

AUTOMATIC CORPUS-BASED TRANSLATION
OF A SPANISH FRAMENET MEDICAL GLOSSARY

COLECCIÓN LINGÜÍSTICA

DIRECTOR DE LA COLECCIÓN

Cano Aguilar, Rafael. Universidad de Sevilla

COMITÉ CIENTÍFICO

Anscombe, Jean-Claude. CNRS y Université Paris 13
Borreguero Zuloaga, Margarita. Universidad Complutense de Madrid
Cabrillana Leal, Concepción. Universidad de Santiago de Compostela
Crespo Güemes, Emilio. Universidad Autónoma de Madrid
Donaire Fernández, María Luisa. Universidad de Oviedo
Fierro Bello, M^a Isabel. CSIC
Geeraerts, Dirk. Universidad de Lovaina
Girón Alconchel, José Luis. Universidad Complutense de Madrid
Kabatek, Johannes. Universidad de Zúrich
Larreta Zulategui, Juan Pablo. Universidad Pablo de Olavide
Martínez Vázquez, Montserrat. Universidad Pablo de Olavide
Moreno Cabrera, Juan Carlos. Universidad Autónoma de Madrid
Martín, Salvador. Universidad de Málaga
Pompei, Anna. Università di Roma III
Schierholz, Stefan. Universidad de Erlangen-Nürnberg
Simone, Raffaele. Università di Roma III
Torrego Salcedo, Esperanza. Universidad Autónoma de Madrid

CONSEJO DE REDACCIÓN

Bruña Cuevas, Manuel. Universidad de Sevilla
Cano Aguilar, Rafael. Universidad de Sevilla
Carrera Díaz, Manuel. Universidad de Sevilla
Comesaña Rincón, Joaquín. Universidad de Sevilla
Falque Rey, Emma. Universidad de Sevilla.
González Ferrín, Emilio. Universidad de Sevilla.
López Serena, Araceli. Universidad de Sevilla
Martos Ramos, José Javier. Universidad de Sevilla
Ruiz Yamuza, Emilia Reyes. Universidad de Sevilla
Salguero Lamillar, Francisco José. Universidad de Sevilla

MARIO CRESPO MIGUEL

AUTOMATIC CORPUS-BASED TRANSLATION
OF A SPANISH FRAMENET MEDICAL
GLOSSARY



Sevilla 2021

Colección Lingüística
Núm.: 65

COMITÉ EDITORIAL:

Araceli López Serena
(Directora de la Editorial Universidad de Sevilla)
Elena Leal Abad
(Subdirectora)

Concepción Barrero Rodríguez
Rafael Fernández Chacón
María Gracia García Martín
Ana Ilundáin Larrañeta
María del Pópulo Pablo-Romero Gil-Delgado
Manuel Padilla Cruz
Marta Palenque Sánchez
María Eugenia Petit-Breuilh Sepúlveda
José-Leonardo Ruiz Sánchez
Antonio Tejedor Cabrera

Esta publicación, que ha contado con el apoyo y financiación del Contrato Programa 2020 del Departamento de Filología de la Universidad de Cádiz, se enmarca en los proyectos «Comunicación especializada y terminografía: usos terminológicos relacionados con los contenidos y perspectivas actuales de la semántica léxica» (Ref. FFI2014-54609-P) del Programa Estatal de Fomento de la Investigación Científica y Técnica de Excelencia. Subprograma Estatal de Generación del Conocimiento (convocatoria 2014 del Ministerio de Economía y Competitividad) y «Bases metodológicas y recursos digitales para la creación de un léxico relacional de usos terminológicos de la semántica léxica (Ter-LexNet)», solicitado en la Convocatoria 2020 de Proyectos de I+D+i del Ministerio de Ciencia e Innovación. Igualmente se inscribe en los proyectos «Lingüística y nuevas tecnologías de la información: la creación de un repositorio electrónico de documentación lingüística» (Ref. FEDER-UCA18-107788), perteneciente a los Proyectos de I+D+i del Programa Operativo FEDER Andalucía 2014-2020, y «Lingüística y Humanidades Digitales: base de datos relacional de documentación lingüística» (Ref. PY18-FR-2511) de la Convocatoria 2018 de Ayudas a proyectos I+D+i (Modalidad «Frontera Consolidado») en el ámbito del Plan Andaluz de Investigación, Desarrollo e Innovación (Junta de Andalucía, PAIDI 2020).

Cualquier forma de reproducción, distribución, comunicación pública o transformación de esta obra sólo puede ser realizada, salvo excepción prevista en la ley, con la autorización de Editorial Universidad de Sevilla. La infracción de los derechos mencionados puede ser constitutiva de delito contra la propiedad intelectual (art. 270 y siguientes del Código Penal).

© EDITORIAL UNIVERSIDAD DE SEVILLA 2021
Porvenir, 27 - 41013 Sevilla.
Tífs.: 954 487 447; 954 487 451; Fax: 954 487 443
Correo electrónico: secpub4@us.es
Web: <<https://www.publius.us.es>>

© Mario Crespo Miguel, 2021

Impreso en papel ecológico
Impreso en España – Printed in Spain

ISBN: 978-84-472-3027-3
Depósito Legal: SE 1510-2021
Maquetación: sputnix diseño editorial – www.sputnix.es
Impresión: Podiprint

Índice

| | |
|--|----|
| CHAPTER 1. INTRODUCTION | 13 |
| 1.1. Motivation | 13 |
| 1.2. Goals | 14 |
| 1.3. Contributions..... | 14 |
| 1.4. Outline of this work..... | 15 |
| CHAPTER 2. LEXICAL RESOURCES IN NATURAL LANGUAGE PROCESSING..... | 17 |
| 2.1. Introduction | 17 |
| 2.2. Semantic resources, thesauri and ontologies | 19 |
| 2.3. Lexical resources in natural language processing..... | 22 |
| 2.4. Lexical resources for the biomedical domain | 37 |
| 2.5. Cross-linguistic induction of lexical knowledge..... | 39 |
| 2.6. General approaches to the induction of linguistic knowledge..... | 43 |
| 2.7. Summary..... | 47 |
| CHAPTER 3. A CORPUS-BASED APPROACH..... | 49 |
| 3.1. Overview | 49 |
| 3.2. A corpus-based approach | 52 |
| 3.3. Statistical hypothesis testing for frame selection: procedure | 56 |
| 3.4. The process in a nutshell | 59 |
| 3.5. Summary..... | 60 |
| CHAPTER 4. EXPERIMENTS REGARDING FRAME SELECTION FOR THE MEDICAL DOMAIN | 61 |
| 4.1. Overview | 61 |
| 4.2. Experiment 1. Frame selection based on hypothesis testing..... | 65 |
| 4.3. Experiment 2. Frame selection based on different levels of statistical significance..... | 70 |
| 4.4. Experiment 3. Medical frame selection by applying scientific texts as a reference corpus | 72 |
| 4.5. Experiment 4. Frame selection by restricting words to compute the hypothesis testing | 74 |
| 4.6. Experiment 5. Frame selection by extending the system with wordnet | 76 |
| 4.7. Discussion..... | 84 |
| 4.8. Summary..... | 85 |

| | |
|--|-----|
| CHAPTER 5. EXPERIMENTS IN DISAMBIGUATING THE FRAME SELECTION | |
| FOR THE MEDICAL DOMAIN | 87 |
| 5.1. Overview. Matching triggers in framenet to synsets in wordnet | 87 |
| 5.2. Experimentation in framenet to wordnet association..... | 94 |
| 5.3. Experimental translation | 109 |
| 5.4. Summary..... | 115 |
| CHAPTER 6 FINAL RESULTS AND CONCLUSIONS..... | 117 |
| 6.1. Translation of our medical selection of frames..... | 117 |
| 6.2. Conclusions..... | 128 |
| APPENDIX A. MANUAL SELECTION OF SYNSETS FOR OUR MEDICAL BENCHMARK..... | 135 |
| APPENDIX B. EXTENSION OF FRAME NET TRIGGERS WITH NEW TERMS..... | 143 |
| REFERENCES..... | 149 |

Tables

| | |
|--|----|
| Table 1. Thematic roles in 3LB-LEX project..... | 30 |
| Table 2. FrameNet status..... | 36 |
| Table 3. Participants or thematic roles of the Bringing and Transport_ intracellular frames..... | 39 |
| Table 4. Representation of the EuroWordNet concept '01238865v' in Spanish and English..... | 41 |
| Table 5. Frame Coincidence with English..... | 43 |
| Table 6. Relative frequencies of units in medical and general corpus..... | 53 |
| Table 7. Number of frames and triggers in FrameNet..... | 67 |
| Table 8. Significant frames after applying the Student t-test to the group of triggers of each frame..... | 68 |
| Table 9. Significant frames at a 90% significance level..... | 70 |
| Table 10. Significant frames at a 99% significance level..... | 71 |
| Table 11. Significant frames after applying t-test over the group of triggers of each frame..... | 73 |
| Table 12. Significant frames after applying the t-test over the group of triggers of each frame..... | 73 |
| Table 13. Significant frames after applying the t-test to the group of triggers of each frame..... | 74 |
| Table 14. Significant frames after removing outliers and extreme outliers..... | 75 |
| Table 15. Significant frames after removing outliers and extreme outliers..... | 76 |
| Table 16. Frames augmented with the WordNet relations of those non-ambiguous words with t-Test..... | 79 |
| Table 17. Frames augmented with the WordNet relations of those non-ambiguous words with Wilcoxon test..... | 80 |
| Table 18. Contingency table..... | 81 |
| Table 19. Frames augmented with the WordNet relations..... | 82 |
| Table 20. Frames augmented with the WordNet relations..... | 83 |
| Table 21. Results from a big bag of filtered words from semantic relations..... | 84 |
| Table 22. Spanish units attached to different synsets..... | 89 |
| Table 23. Synsets associated with FrameNet triggers..... | 89 |
| Table 24. Synsets associated with FrameNet triggers..... | 91 |
| Table 25. Example of non-ambiguous triggers in our selection..... | 91 |
| Table 26. Synsets associated with FrameNet triggers..... | 91 |
| Table 27. Synsets associated with mark.v according to WordNet..... | 93 |

| | |
|--|-----|
| Table 28. Number of triggers in manual disambiguation benchmark according to associated synsets..... | 93 |
| Table 29. Results with no experimentation..... | 95 |
| Table 30. Coincidence in synsets between two triggers of a single frame. | 96 |
| Table 31. Results for experiment A..... | 97 |
| Table 32. Results for experiment B. | 101 |
| Table 33. Results for experiment C..... | 103 |
| Table 34. Results for experiment D..... | 104 |
| Table 35. Derivate relations among synsets in WordNet..... | 105 |
| Table 36. Results for experiment E..... | 106 |
| Table 37. Triggers and frequency distribution of semantically related words..... | 108 |
| Table 38. Results for experiment F..... | 109 |
| Table 39. Number of triggers per Frame and number of correct items in Spanish..... | 126 |
| Table 40. Triggers in the original FrameNet and after attaching new terms from WordNet. | 128 |

Figures

| | |
|---|----|
| Figure 1. Lexical representation of synsets in WordNet..... | 22 |
| Figure 2. Semantic relations among synsets in WordNet..... | 24 |
| Figure 3. View of WordNet..... | 24 |
| Figure 4. Synset 03532413n in EuroWordNet..... | 25 |
| Figure 5. Example of the unit PLAY in VerbNet..... | 26 |
| Figure 6. Verbal class in VerbNet..... | 28 |
| Figure 7. PropBank definition of unit 'accept.01'..... | 29 |
| Figure 8. AnCora Annotation..... | 31 |
| Figure 9. Adesse analysis for the unit 'acceptar'..... | 32 |
| Figure 10. Lexical unit 'accept' in two different frames..... | 33 |
| Figure 11. Domain of Communication and Cognition..... | 34 |
| Figure 12. View of a frame in FrameNet..... | 35 |
| Figure 13. Verb information in BioProp..... | 38 |
| Figure 14. Traditional representation of a same concept in Spanish, French and German..... | 40 |
| Figure 15. Outlook of FrameNet..... | 42 |
| Figure 16. Representation of sentence alignment and transfer of information between languages..... | 44 |
| Figure 17. Outlook of the triggers and frames sorted by Medical condition domain in FrameNet..... | 50 |
| Figure 18. Illustration of a medical situation..... | 50 |
| Figure 19. Alignment of FrameNet triggers with WordNet..... | 51 |
| Figure 20. Comparison of the triggers of frame Medical conditions in two different corpora..... | 53 |
| Figure 21. Representation of the meanings of the substantive 'Cold' according to EuroWordNet..... | 54 |
| Figure 22. Representation of the same concept in EuroWordNet for English and Spanish..... | 54 |
| Figure 23. Frequency distributions of synset units associated with 'cold. n'..... | 55 |
| Figure 24. COCA corpus genre comparison..... | 56 |
| Figure 25. View of different frame triggers or lexical units in two different frames..... | 57 |
| Figure 26. Selection of the corpus..... | 59 |
| Figure 27. Process of selecting candidates for the medical selection..... | 59 |
| Figure 28. Trigger translation by using WordNet..... | 60 |
| Figure 29. F-score formula..... | 64 |

| | |
|---|-----|
| Figure 30. Positive-negative matrix. Image taken from Powers (2011)..... | 64 |
| Figure 31. Student test for Matched pairs in Triola (2007)..... | 66 |
| Figure 32. Test Statistic for Wilcoxon in Triola (2007)..... | 67 |
| Figure 33. Way of adding new terms to the frame..... | 77 |
| Figure 34. Way of adding WordNet lexical relations to the group of triggers..... | 78 |
| Figure 35. Log-likelihood formula..... | 81 |
| Figure 36. Log-likelihood formula..... | 82 |
| Figure 37. Trigger translation by using EuroWordNet..... | 88 |
| Figure 38. Representation of senses of 'cold' according to EuroWordNet..... | 88 |
| Figure 39. Process of disambiguating ambiguous triggers..... | 90 |
| Figure 40. Translation and extension of the frame..... | 90 |
| Figure 41. Unit-synset associations. | 92 |
| Figure 42. F-score formula..... | 94 |
| Figure 43. Ideal process of synset pruning..... | 95 |
| Figure 44. Semantic coincidences between senses of unit 'cure.v' and rest of senses of the predicates in the frame. | 98 |
| Figure 45. Illustration of how to include lexical relations. | 99 |
| Figure 46. General procedure for Experiment B. | 99 |
| Figure 47. Illustration of how frequencies are taken..... | 100 |
| Figure 48. Wilcoxon test on senses..... | 100 |
| Figure 49. Senses of 'cancer' and number of semantic matches..... | 102 |
| Figure 50. General procedure for Experiment C..... | 102 |
| Figure 51. Senses of 'stress.n' and number of semantic matches..... | 104 |
| Figure 52. General procedure of Experiment E..... | 106 |
| Figure 53. General Procedure for Experiment F. | 108 |
| Figure 54. Translation method using EuroWordNet..... | 110 |
| Figure 55. Way of transferring new terms by using EuroWordNet..... | 127 |

CHAPTER 1

INTRODUCTION

In this chapter, we will briefly give an overview of this book. Our task, the Automatic Corpus-based Induction of a Spanish FrameNet Medical Glossary, is motivated by the demand for more linguistic resources for the study of languages and the improvement of those already existing. The main contributions include 1) the development of a reliable way of selecting the FrameNet frames that conceptually reflect the domain of medicine, 2) the accurate translation of such frame selections into Spanish, 3) the improvement of the coverage of these frames with new triggers and semantic relations provided by EuroWordNet.

1.1. Motivation

The purpose of language is communication. Human beings use languages as a means of interacting with one another and to talk about the world. Languages are therefore meaningful and expressive instruments. Linguistic analysis and study must always bear in mind that languages would not exist beyond the goal of communication. However, languages differ from one another. They can be considered as different versions of the same “instrument of communication” (Dummet, 1991). The organization and structure is completely different from one language to another, but they all have the same goal.

Broadly speaking, the act of communicating is divided into a structural part (words, syntactic structures, pronunciation, etc.) and a semantic part (the meaning conveyed). Likewise, linguistic study consists of different areas: whereas phonological, morphological and syntactic analyses focus on formal parts of languages, semantics deals with the analysis of the meaning expressed by that formal part. A complete analysis of a language must include all these areas and the idea that everything in languages exists for the sake of communication. In natural language processing (henceforth, NLP), the number of projects dealing with semantics is smaller than projects related to other fields like morphology or syntax. The reason for such a shortage in computational linguistics can be attributable to the fact that meaning cannot be observed directly.

Semantics, in relation to morphology or syntax, is a more difficult linguistic level to formalize and, hence, be automatically processed.

In this work, we will review some of the most important projects devoted to describing the semantic properties of languages and we will determine if the information they provide can be projected and translated cross-linguistically. These projects include FrameNet (Baker *et al.*, 1998), PropBank (Palmer *et al.*, 2005), WordNet (Miller *et al.*, 1993) or VerbNet (Kipper *et al.*, 2000). Finally, we will transfer part of FrameNet into Spanish. In particular, we will focus on the FrameNet frames most closely related to the domain of medicine.

1.2. Goals

Most of the semantically-oriented projects have been created for English, mainly because most modern approaches to computational lexical semantics emerged in the United States. This situation is changing over time and some of these projects have been subsequently extended to other languages; however, in all cases, much time and effort need to be invested in creating such resources. Because of this, one of the main purposes of this work is to investigate the possibility of extending these resources to other languages like Spanish. As we will see, the special structure of FrameNet offers an opportunity to create similar resources for other languages. FrameNet aims to explain how languages account for daily situations linguistically.

Focusing on the frames that best represent the domain of medicine, we present a statistical method which, assisted by the word associations proposed by WordNet, can create a medical FrameNet selection and translation for Spanish and can also improve the trigger coverage of the English FrameNet. Results will be checked manually to evaluate the reliability of the system.

1.3. Contributions

We have developed a method of matching frame predicates with WordNet synsets by using the contextual information provided by a representative corpus (*COCA corpus*). This approach has been used to disambiguate the FrameNet triggers according to WordNet. Once the matching has been done, all the information that WordNet provides can be used both to translate the unit into other languages and, in this case, to extend the coverage of FrameNet with new units.

More than 90% of the triggers in our frame selection match one or two synsets in WordNet and 95.6% of the translated words were correct in Spanish. This approach

provides us with a reliable way to transfer FrameNet triggers to other languages. In addition to this, our medical frame selection was widened with new units by 204%. This approach could be also used to improve the range of the current English FrameNet.

1.4. Outline of this work

This book is structured as follows.

In Chapter 2, we will present some of the most important resources in Lexical Semantics. We will discuss in detail their linguistic representations applied and features selected. We will pay special attention to those resources developed for Spanish. After that, we will discuss some of the techniques for the cross-linguistic transfer of information.

In Chapter 3, we will discuss our corpus-based approach. Firstly, we will describe how different text domains select different frames. Accordingly, texts about 'communication' will select frames such as *conversation*, *questioning*, *statement*, etc. and frames such as *judgment* or *categorization* will show up in texts about 'cognition'. Our approach aims to choose the range of frames that account for the medical domain. Roughly speaking, our approach uses the information regarding word distribution from a medical corpus to select a representative set of frames and translate them into other languages. Translation is supported by EuroWordNet, the extension of the Princeton WordNet for some European languages. Finally, this chapter briefly introduces the concept of a Statistical Hypothesis Testing, a key element in this work.

In Chapter 4, we will show the different experiments conducted for medically-oriented frame selection. Frame gathering is based on the results after carrying out Statistical Hypothesis Testing on the FrameNet frame triggers. Output was compared to a manual benchmark to check the suitability of the selection.

In Chapter 5, we will describe several experiments that were carried out to match the triggers of the frames selected with synsets of EuroWordNet. Both semantic properties of words and frequency distribution are applied to attach a particular trigger to the WordNet synsets. Various experiments were conducted and, at the end, the results are evaluated.

In Chapter 6, we summarize and discuss the basic results of the present work and outline some important future directions.