, 1993] Dumas, J. and Redish, J. (1993). A Practical Guide to Usability Testing. Exeter, UK: Intellect.
- [Dunne, 2006] Dunne, K. (2006). Perspectives on localization. *A Copernican revolution In Dunne, K. J. (ed.)*, pages 1–12.
- [Dzbor et al., 2009] Dzbor, M., Suárez-Figueroa, M. C., Blomqvist, E., Lewen, H., Espinoza, M., Gómez-Pérez, A., and Palma, R. (2009). Multilingual Ontologies for Networked Knowledge. Monnet Project Deliverable 5.6.2.
- [Ehrig, 2007] Ehrig, M. (2007). *Ontology Alignment. Bridging the Semantic Gap*. Springer Science+Business Media, LLC.
- [Englund, 2005] Englund, B. (2005). *Expertise and Explication in the Translation Process*. Benjamins Translation Library, Amsterdam.
- [Espinoza et al., 2008a] Espinoza, M., Gómez-Pérez, A., and Mena, E. (2008a). Enriching an ontology with multilingual information. In *Proc. of 5th European Semantic Web Conference (ESWC’08), Tenerife, (Spain)*.
- [Espinoza et al., 2008b] Espinoza, M., Gómez-Pérez, A., and Mena, E. (2008b). Labeltranslator - automatically localizing an ontology. In *Proc. of 5th European Semantic Web Conference (ESWC’08), Tenerife, (Spain)*.
- [Espinoza et al., 2009a] Espinoza, M., Gómez-Pérez, A., and Montiel-Ponsoda, E. (2009a). Multilingual and localization support for ontologies. In *Proc. of 6th European Semantic Web Conference (ESWC’09), Heraklion, (Greece)*.
- [Espinoza et al., 2006a] Espinoza, M., Gracia, J., Trillo, R., and Mena, E. (2006a). Discovering the semantics of keywords: An ontology-based approach. In *The 2006 International Conference on Semantic Web and Web Services (SWWS’06), Las Vegas, Nevada (USA)*. CSREA Press.
- [Espinoza et al., 2007] Espinoza, M., Gracia, J., Trillo, R., and Mena, E. (2007). Discovering the semantics of user keywords. In *4th European Semantic Web Conference (ESWC-2007), Innsbruck, Austria, June 2007*.

- [Espinoza and Mena, 2007] Espinoza, M. and Mena, E. (2007). Discovering web services using semantic keywords. In *5th IEEE International Conference on Industrial Informatics (INDIN-2007), Vienna (Austria), IEEE, ISBN 1-4244-0864-4, pp. 725-730.*
- [Espinoza et al., 2012] Espinoza, M., Montiel-Ponsoda, E., Aguado de Cea, G., and Gómez-Pérez, A. (2012). *Ontology Engineering in a Networked World*, chapter 8. Springer.
- [Espinoza et al., 2009b] Espinoza, M., Montiel-Ponsoda, E., and Gómez-Pérez, A. (2009b). Ontology localization. In *Proc. of 5th international conference on Knowledge capture (KCAP-09), Redondo Beach, California (USA), ACM, ISBN 978-1-60558-658-8, pp. 33-40.*
- [Espinoza et al., 2010] Espinoza, M., Montiel-Ponsoda, E., Gracia, J., and Aguado de Cea, G. (2010). Multilingual Ontologies for Networked Knowledge. Monnet Project Deliverable 2.2.1.
- [Espinoza et al., 2006b] Espinoza, M., Trillo, R., Gracia, J., and Mena, E. (2006b). Discovering and merging keyword senses using ontology matching. In *1st International Workshop on Ontology Matching (OM-2006) @ 5th International Semantic Web Conference ISWC-2006, Athens, Georgia (USA), CEUR-WS, ISSN 1613-0073, volume 225, pp. 1-5, November 2006.*
- [Esselink, 1998] Esselink, B. (1998). *A Practical Guide to Software Localization: For Translators, Engineers and Project Managers*. John Benjamins.
- [Esselink, 2000] Esselink, B. (2000). *A Practical Guide to Localization*. John-Benjamins.
- [Euzenat, 2001] Euzenat, J. (2001). Towards a principled approach to semantic interoperability. In *Workshop on Ontologies and Information Sharing, IJCAI01.*
- [Euzenat and Shvaiko, 2007] Euzenat, J. and Shvaiko, P. (2007). *Ontology matching*. Springer-Verlag, Heidelberg (DE).
- [Farquhar et al., 1996] Farquhar, A., Fikes, R., and Rice, J. (1996). The ontolingua server: a tool for collaborative ontology construction. In *International Journal of Human-Computer Studies*.
- [Fernández-López et al., 1999] Fernández-López, M., Gómez-Pérez, A., Pazos-Sierra, J., and Pazos-Sierra, A. (1999). Building a chemical ontology using methontology and the ontology design environment. In *IEEE Intelligent Systems, 14(1):3746.*

BIBLIOGRAPHY

- [Fiscus, 1997] Fiscus, J. (1997). A post-processing system to yield reduced word error rates: Recogniser output voting error reduction (rover). In *Proceedings 1997 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 347–352, Santa Barbara, CA.
- [Flanagan, 1994] Flanagan, M. (1994). Error classification for mt evaluation. *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, pages 65–72.
- [Flied et al., 2007] Flied, G., Kop, C., and Vhringer, J. (2007). From owl class and property labels to human understandable natural language. In *Proceeding of 12th International Conference on Applications of Natural Language to Information Systems, NLDB 2007*.
- [Fordyce, 2007] Fordyce, C. S. (2007). EuroMatrix, Statistical and Hybrid Machine Translation Between All European Languages. Deliverable 1.3 Survey of Machine Translation Evaluation.
- [Fox, 1992] Fox, M. S. (1992). The tove project towards a common-sense model of the enterprise. In *IEA/AIE '92: Proceedings of the 5th international conference on Industrial and engineering applications of artificial intelligence and expert systems*, pages 25–34, London, UK. Springer-Verlag.
- [Fu et al., 2009a] Fu, B., Brennan, R., and O’Sullivan, D. (2009a). Cross-lingual ontology mapping - an investigation of the impact of machine translation. In *Proceedings of the 4th Annual Asian Semantic Web Conference, LNCS 5926, pp. 1-15, Shanghai, China, December*.
- [Fu et al., 2009b] Fu, B., Brennan, R., and O’Sullivan, D. (2009b). Multilingual ontology mapping: Challenges and a proposed framework. In *Workshop on Matching and Meaning - Automated Development, Evolution and Interpretation of Ontologies: 32-35, Edinburgh, UK, April*.
- [Gamallo, 2007] Gamallo, P. (2007). Learning bilingual lexicons from comparable english and spanish corpora. In *Proceedings of Machine Translation Summit XI, Copenhagen, Denmark, pp. 191–198*.
- [Gandrabur and Foster, 2003] Gandrabur, S. and Foster, G. (2003). Confidence estimation for text prediction. In *Proceedings of the Conference on Natural Language Learning (CoNLL), pages 95102, Edmonton, Canada*.
- [Gao et al., 2002] Gao, J., Zhou, M., Nie, J.-Y., He, H., and Chen, W. (2002). Resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependence relations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '02*, pages 183–190, New York, NY, USA. ACM.

- [García-Castro, 2008] García-Castro, R. (2008). *Benchmarking Semantic Web Technology*. PhD thesis, Facultad de Informática, Universidad Politécnica de Madrid, Madrid, Spain,.
- [Garside, 1987] Garside, R. (1987). The CLAWS word-tagging system. In Garside, R., Leech, G., and Sampson, G., editors, *The Computational Analysis of English: a corpus-based approach*, pages 30–41. Longman.
- [Gimpel and Smith, 2008] Gimpel, K. and Smith, N. A. (2008). Rich source-side context for statistical machine translation. In *StatMT '08: Proceedings of the Third Workshop on Statistical Machine Translation*, pages 9–17, Morristown, NJ, USA. Association for Computational Linguistics.
- [Global Vision, 2007] Global Vision (2007). Much talk about Machine Translation, <http://www.globalvis.com/>.
- [Gómez-Pérez et al., 2004] Gómez-Pérez, A., Fernández-López, M., and et al. (2004). *Ontological Engineering: with examples from the areas of Knowledge Management, e-commerce and the Semantic Web*. Springer-Verlag.
- [Gotti et al., 2005] Gotti, F., Langlais, P., Macklovitch, E., Bourigault, D., Robichaud, B., and Coulombe, C. (2005). 3gtm: A third-generation translation memory. In *3rd Computational Linguistics in the North-East (CLiNE) Workshop*, Gatineau, Québec.
- [Goutte, 2006] Goutte, C. (2006). Automatic evaluation of machine translation quality.
- [Gracia and Mena, 2009] Gracia, J. and Mena, E. (2009). Multiontology semantic disambiguation in unstructured web contexts. In *Proc. of Workshop on Collective Knowledge Capturing and Representation (CKCaR'09) at K-CAP'09, Redondo Beach, California (USA)*. CEUR-WS, ISSN 1613-0073.
- [Gracia et al., 2006] Gracia, J., Trillo, R., Espinoza, M., and Mena, E. (2006). Querying the web: A multiontology disambiguation method. In *Sixth International Conference on Web Engineering (ICWE'06), Palo Alto, California (USA), ISBN 1-59593-352-2, pp. 241–248*. ACM.
- [Grefenstette, 1998] Grefenstette, G. (1998). *Cross-Language Information Retrieval*. Springer US.
- [Grefenstette, 1999] Grefenstette, G. (1999). The www as a resource for example-based mt tasks. In *ASLIB99 Translating and the Computer 21*.
- [Gruber, 1995] Gruber, T. (1995). Towards principles for the design of ontologies used for knowledge sharing. In *The International Journal of Human-Computer studies*, 43(5/6) : pp. 907-928.

BIBLIOGRAPHY

- [Guarino, 1998] Guarino, N. (1998). *Formal Ontology in Information Systems: Proceedings of the 1st International Conference June 6-8, 1998, Trento, Italy*. IOS Press, Amsterdam, The Netherlands, The Netherlands.
- [Guyot et al., 2005] Guyot, J., Radhouani, S., and Falquet, G. (2005). Ontology-based multilingual information retrieval. In *CLEF*.
- [Habash and Sadat, 2006] Habash, N. and Sadat, F. (2006). Arabic preprocessing schemes for statistical machine translation. In *NAACL '06: Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers on XX*, pages 49–52, Morristown, NJ, USA. Association for Computational Linguistics.
- [Hasse et al., 2008] Hasse, P., Lewen, H., Studer, R., and Erdmann, M. (2008). The neon ontology engineering toolkit. In *WWW 2008 Developers Track*.
- [Hearne and Way, 2011] Hearne, M. and Way, A. (2011). Statistical machine translation: A guide for linguists and translators. *Language and Linguistics Compass*, 5(5):205–226.
- [Hewlett et al., 2005] Hewlett, D., Kalyanpur, A., Kovlovski, V., and Halaschek-Wiener, C. (2005). Effective natural language paraphrasing of ontologies on the semantic web. In *End User Semantic Web Interaction Workshop, Galway, Ireland*.
- [Hirst, 2003] Hirst, G. (2003). Ontology and the lexicon. In *Handbook on Ontologies in Information Systems*, pages 209–230. Springer.
- [Hovy et al., 2001] Hovy, E., Ide, N., Frederking, R., Mariani, J., and Zampolli, A. (2001). *Multilingual Information Management*. Pisa, Italy: Giardini Editori e Stampatori and Kluwer Academic Publishers.
- [Huang and Papineni, 2007] Huang, F. and Papineni, K. (2007). Hierarchical system combination for machine translation. In *EMNLP-CoNLL*, pages 277–286.
- [Hudik and Ruopp, 2011] Hudik, T. and Ruopp, A. (2011). The integration of moses into localization industry. In *15th Annual Conference of the EAMT*, pages 47–53.
- [Hutchins, 2007] Hutchins, J. (2007). Machine translation: problems and issues. Presentation. Chelyabinsk, Russia. 18 slides.
- [Ide and Veronis, 1998] Ide, N. and Veronis, J. (1998). Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics. Special Issue on Word Sense Disambiguation*.

- [IEEE, 2000] IEEE (2000). The Authoritative Dictionary of IEEE Standard Terms. Seventh edition, December.
- [ISO, 1985] ISO (1985). Documentation – Guidelines for the establishment and development of multilingual thesauri.
- [ISO, 1986] ISO (1986). Documentation – Guidelines for the establishment and development of monolingual thesauri.
- [ISO, 1992] ISO (1992). Guidance on usability specification and measures, ISO, CD 9241-11.
- [Jang et al., 1999] Jang, M.-G., Myaeng, S. H., and Park, S. Y. (1999). Using mutual information to resolve query translation ambiguities and query term weighting. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 223–229, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Jayaraman and Lavie, 2005] Jayaraman, S. and Lavie, A. (2005). Multi-engine machine translation guided by explicit word matching. In *Proc. of EAMT*, pages 143–152.
- [Jevsikova, 2009] Jevsikova, T. (2009). *Internet Software Localization*. PhD thesis, Vilnius University, Vilnius, Lithuania,.
- [Kansai et al., 1996] Kansai, M. K., Kitamura, M., and Matsumoto, Y. (1996). Automatic extraction of word sequence correspondences in parallel corpora. In *Proc. of the 4th Annual Workshop on Very Large Corpora (WVLC-4)*, pages 79–87.
- [Kersten et al., 2002] Kersten, G. E., Kersten, M., and Rakowski, W. M. (2002). Software and culture: Beyond the internationalization of the interface. *JGIM*, 10(4):86–101.
- [Kilgariff and Rosenzweig, 2000] Kilgariff, A. and Rosenzweig, J. (2000). Framework and results for english senseval. *Special Issue on SENSEVAL. Computers and the Humanities*, pages 15–48.
- [Kim and Kim, 1997] Kim, S. and Kim, Y. (1997). Sentence segmentation for efficient english syntactic analysis. *Journal of Korea Information Science Society*. v24 i8.
- [Kim et al., 2001] Kim, S.-D., Zhang, B.-T., and Kim, Y. T. (2001). Learning-based intrasentence segmentation for efficient translation of long sentences. *Machine Translation*, 16(3):151–174.

BIBLIOGRAPHY

- [Kim and Oh, 2008] Kim, Y.-S. and Oh, Y.-J. (2008). Intra-sentence segmentation based on support vector machines in english-korean machine translation systems. *Expert Syst. Appl.*, 34(4):2673–2682.
- [Kirakowski and Corbett, 1993] Kirakowski, J. and Corbett, M. (1993). Sumi: The software usability measurement inventory. *British Journal of Educational Technology*.
- [Klein, 2001] Klein, M. (2001). Combining and relating ontologies: An analysis of problems and solutions.
- [Koehn, 2005] Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86. AAMT.
- [Koehn and Knight, 2003] Koehn, P. and Knight, K. (2003). Empirical methods for compound splitting. In *EACL '03: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*, pages 187–193, Morristown, NJ, USA. Association for Computational Linguistics.
- [Koehn and Monz, 2006] Koehn, P. and Monz, C. (2006). Manual and automatic evaluation of machine translation between european languages. In *Proceedings of the Workshop on Statistical Machine Translation*, StatMT '06, pages 102–121, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Kuhns, 2007] Kuhns, R. J. (2007). Advanced leveraging: The new generation of tms.
- [Lagoudaki, 2006] Lagoudaki, E. (2006). Translation memory systems: Enlightening users perspective.
- [Landers, 2001] Landers, C. (2001). *Literary Translation: A Practical Guide*. Topics in translation. Multilingual Matters.
- [Lassila and McGuinness, 2001] Lassila, O. and McGuinness, D. L. (2001). The role of frame-based representation on the semantic web. In *Knowledge Systems Laboratory Report KSL-01-02, Stanford University, 2001*; Also appeared as *Linking Electronic Articles in Computer and Information Science, Vol. 6 (2001), No. 005, Linking University*.
- [Lavie et al., 2004] Lavie, A., Sagae, K., and Jayaraman, S. (2004). The significance of recall in automatic metrics for mt evaluation. In *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA-2004)*.

- [Laviosa, 1997] Laviosa, S. (1997). How comparable can “comparable corpora” be? In *Target* 9(2):289-319.
- [Lenat and Guha, 1989] Lenat, D. B. and Guha, R. V. (1989). *Building Large Knowledge-Based Systems; Representation and Inference in the Cyc Project*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.
- [Lesk, 1986] Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, SIGDOC '86, pages 24–26, New York, NY, USA. ACM.
- [Levenshtein, 1965] Levenshtein, V. (1965). Binary codes capable of correcting deletions, insertions, and reversals. doklady akademii nauk sssr, 163(4):845848,. In *Russian. English Translation in Soviet Physics Doklady*, 10(8) p. 707710,.
- [Levin., 1993] Levin., B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago, IL, USA.
- [Li et al., 2003] Li, H., Cao, Y., and Li, C. (2003). Using bilingual web data to mine and rank translations. *IEEE Intelligent Systems*, 18:54–59.
- [Liang and Sini, 2006] Liang, A. and Sini, M. (2006). Mapping agrovoc and the chinese agricultural thesaurus: Definitions, tools, procedures. In *New Review of Hypermedia and Multimedia*, pp. 51 – 62, 12 (1).
- [Liang et al., 2005] Liang, A., Sini, M., Chang, C., Li, S., Lu, W., He, C., and Keizer, J. (2005). The mapping schema from chinese agricultural thesaurus to agrovoc. In *6th Agricultural Ontology Service (AOS) Workshop on Ontologies: the more practical issues and experiences*.
- [Litkowski, 2005] Litkowski, K. C. (2005). Computational Lexicons and Dictionaries. In *Encyclopedia of Language and Linguistics (2nd ed.)*. Elsevier Publishers.
- [Llitjs, 2009] Llitjs, A. F. (2009). *Automatic Improvement of Machine Translation Systems*. VDM Verlag, Saarbrücken, Germany.
- [Lu et al., 2002] Lu, W.-H., Chien, L.-F., and Lee, H.-J. (2002). Translation of web queries using anchor text mining. *ACM Transactions on Asian Language Information Processing (TALIP)*, 1(2):159–172.
- [Malaisé et al., 2007] Malaisé, V., Isaac, A., Gazendam, L., and Brugman, H. (2007). Anchoring dutch cultural heritage thesauri to wordnet: Two case studies. In *Proceedings of the Workshop on Language Technology*

BIBLIOGRAPHY

- for *Cultural Heritage Data (LaTeCH 2007)*., pages 57–64, Prague, Czech Republic. Association for Computational Linguistics.
- [Malmkjær, 2000] Malmkjær, K. (2000). Multidisciplinarity in process research. In *S. Tirkkonen-Condit & R. Jaaskelainen (eds.), Tapping and Mapping the Process of Translation: Outlooks on Empirical Research.*, pages 163–170.
- [Martin et al., 2008] Martin, H., De Leenheer, P., de Moor, A., and Sure, Y. (2008). *Ontology Management, Semantic Web, Semantic Web Services, and Business Applications*. Semantic Web and Beyond. Springer-Verlag, Heidelberg.
- [Maruyama, 1992] Maruyama, H.; Watanabe, H. (1992). Tree cover search algorithm for example-based translation. In *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation, Montreal, 173184*.
- [Massion, 2005] Massion, F. (2005). Translation-memory-systeme im vergleich.
- [Matusov et al., 2006] Matusov, E., Ueffing, N., and Ney, H. (2006). Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment. In *Cambridge University Engineering Department*, pages 33–40.
- [McCrae et al., 2012] McCrae, J., Davis, B., and Gracia, J. (2012). Enriching the web with ontology-lexica.
- [McCrae et al., 2011a] McCrae, J., Espinoza, M., Montiel-Ponsoda, E., Aguado-de Cea, G., and Cimiano, P. (2011a). Combining statistical and semantic approaches to the translation of ontologies and taxonomies. In *Proceedings of the Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation, SSST-5*, pages 116–125, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [McCrae et al., 2011b] McCrae, J., Spohr, D., and Cimiano, P. (2011b). Linking lexical resources and ontologies on the semantic web with lemon. In *Proceedings of the 8th extended semantic web conference on The semantic web: research and applications - Volume Part I, ESWC’11*, pages 245–259, Berlin, Heidelberg. Springer-Verlag.
- [McEnery, 2003] McEnery, A. (2003). Corpus linguistics. In *R. Mitkov (ed.) Oxford handbook of computational linguistics.*, pages 448–63.
- [Melamed et al., 2003] Melamed, I. D., Green, R., and Turian, J. P. (2003). Precision and recall of machine translation. In *NAACL ’03: Proceedings*

- of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, pages 61–63, Morristown, NJ, USA. Association for Computational Linguistics.
- [Mihalcea, 2002] Mihalcea, R. F. (2002). Word sense disambiguation with pattern learning and automatic feature selection. *Nat. Lang. Eng.*, 8(4):343–358.
- [Miles et al., 2005] Miles, A., Matthews, B., Beckett, D., Brickley, D., Wilson, M., and Rogers, N. (2005). Skos: A language to describe simple knowledge structures for the web. In *Proceedings of XTech 2005*.
- [Miller, 1995] Miller, G. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11).
- [Miller and Matthews, 2001] Miller, K. and Matthews, B. (2001). Having the right connections: the limber project. *Journal of Digital information*, vol. 1 issue 8.
- [Mitkov and Corpas, 2008] Mitkov, R. and Corpas, G. (2008). Improving third generation translation memory systems through identification of rhetorical predicates. In *LangTech Proceedings*.
- [Montiel-Ponsoda, 2011a] Montiel-Ponsoda, E. (2011a). *Multilingualism in Ontologies*. PhD thesis, Universidad Politécnica de Madrid, Madrid, España.
- [Montiel-Ponsoda, 2011b] Montiel-Ponsoda, E. (2011b). *Multilingualism in Ontologies - Building Patterns and Representation Models*. LAP Lambert Academic Publishing.
- [Montiel-Ponsoda et al., 2008] Montiel-Ponsoda, E., Aguado, G., Gómez-Pérez, A., and Peters, W. (2008). Modelling multilinguality in ontologies. In *Coling 2008: Companion volume - Posters and Demonstrations, Manchester, UK*.
- [Montiel-Ponsoda et al., 2011] Montiel-Ponsoda, E., de Cea, G. A., Gómez-Pérez, A., and Peters, W. (2011). Enriching ontologies with multilingual information. *Natural Language Engineering*, 17(3):283–309.
- [MSDN, 2012] MSDN (last accessed December 2012). Chapter 1 Understanding Internationalization <http://msdn.microsoft.com/en-us/library/cc194758.aspx>.
- [Mudur and Sharma, 2002] Mudur, S. P. and Sharma, R. (2002). A reference model for software localisation.

BIBLIOGRAPHY

- [Muntés et al., 2012] Muntés, V., Paladini, P., España-Bonet, C., and Màrquez, L. (2012). Context-aware machine translation for software localization. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT12)*, pages 77–80, Trento, Italy.
- [Müuller, 2009] Müuller, E. (2009). Building quality into the localization process. *Multilingual Localization: Getting Started Guide*.
- [Nagao, 1984] Nagao, M. (1984). A framework of a mechanical translation between japanese and english by analogy principle. In *Proc. of the international NATO symposium on Artificial and human intelligence*, pages 173–180, New York, NY, USA. Elsevier North-Holland, Inc.
- [Nie, 2010] Nie, J.-Y. (2010). Cross-language information retrieval. *Synthesis Lectures on Human Language Technologies*, 3(1):1–125.
- [Nie et al., 2001] Nie, J.-Y., Simard, M., and Foster, G. (2001). Multilingual information retrieval based on parallel texts from the web. In *CLEF '00: Revised Papers from the Workshop of Cross-Language Evaluation Forum on Cross-Language Information Retrieval and Evaluation*, pages 188–201, London, UK. Springer-Verlag.
- [Nomoto, 2004] Nomoto, T. (2004). Multi-engine machine translation with voted language model. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 494, Morristown, NJ, USA. Association for Computational Linguistics.
- [Nord, 1997] Nord, C. (1997). Translating as a purposeful activity. Functional approaches explained. UK: St. Jerome.
- [Nord, 2005] Nord, C. (2005). *Text Analysis in Translation; Theory, Methodology and Didactic Application of a Model for Translation-Oriented Text Analysis*. GA: Rodopi., Amsterdam - Atlanta, second edition.
- [Noy et al., 2001] Noy, N. F., Sintek, M., Decker, S., Crubézy, M., Ferguson, R. W., and Musen, M. A. (2001). Creating semantic web contents with protégé-2000. *IEEE Intelligent Systems*, 16(2):60–71.
- [Oard, 1997] Oard, D. W. (1997). Alternative approaches for cross-language text retrieval. In *AAAI Symposium on cross-language text and speech retrieval. American Association for Artificial Intelligence*.
- [Oard and Hackett, 1997] Oard, D. W. and Hackett, P. (1997). Document translation for cross-language text retrieval at the university of maryland. In *The Sixth Text REtrieval Conference (TREC-6). National Institutes of Standards and Technology*.

- [Och, 2002] Och, F. J. (2002). *Statistical Machine Translation: From Single-Word Models to Alignment Templates*. PhD thesis, RWTH Aachen University, Aachen, Germany.
- [Och and Ney, 2003] Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.
- [OGC, 1996] OGC, editor (1996). *The OpenGIS Guide - Introduction to Interoperable Geoprocessing and the OpenGIS Specification*. Open GIS Consortium, Inc, Boston.
- [Ogden and Davis, 2000] Ogden, W. C. and Davis, M. W. (2000). Improving cross-language text retrieval with human interactions.
- [O’Sullivan, 2001] O’Sullivan (2001). *A Paradigm for Creating Multilingual Interfaces*. PhD thesis, University of Limerick, Ireland.
- [Palma et al., 2011] Palma, R., Corcho, Ó., Gómez-Pérez, A., and Haase, P. (2011). A holistic approach to collaborative ontology development based on change management. *J. Web Sem.*, 9(3):299–314.
- [Palma et al., 2008] Palma, R., Haase, P., Corcho, O., Gómez-Pérez, A., and Ji., Q. (2008). An editorial workflow approach for collaborative ontology development. In *3rd Asian Semantic Web Conference. ASWC 08. Bangkok, Thailand*.
- [Palmer and Wu, 1995] Palmer, M. S. and Wu, Z. (1995). Verb semantics for english-chinese translation. *Machine Translation*, 10(1-2):59–92.
- [Papineni et al., 2002] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *ACL ’02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA. Association for Computational Linguistics.
- [Park, 2001] Park, S. B. (2001). Computing consensus translation from multiple machine translation systems. In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU-2001)*, pages 351–354.
- [Paul et al., 2005] Paul, M., Doi, T., Hwang, Y., Imamura, K., Okuma, H., and Sumita, E. (2005). Nobody is perfect: Atr’s hybrid approach to spoken language translation. In *Proc. of IWSLT*, pages 55–62.
- [Pazienza et al., 2005] Pazienza, M., Stellato, A., Zanzotto, F., Henriksen, L., and Paggio, P. (2005). Ontology mapping to support ontology based question answering. In *Proceedings of the Meaning 05 Workshop*.

BIBLIOGRAPHY

- [Pazienza and Stellato, 2006] Pazienza, M. T. and Stellato, A. (2006). Exploiting linguistic resources for building linguistically motivated ontologies in the semantic web. In *Second Workshop on Interfacing Ontologies and Lexical Resources for Semantic Web Technologies (OntoLex2006)*, held jointly with *LREC2006*, May 24-26, 2006, Genoa, (Italy).
- [Pease and Niles, 2002] Pease, A. and Niles, I. (2002). Ieee standard upper ontology: a progress report. *Knowl. Eng. Rev.*, 17(1):65–70.
- [Pedersen et al., 2005] Pedersen, T., Banerjee, S., and Patwardhan, S. (2005). Maximizing Semantic Relatedness to Perform Word Sense Disambiguation. Research Report UMSI 2005/25, University of Minnesota Supercomputing Institute.
- [Pekar and Mitkov, 2007] Pekar, V. and Mitkov, R. (2007). New generation translation memory: Content-sensitive matching. In *Proceedings of the 40th Anniversary Congress of the Swiss Association of Translators, Terminologists and Interpreters*, pages 16–30.
- [Peters and Sheridan, 2000] Peters, C. and Sheridan, P. (2000). Multilingual information access. In *ESSIR*, pages 51–80.
- [Pinto et al., 2004] Pinto, S., Staab, S., Sure, Y., and Tempich, C. (2004). Ontoedit empowering swap: a case study in supporting distributed, loosely-controlled and evolving engineering of ontologies (diligent). In *Proceedings of the First European Semantic Web Symposium, ESWS 2004, Heraklion, Crete, Greece*, pages 16–30.
- [Prikladnicki et al., 2008] Prikladnicki, R., Damian, D., and Audy, J. (2008). Patterns of evolution in the practice of distributed software development: Quantitative results from a systematic review. In *Evaluation and Assessment in Software Engineering (EASE)*, Bari, Italy,.
- [Prins and van den Broek, 2004] Prins, J. and van den Broek, P. (2004). Semantic search ilse media towards a dutch semantic web infrastructure. Master’s thesis, Vrije Universiteit Amsterdam, The Netherlands,.
- [Pustejovsky, 1991] Pustejovsky, J. (1991). The generative lexicon. *Computational Linguistics*, 17(4):409–441.
- [Qu et al., 2012] Qu, J., Shimazu, A., and Nguyen, M. (2012). Oov term translation, context information and definition extraction based on oov term type prediction. In *Advances in Natural Language Processing*, volume 7614 of *Lecture Notes in Computer Science*, pages 76–87. Springer Berlin Heidelberg.

- [Quirk, 2004] Quirk, C. (2004). Training a sentence-level machine translation confidence metric. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, pages 825–828, Lisbon, Portugal.
- [Rada et al., 1989] Rada, R., Mili, H., Bicknell, E., and Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics*, 19(1):17–30.
- [Raghavan and Wong, 1986] Raghavan, V. V. and Wong, S. K. M. (1986). A critical analysis of vector space model for information retrieval. *J. Am. Soc. Inf. Sci.*, 37(5):279–287.
- [Rajlich and Bennett, 2000] Rajlich, V. T. and Bennett, K. H. (2000). A staged model for the software life cycle. *Computer*, 33(7):66–71.
- [Rapp, 1997] Rapp, R. (1997). Text-detektor. fehlertolerantes retrieval ganz einfach. In *Magazin fr Computertechnik*, 4/97, 386–392.
- [Reinke, 2013] Reinke, U. (2013). State of the art in translation memory technology. *Translation: Computation, Corpora, Cognition*, 3(1).
- [Resnik, 1999] Resnik, P. (1999). Disambiguating noun groupings with respect to wordnet senses. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 54–68. Association for Computational Linguistics.
- [Rosti et al., 2007a] Rosti, A.-V. I., Ayan, N. F., Xiang, B., Matsoukas, S., Schwartz, R. M., and Dorr, B. J. (2007a). Combining outputs from multiple machine translation systems. In *HLT-NAACL*, pages 228–235.
- [Rosti et al., 2007b] Rosti, A.-V. I., Matsoukas, S., and Schwartz, R. M. (2007b). Improved word-level system combination for machine translation. In *ACL*.
- [Rosti et al., 2008] Rosti, A.-V. I., Zhang, B., Matsoukas, S., and Schwartz, R. (2008). Incremental hypothesis alignment for building confusion networks with application to machine translation system combination. In *StatMT '08: Proceedings of the Third Workshop on Statistical Machine Translation*, pages 183–186, Morristown, NJ, USA. Association for Computational Linguistics.
- [Ruopp, 2010] Ruopp, A. (2010). The mooses for localization open source project. In *Conference of the AMTA*.
- [Ruparelia, 2010] Ruparelia, N. B. (2010). Software development lifecycle models. *SIGSOFT Softw. Eng. Notes*, 35(3):8–13.

BIBLIOGRAPHY

- [Sánchez and Moreno, 2008] Sánchez, D. and Moreno, A. (2008). Learning non-taxonomic relationships from web documents for domain ontology construction. *Data Knowl. Eng.*, 64:600–623.
- [Sato and Saito, 2002] Sato, K. and Saito, H. (2002). Extracting word sequence correspondences with support vector machines. In *Proceedings of the 19th international conference on Computational linguistics*, pages 1–7, Morristown, NJ, USA. Association for Computational Linguistics.
- [Schober et al., 2007] Schober, D., Kusnierczyk, W., Lewis, S. E., Jane Lomax, M. o. t. M., Groups, P. O. W., Mungall, C., Rocca-Serra, P., Smith, B., and Sansone, S.-A. (2007). Towards naming conventions for use in controlled vocabulary and ontology engineering. In *Bioontology SIG proceedings (ISMB 2007)*.
- [Schwenk and Gauvain, 2000] Schwenk, H. and Gauvain, J.-L. (2000). In proceedings of the iee international conference on speech and language proceesing (icslp), volume 2, pages 915918, beijin, october. ieee. In *Bioontology SIG proceedings (ISMB 2007)*.
- [Segev and Gal, 2008] Segev, A. and Gal, A. (2008). Enhancing portability with multilingual ontology-based knowledge management. *Decis. Support Syst.*
- [Sheu, 1997] Sheu, P.-Y. (1997). Software life cycle models. In *Software Engineering and Environment*, Software Science and Engineering, pages 1–7. Springer US.
- [Sim et al., 2007] Sim, K. C., Byrne, W. J., Gales, M. J. F., Sahbi, H., and Woodl, P. C. (2007). Consensus network decoding for statistical machine translation system combination. In *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*.
- [Sosnovsky and Gavrilova, 2006] Sosnovsky, S. and Gavrilova, T. (2006). Development of educational ontology for c-programming. *International Journal Information Theories and Applications, Volume 13*.
- [Sowa, 1999] Sowa, J. F. (1999). *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Course Technology.
- [Staab et al., 2001] Staab, S., Studer, R., Schnurr, H.-P., and Sure., Y. (2001). Knowledge processes and ontologies. In *IEEE Intelligent Systems*, 16(1):26–34.
- [Starren and Thelen, 1988] Starren, P. and Thelen, M. (1988). General dictionaries and students of translation: A report on the use of dictionaries in the translation process. In *Proceedings BudaLEX88*.

- [Steinberger et al., 2002] Steinberger, R., Pouliquen, B., and Hagman, J. (2002). Cross-lingual document similarity calculation using the multilingual thesaurus eurovoc. In *Computational Linguistics and Intelligent Text Processing, Third International Conference, CICLing'2002*.
- [Stroppa et al., 2007] Stroppa, N., van den Bosch, A., and Way, A. (2007). Exploiting source similarity for smt using context-informed features. In *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation Skvde, Sweden, pp.231-240*.
- [Studer et al., 1998] Studer, R., Benjamins, V. R., and D., F. (1998). Knowledge engineering principles and methods. In *IEEE Transactions on Knowledge and Data Engineering* 25(1-2), 161-197.
- [Sturm, 2002] Sturm, C. (2002). Tlcc - towards a framework for systematic and successful product internationalization. In *4th International Workshop on Internationalisation of Products and Systems*, pages 61–70, Austin/Texas, USA. Springer-Verlag.
- [Suarez-Figueroa, 2013] Suarez-Figueroa, M. (2013). *NeOn Methodology for Building Ontology Networks: Specification, Scheduling and Reuse*. Dissertations in Artificial Intelligence / Dissertationen Zur Kunstlichen Intelligenz. IOS Press, Incorporated.
- [Suárez-Figueroa and Gómez-Pérez, 2008] Suárez-Figueroa, M. and Gómez-Pérez, A. (2008). Towards a glossary of activities in the ontology engineering field. In *LREC*.
- [Suárez-Figueroa, 2010] Suárez-Figueroa, M. C. (2010). *NeOn Methodology for Building Ontology Networks: Specification, Scheduling and Reuse*. PhD thesis, Universidad Politécnica de Madrid, Madrid, Spain.
- [Suárez-Figueroa and Gómez-Pérez, 2008] Suárez-Figueroa, M. C. and Gómez-Pérez, A. (2008). First attempt towards a standard glossary of ontology engineering terminology. In *Proc. of 8th International Conference on Terminology and Knowledge Engineering (TKE'08)*.
- [Sure et al., 2002] Sure, Y., Angele, J., and Staab, S. (2002). On-toedit: Guiding ontology development by methodology and inferencing. In *On the Move to Meaningful Internet Systems, 2002 - DOA/CoopIS/ODBASE 2002 Confederated International Conferences DOA, CoopIS and ODBASE 2002*, pages 1205–1222, London, UK. Springer-Verlag.
- [Teixeira, 2011] Teixeira, C. (2011). Knowledge of provenance and its effects on translation performance in an integrated tm/mt environment. In *Proceedings of the 8th International NLPCS Workshop - Special theme:*

BIBLIOGRAPHY

- Human-Machine Interaction in Translation*. Copenhagen Studies in Language.
- [Tjoa et al., 2005] Tjoa, A. M., Andjomshoaa, A., Shayeganfar, F., and Wagner, R. (2005). Semantic web challenges and new requirements. In *Proceedings of the 16th International Workshop on Database and Expert Systems Applications (DEXA'05)*.
- [TreeTagger, 1997] TreeTagger (1997). <http://www.ims.uni-stuttgart.de/projekte/corplex/>.
- [Trillo et al., 2007] Trillo, R., Gracia, J., Espinoza, M., and Mena, E. (2007). Discovering the semantics of user keywords. *Journal on Universal Computer Science. Special Issue: Ontologies and their Applications*.
- [Trojahn et al., 2008] Trojahn, C., Quaresma, P., and Bieira, R. (2008). A framework for multilingual ontology mapping. In *Proceedings of the 6th edition of the Language Resources and Evaluation Conference*, pp. 1034-1037.
- [Tudorache et al., 2008] Tudorache, T., Noy, N., Tu, S., and Musen., M. (2008). Supporting collaborative ontology development in protege. In *International Semantic Web Conference*.
- [Ueffing et al., 2003] Ueffing, N., Macherey, K., and Ney, H. (2003). Confidence measures for statistical machine translation. In *Proc. MT Summit IX*, pages 394–401. Springer-Verlag.
- [Uschold and Gruninger, 1996] Uschold, M. and Gruninger, M. (1996). Ontologies. principles, methods and applications. In *Knowledge Engineering Review*, 11(2), 93-155.
- [Uschold et al., 1998] Uschold, M., King, M., Moralee, S., and Zorgios, Y. (1998). The enterprise ontology. *Knowl. Eng. Rev.*, 13(1):31–89.
- [Utsuro et al., 2003] Utsuro, T., Koduma, Y., Watanabe, T., Nishizaki, H., and Nakagawa, S. (2003). Confidence of Agreement Among Multiple LVCSR Models and Model Combination by SVM. In *Int. Conf. on Acoustics, Speech and Signal Processing*.
- [van Heijst et al., 1997] van Heijst, G., Schreiber, A. T., and Wielinga, B. J. (1997). Using explicit ontologies in kbs development. *Int. J. Hum.-Comput. Stud.*, 46(2-3):183–292.
- [Varó and Linares, 1997] Varó, E. A. and Linares, M. M. (1997). *Diccionario de Lingüística Moderna*. Ariel.
- [Vas, 2007] Vas, R. (2007). Educational ontology and knowledge testing. *The Electronic Journal of Knowledge Management*, 5(1):123–130.

- [Vashee, 2007] Vashee, K. (2007). Statistical machine translation and translation memory: An integration made in heaven! *ClientSide News Magazine*, 7(6):18–20.
- [Vossen, 1997] Vossen, P. (1997). Eurowordnet: a multilingual database for information retrieval. In *Proceedings of the workshop on Cross-language Information Retrieval, Zurich*.
- [Vossen, 2004] 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.
- [Way, 2001] Way, A. (2001). *Lfg-dot: A hybrid architecture for robust MT*. PhD thesis, University of Essex, UK.
- [Wilks, 2009] Wilks, Y. (2009). *Machine Translation: Its Scope and Limits*. Springer, New York.
- [Wilks and Stevenson, 1997] Wilks, Y. and Stevenson, M. (1997). Sense tagging: Semantic tagging with a lexicon. In *Proceedings of the SIGLEX Workshop*, pages 47–51.
- [Willett and Angell, 1983] Willett, P. and Angell, R. (1983). Automatic spelling correction using a trigram similarity measure. *Information Processing & Management* 19, pages 255–261.
- [Wong and Yao, 1995] Wong, S. K. M. and Yao, Y. Y. (1995). On modeling information retrieval with probabilistic inference. *ACM Trans. Inf. Syst.*, 13:38–68.
- [Wright, 2011] Wright, D. (2011). *Software Life Cycle Management*. It Governance Ltd.
- [Wright and Budin, 1997] Wright, S. and Budin, G. (1997). *Handbook of Terminology Management: Basic aspects of terminology management*. Number v. 1 in Handbook of Terminology Management. Lightning Source Incorporated.
- [Wu et al., 2008a] Wu, D., He, D., Ji, H., and Grishman, R. (2008a). A study of using an out-of-box commercial mt system for query translation in clir. In *CIKM-iNEWS*, pages 71–76.
- [Wu et al., 2008b] Wu, D., He, D., Ji, H., and Grishman, R. (2008b). A study of using an out-of-box commercial mt system for query translation in clir. In *Proceedings of the 2nd ACM workshop on Improving non english web searching, iNEWS '08*, pages 71–76, New York, NY, USA. ACM.

BIBLIOGRAPHY

- [Yang and Li, 2003] Yang, C. and Li, K. (2003). Automatic construction of english/chinese parallel corpora. In *Journal of the American Society for Information Science and Technology*.
- [Yang, 2007] Yang, Y. (2007). Extending the user experience to localized products. In *Second International Conference on Usability and Internationalization, UI-HCII 2007, Held as Part of HCI International, Beijing, China, July*.
- [Zhang et al., 2005] Zhang, Y., Huang, F., and Vogel, S. (2005). Mining translations of oov terms from the web through cross-lingual query expansion. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 669–670, New York, NY, USA. ACM.
- [Zhang et al., 2002] Zhang, Z., Zhang, C., and Ong, S. (2002). Building an ontology for financial investment. In *Intelligent Data Engineering and Automated Learning (IDEAL 2000), Data Mining, Financial Engineering, and Intelligent Agents, Second International Conference*.

Appendix A

Ontology Localization Framework Evaluation

Software Usability Measurement Inventory (SUMI) questionnaire to assess the usability of the LabelTranslator system.

A.1 Efficiency

1. This software responds too slowly to inputs.
 - Agree
 - Undecided
 - Disagree
2. I would recommend this software to my colleagues.
 - Agree
 - Undecided
 - Disagree
3. The instructions and prompts are helpful.
 - Agree
 - Undecided
 - Disagree
4. The software has at some time stopped unexpectedly.
 - Agree
 - Undecided
 - Disagree

5. Learning to operate this software initially is full of problems.
 - Agree
 - Undecided
 - Disagree
6. I sometimes dont know what to do next with this software.
 - Agree
 - Undecided
 - Disagree
7. I enjoy my sessions with this software.
 - Agree
 - Undecided
 - Disagree
8. I find that the help information given by this software is not very useful.
 - Agree
 - Undecided
 - Disagree
9. If this software stops, it is not easy to restart it.
 - Agree
 - Undecided
 - Disagree
10. It takes too long to learn the software commands.
 - Agree
 - Undecided
 - Disagree

A.2 Affect

1. I sometimes wonder if Im using the right command.
 - Agree
 - Undecided
 - Disagree

2. Working with this software is satisfying.
 - Agree
 - Undecided
 - Disagree
3. The way that system information is presented is clear and understandable.
 - Agree
 - Undecided
 - Disagree
4. I feel safer if I use only a few familiar commands or operations.
 - Agree
 - Undecided
 - Disagree
5. The software documentation is very informative.
 - Agree
 - Undecided
 - Disagree
6. This software seems to disrupt the way I normally like to arrange my work.
 - Agree
 - Undecided
 - Disagree
7. Working with this software is mentally stimulating.
 - Agree
 - Undecided
 - Disagree
8. There is never enough information on the screen when its needed.
 - Agree
 - Undecided
 - Disagree
9. I feel in command of this software when I am using it.

- Agree
 - Undecided
 - Disagree
10. I prefer to stick to the facilities that I know best.
- Agree
 - Undecided
 - Disagree

A.3 Helpfulness

1. I think this software is inconsistent.
- Agree
 - Undecided
 - Disagree
2. I would not like to use this software every day.
- Agree
 - Undecided
 - Disagree
3. I can understand and act on the information provided by this software.
- Agree
 - Undecided
 - Disagree
4. This software is awkward when I want to do something which is not standard.
- Agree
 - Undecided
 - Disagree
5. There is too much to read before you can use the software.
- Agree
 - Undecided
 - Disagree

A.4. CONTROL

6. Tasks can be performed in a straightforward manner using this software.
 - Agree
 - Undecided
 - Disagree
7. Using this software is frustrating.
 - Agree
 - Undecided
 - Disagree
8. The software has helped me overcome any problems I have had in using it.
 - Agree
 - Undecided
 - Disagree
9. The speed of this software is fast enough.
 - Agree
 - Undecided
 - Disagree
10. I keep having to go back to look at the guides.
 - Agree
 - Undecided
 - Disagree

A.4 Control

1. It is obvious that user needs have been fully taken into consideration.
 - Agree
 - Undecided
 - Disagree
2. There have been times in using this software when I have felt quite tense.
 - Agree

- Undecided
 - Disagree
3. The organisation of the menus or information lists seems quite logical.
- Agree
 - Undecided
 - Disagree
4. The software allows the user to be economic of keystrokes.
- Agree
 - Undecided
 - Disagree
5. Learning how to use new functions is difficult.
- Agree
 - Undecided
 - Disagree
6. There are too many steps required to get something to work.
- Agree
 - Undecided
 - Disagree
7. I think this software has made me have a headache on occasion.
- Agree
 - Undecided
 - Disagree
8. Error prevention messages are not adequate.
- Agree
 - Undecided
 - Disagree
9. It is easy to make the software do exactly what you want.
- Agree
 - Undecided
 - Disagree

10. I will never learn to use all that is offered in this software.

- Agree
- Undecided
- Disagree

A.5 Learnability

1. The software hasnt always done what I was expecting.

- Agree
- Undecided
- Disagree

2. The software has a very attractive presentation.

- Agree
- Undecided
- Disagree

3. Either the amount or quality of the help information varies across the system.

- Agree
- Undecided
- Disagree

4. It is relatively easy to move from one part of a task to another.

- Agree
- Undecided
- Disagree

5. It is easy to forget how to do things with this software.

- Agree
- Undecided
- Disagree

6. This software occasionally behaves in a way which cant be understood.

- Agree
- Undecided
- Disagree

7. This software is really very awkward.

- Agree
- Undecided
- Disagree

8. It is easy to see at a glance what the options are at each stage.

- Agree
- Undecided
- Disagree

9. Getting data files in and out of the system is not easy.

- Agree
- Undecided
- Disagree

10. I have to look for assistance most times when I use this software.

- Agree
- Undecided
- Disagree

Appendix B

Localization User Guides

B.1 User Guide for Localization Managers

The Localization Manager's User Guide demonstrates i) how to install the environment for the collaborative ontology localization scenario, ii) how to set preferences of the ontology localization activity, iii) how to import the ontology to be localized, iv) how to set localization parameters, and v) how to select ontology labels to be localized.

B.1.1 Installing the Environment.

Neon Toolkit installation

1. Open a Web browser and install Neon Toolkit from <http://neon-toolkit.org/wiki/download>. For Windows, Linux or Mac operating system choose Basic (Installer) version.
2. Fill fields marked with asterisk in the license Web site.

components that are being licensed under an "open source license". Ontoprise is licensed under the present license agreement. The licensor in such a case is source software. [licenses-3rdparty] contains a list of components of the licenses of such licenses. This software is provided "as is", without a warranty or warranties are hereby excluded. Ontoprise will not be liable for any damages ontoprise be liable for any lost revenue, profit or data, or for indirect, special agreement is subject to German law, excluding is conflicts of law provisions. disputes arising out of or in connection with this license agreement. Copyright

Fields marked with an asterisk (*) are required.

First Name: *

Last Name: *

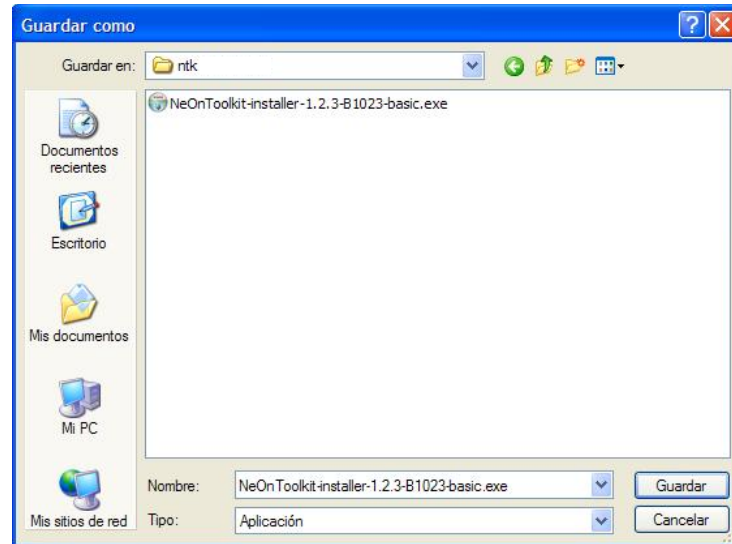
Organisation: *

Email: *

☐ Notify me via e-mail of new NeOn Toolkit releases.

☒ I agree to the above terms.

3. Wait while Neon Toolkit is downloaded and then save the file into any directory (installation directory).



4. Execute the install file from installation directory and complete the Neon Toolkit setup wizard.

LabelTranslator plugins installation

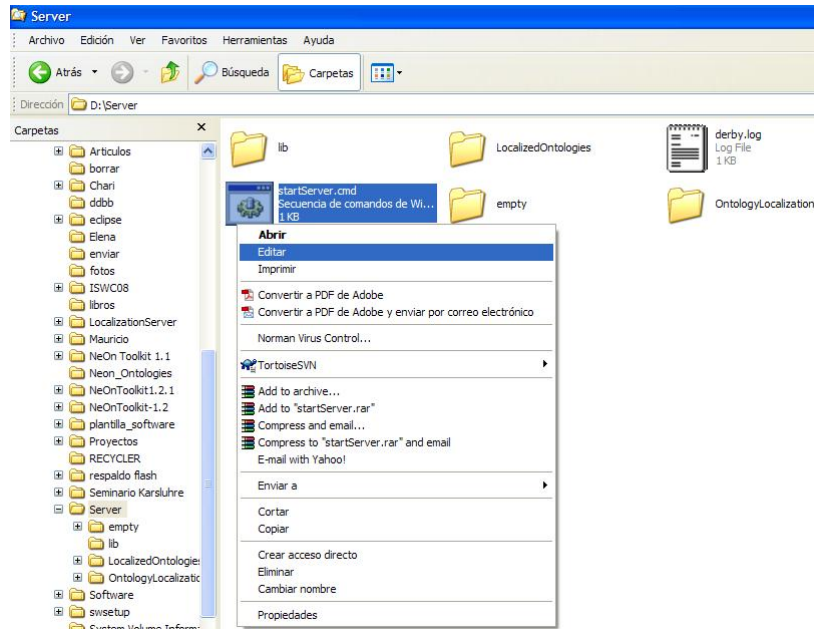
1. Open a Web browser and download LabelTranslator plugins from <http://delicias.dia.fi.upm.es/repos/collaborativelabeltranslator/plugins>
2. Input login and password to access to plugins repository.
3. Select *plugins.zip* file and save the file into any directory.
4. Wait while LabelTranslator plugins are downloaded and unpackged the plugins into Neon Toolkit installation directory (see step 3 in Neon Toolkit installation) .
5. Check the installed files into Neon Toolkit plugins directory.

Localization server installation

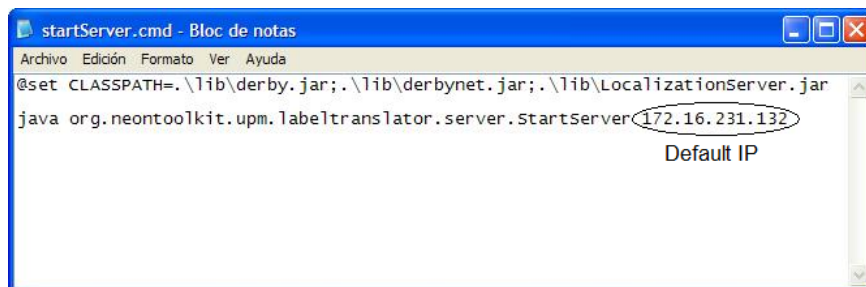
1. Open a Web browser and download Localization Server from <http://delicias.dia.fi.upm.es/repos/collaborativelabeltranslator/server>
2. Select *server.zip* file and save the file into any directory.
3. Wait while *server.zip* file is downloaded and unpackged the file into any directory of the server machine.

B.1. USER GUIDE FOR LOCALIZATION MANAGERS

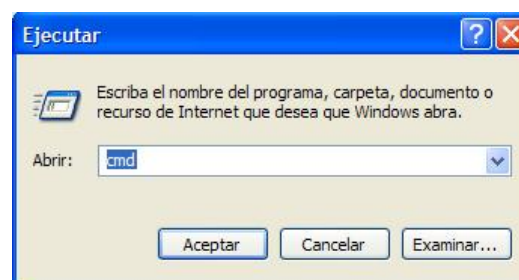
4. In the installation directory of the Localization Server, edit the IP address of the *starServer.cmd* file.



5. Change the default IP by the IP of the machine.



6. For find out the IP of the machine open a window command selecting Start/Execute. In the dialog window, input the *cmd* command



7. In the new command window input the *ipconfig* command



```

C:\WINDOWS\system32\cmd.exe
Microsoft Windows XP [Versión 5.1.2600]
(C) Copyright 1985-2001 Microsoft Corp.

C:\Documents and Settings\Mauricio>ipconfig

Configuración IP de Windows

Adaptador Ethernet Conexiones de red inalámbricas :

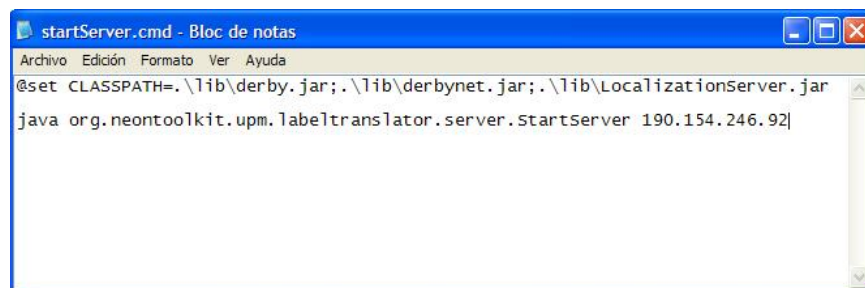
    Estado de los medios. . . .: medios desconectados

Adaptador Ethernet Conexión de área local :

    Sufijo de conexión específica DNS : cpe.satnet.net
    Dirección IP. . . . . : 190.154.246.92
    Máscara de subred . . . . . : 255.255.255.0
    Puerta de enlace predeterminada : 190.154.246.1

C:\Documents and Settings\Mauricio>
  
```

8. Replace the obtained IP address in the *startServer.cmd* file.

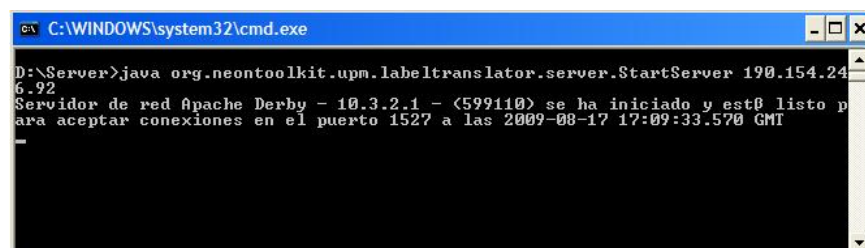


```

startServer.cmd - Bloc de notas
Archivo Edición Formato Ver Ayuda

@set CLASSPATH=.\lib\derby.jar;.\lib\derbynet.jar;.\lib\LocalizationServer.jar
java org.neontoolkit.upm.labeltranslator.server.StartServer 190.154.246.92
  
```

9. Close the command window (see step 7).
10. Save the changes into *startServer.cmd* file and close the window.
11. Execute the Localization Server by double-click on *startServer.cmd* file.
12. Do not close the new window command for maintaining the execution of the Localization Server (Note: You can only minimize the window).



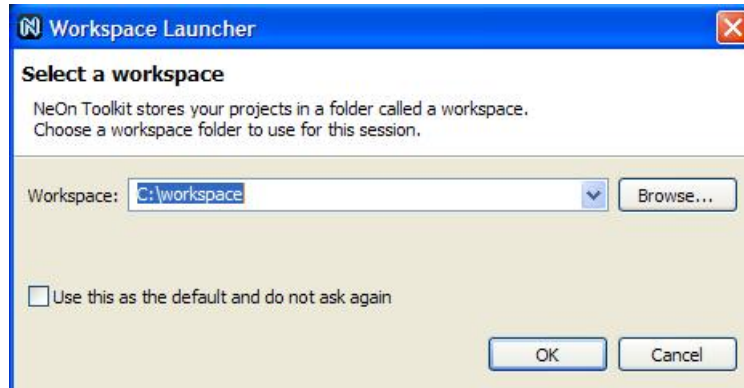
```

C:\WINDOWS\system32\cmd.exe

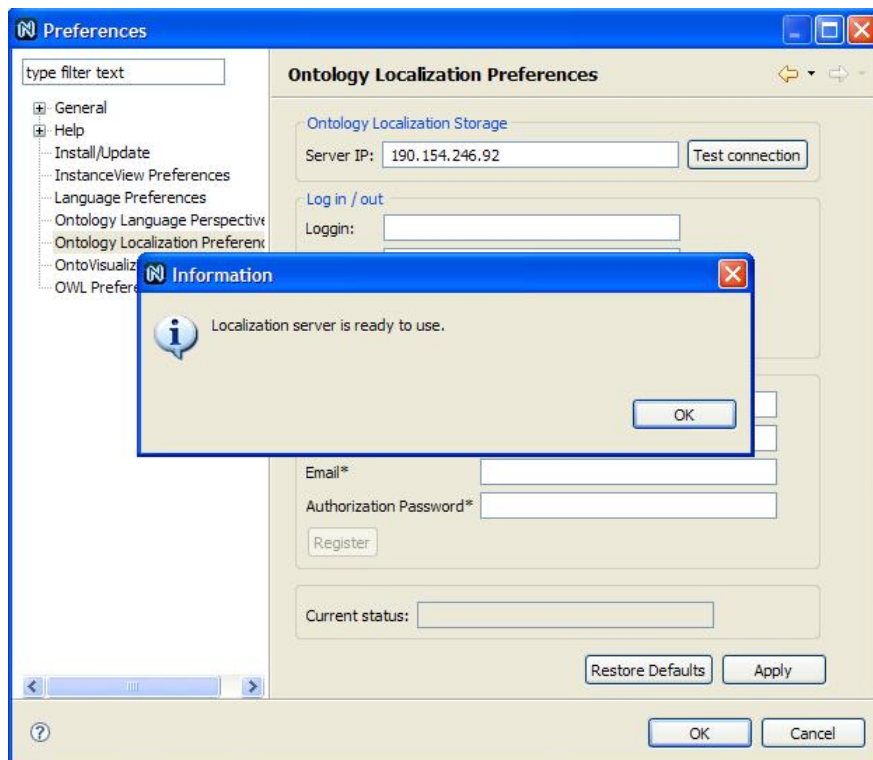
D:\Server>java org.neontoolkit.upm.labeltranslator.server.StartServer 190.154.246.92
Servidor de red Apache Derby - 10.3.2.1 - (599110) se ha iniciado y está listo para aceptar conexiones en el puerto 1527 a las 2009-08-17 17:09:33.570 GMT
  
```

B.1.2 Setting-up Ontology Localization Preferences.

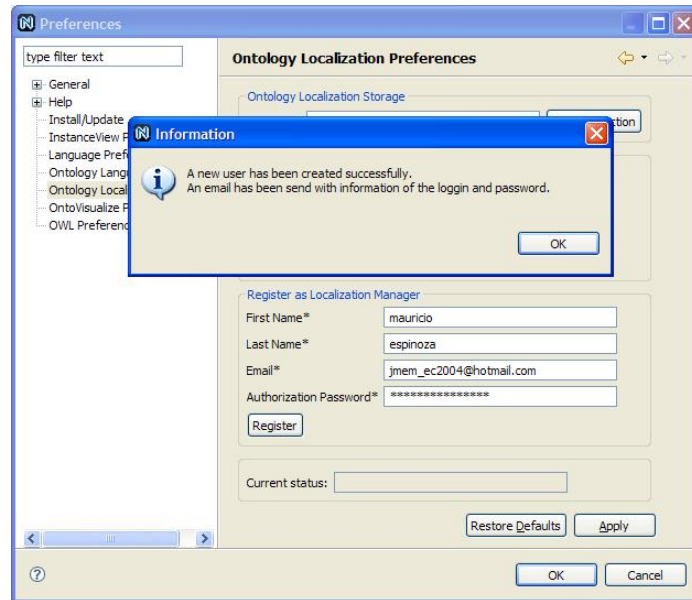
1. Execute Neon Toolkit application and select a workspace.



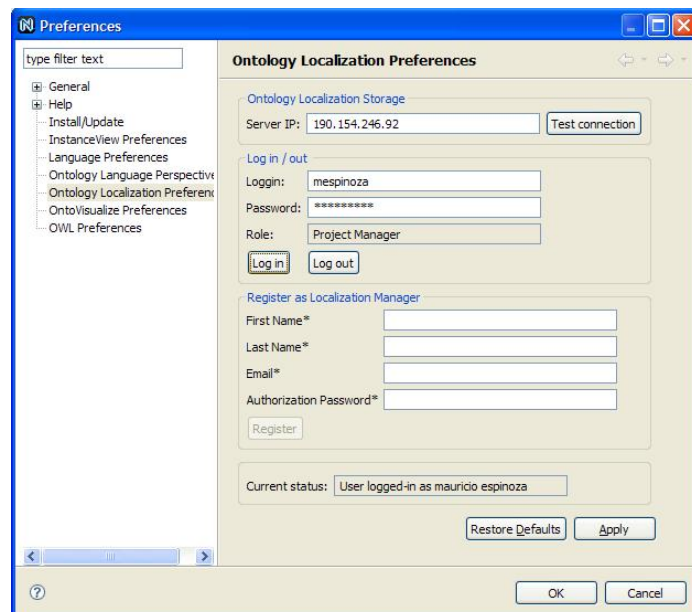
2. Select the Ontology Localization Preferences in the Window/Preferences menu.
3. Input the IP address where the Localization Server is executing (see step 8 in the previous section) and test the connection with the Localization Server.



4. Create a new Localization Manager, filling the fields marked with asterisk in the Ontology Localization Preferences window.

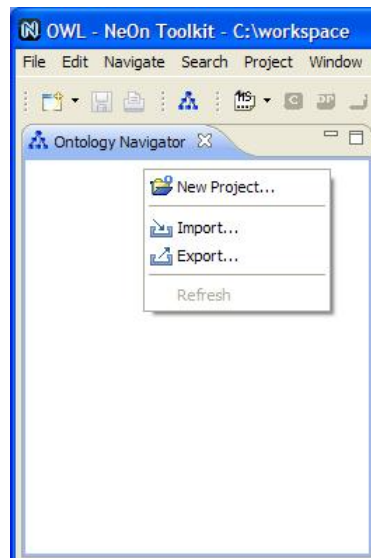


5. Read the email for obtaining the login and password of the localization manager account.
6. Fill the *login* and *password* fields in the Ontology Localization Preferences window and then click on *Login* button.

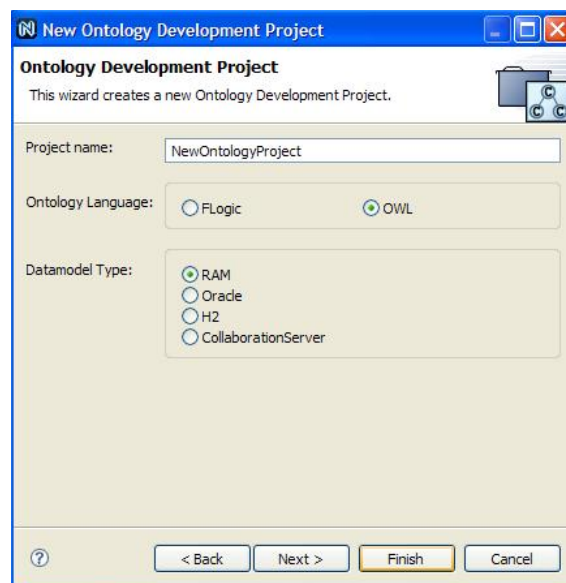


B.1.3 Importing Ontology to be Localized.

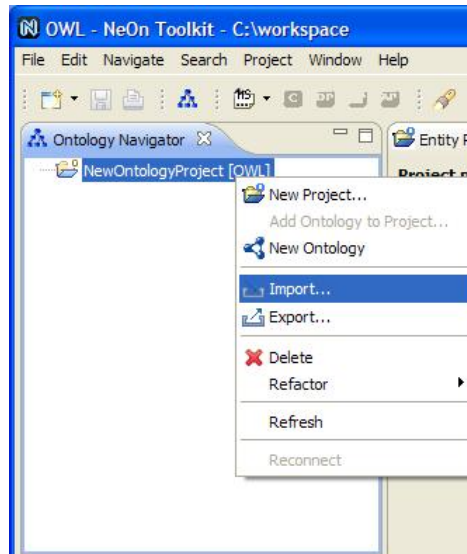
1. In the Neon Toolkit, by right click on the Ontology Navigator select the New Project option.



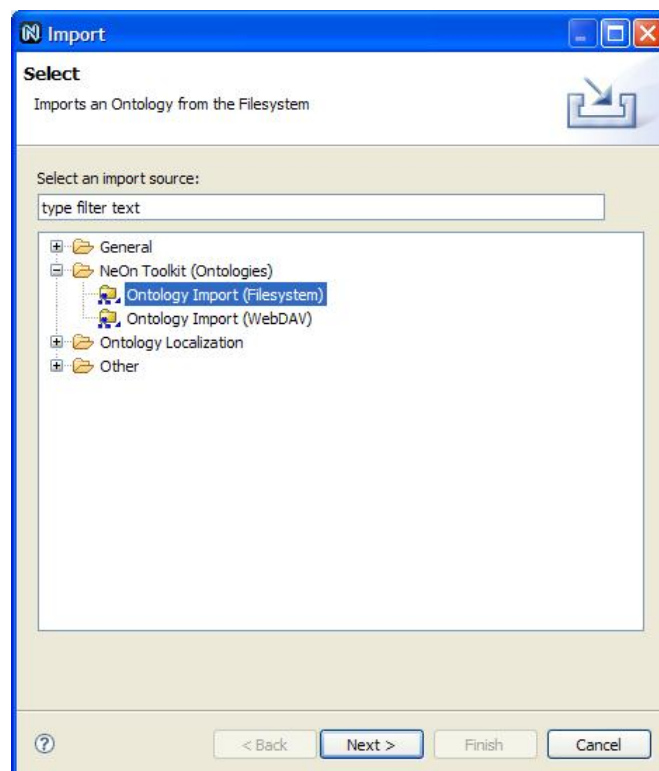
2. In the wizard follow the instructions, choosing *OWL* as ontology language and *RAM* as datamodel type. Click on *Finish* button.



3. By right click on new project select the *Import...* option.

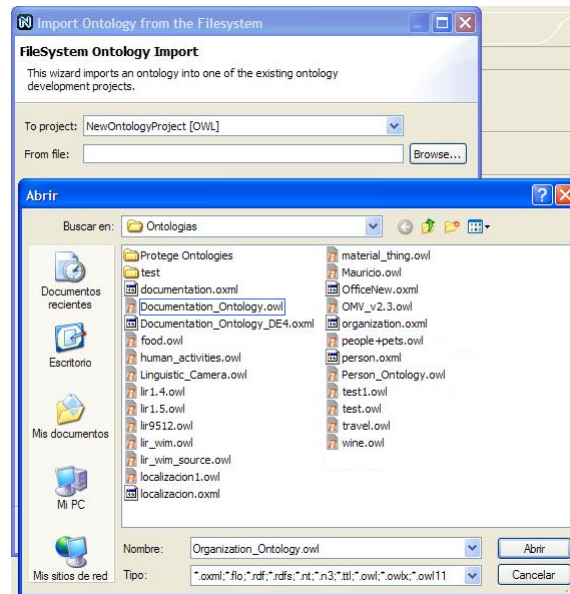


4. Select the *Ontology Import (FileSystem)* and follow the instructions of the wizard.



5. Choose the *Documentation_Ontology.owl* file.

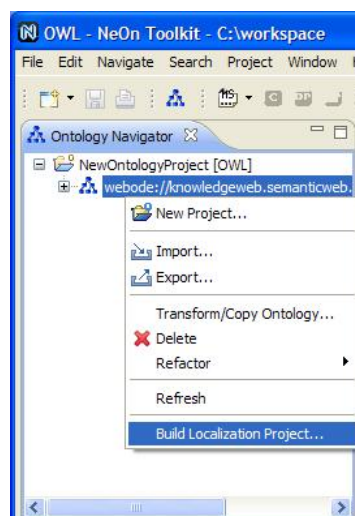
B.1. USER GUIDE FOR LOCALIZATION MANAGERS



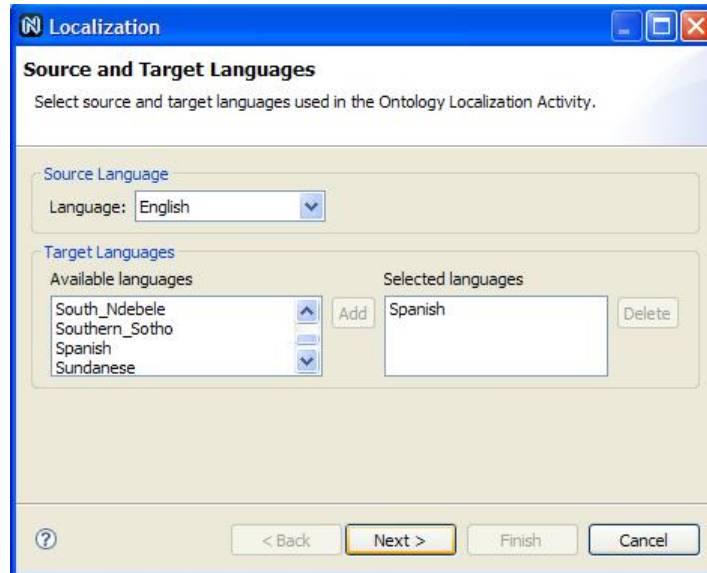
6. On the Ontology Navigator, select the imported ontology and click on save icon to finish the importation.

B.1.4 Setting-up Localization Parameters.

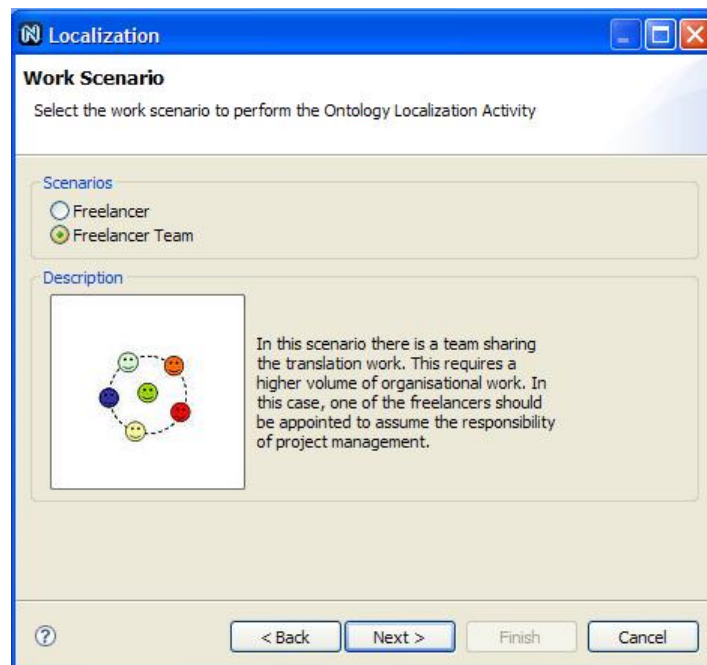
1. In the Window/Open Perspective menu, change the perspective to localization.
2. By right click on the imported ontology select the *Build Localization Project...* option, to configure the parameters of a new ontology localization project.



3. Select source and target languages. For our experiment choose English and Spanish as source and target languages respectively.



4. Select as work scenario *Freelancer Team*. In this scenario there is a team sharing the translation work.



5. Add the actors for executing the localization activity.

Localization

Users

Adds users to Ontology Localization Activity

New user

First Name* Last Name*

Email*

Role* ☒ Translator ☒ Reviewer

Language skills* ☐ English-Southern Sotho ☒ English-Spanish ☐ English-Sundanese ☐ English-Swahili

Fields marked with * are mandatory.

Available Users

Name	Email	Role	Skills

? < Back Next > Finish Cancel

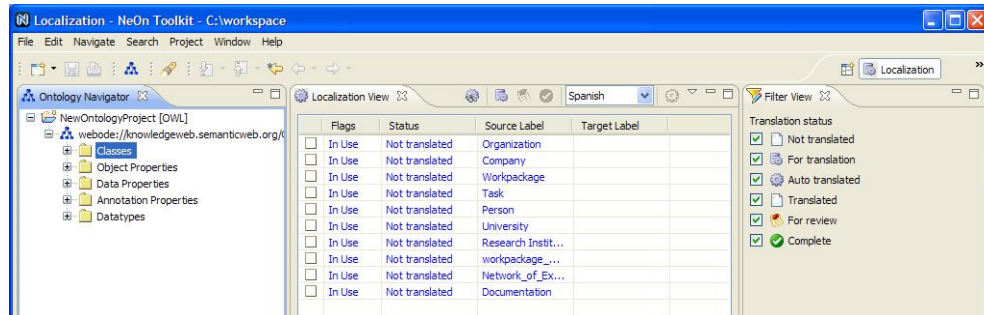
6. Select the users for executing both translation and revision tasks.

[illegible]

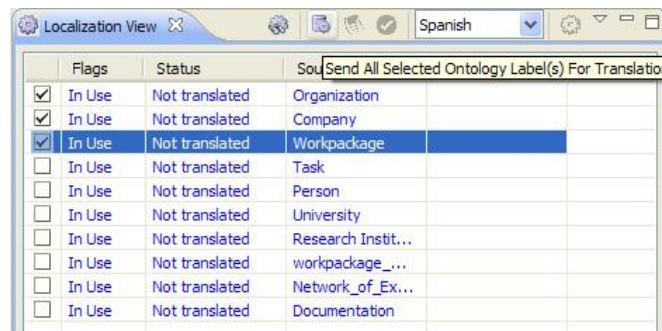
- To finish the configuration, click on *Finish* button and wait while the ontology project is created.

B.1.5 Selecting Ontology Labels to be Translated.

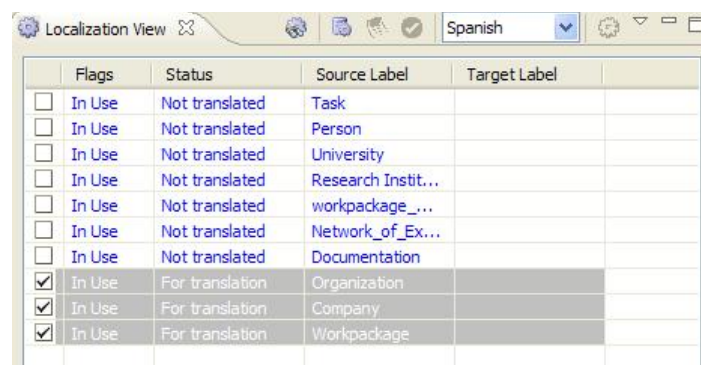
1. In the Ontology Navigator, select the classes, object properties, or data properties to be localized.



2. Send the selected ontology labels to translation.



3. The status of the selected labels is changed to *For Translation* and the labels are disabled.

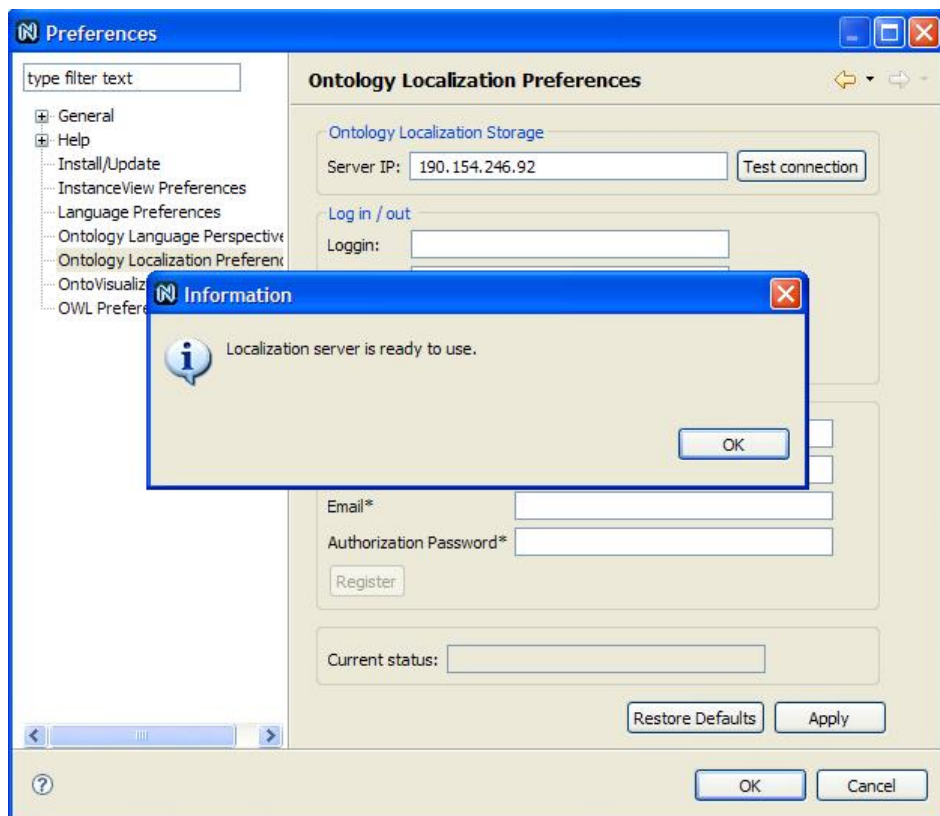


B.2 User Guide for Translators

This user guide demonstrates how to translate automatically ontology labels. The guide is divided into two parts. The first part demonstrates how to login in the system as Translator. The second part contains a complete reference describing the whole process to translate the ontology labels sent by the Localization Manager.

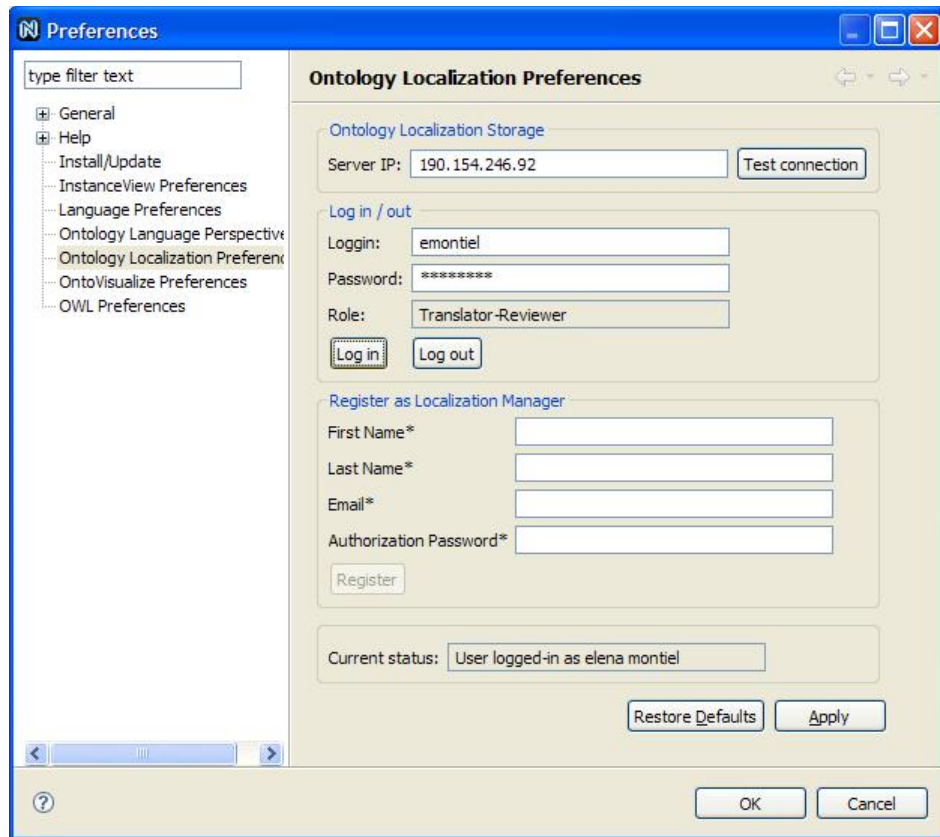
B.2.1 Setting-up Translation Preferences.

1. Start Neon Toolkit.
2. Select Ontology Localization Preferences in the Window/Preferences menu.
3. Input the IP address where the Localization Server is executing (see step 8 in the Localization Server installation) and test the connection with the Localization Server.



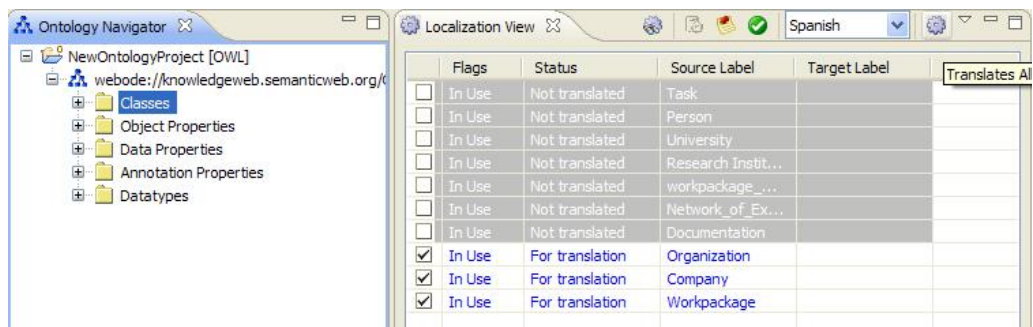
4. Read the email for obtaining the login and password of the translator account.

5. Fill the *login* and *password* fields in the Ontology Localization Preferences window and then click on *Login* button.



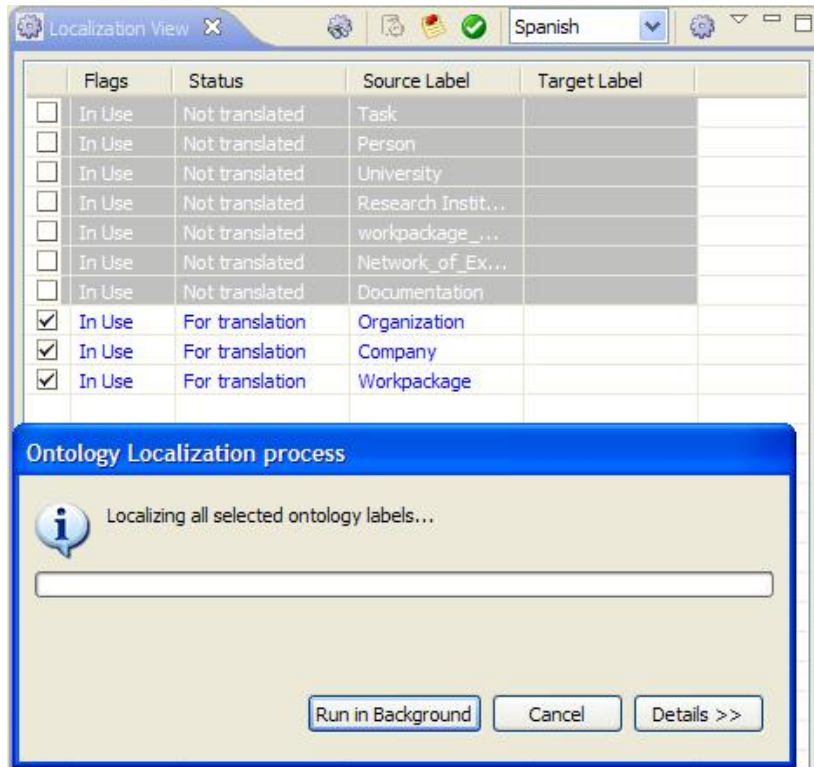
B.2.2 Translating Ontology Labels.

1. In the Ontology Navigator, select the classes, object properties, or data properties to be translated.



B.2. USER GUIDE FOR TRANSLATORS

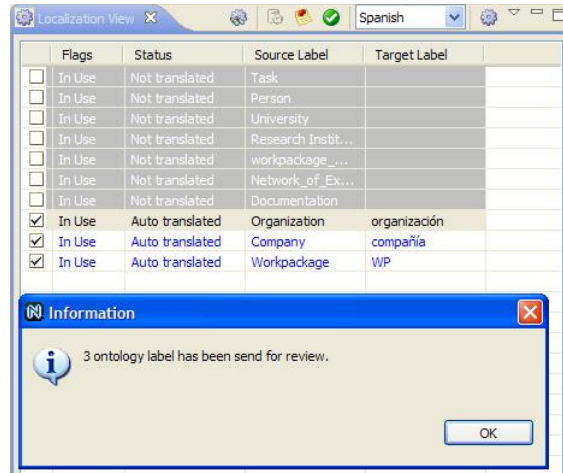
2. Click on *Translator* icon to translate automatically the selected labels and wait while the labels are translated.



3. Select the more appropriate translations from the list of translations obtained by the system.



4. Send the ontology labels to *Revision* status.



5. The status is changed to *For Review*.



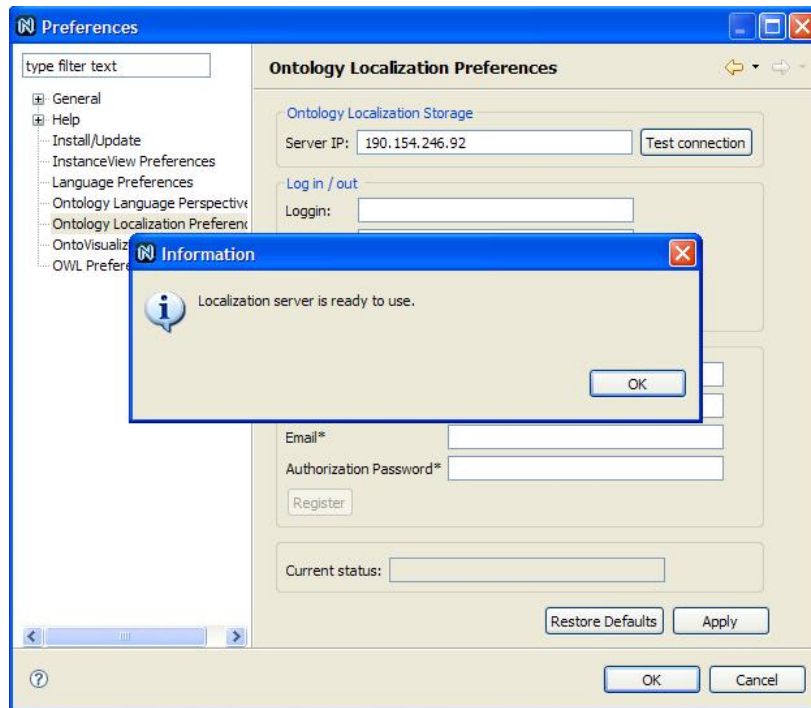
B.3 User Guide for Reviewers

This user guide demonstrates how to review the translations sent by the Translator(s). The guide is divided into two parts. The first part demonstrates how to login in the system as Reviewer. The second part describes the whole process to edit the translations that contain errors.

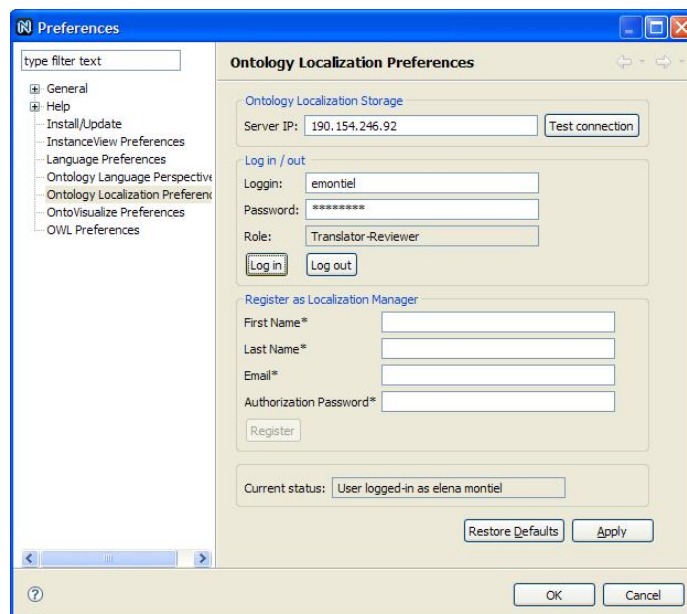
B.3.1 Setting-up Revision Preferences.

1. Start Neon Toolkit.
2. Select the Ontology Localization Preferences in the Window/Preferences menu.
3. Input the IP address where the Localization Server is executing (see step 8 in the Localization Server installation) and test the connection with the Localization Server.

B.3. USER GUIDE FOR REVIEWERS



4. Read the email for obtaining the login and password of the reviewer account.
5. Fill the *login* and *password* fields in the Ontology Localization Preferences window and then click on *Login* button.



B.3.2 Reviewing Translations.

1. In the Localization View, review the ontology labels sent by the Translator and edit the translations that need to be corrected.
2. Send the revised labels to *Complete* status.

