The Structure of the Lexicon in the Task of the Automatic Acquisition of Lexical Information

Lauren Michele Romeo

TESI DOCTORAL UPF / ANY 2015

DIRECTOR DE LA TESI
Prof. Núria Bel
Institut Universitari de Lingüística Aplicada

upf.
Universitat
Pompeu Fabra
Barcelona

This thesis is dedicated to the ever-patient and always-loving Alberto, as it marks our end to one adventure and the beginning of many more...

Acknowledgments

The accomplishment of this thesis is the result of many fruitful collaborations, an uncountable number of insightful conversations and an unwavering, unbreakable support system (both professionally and personally), from which I have many people that I need to thank (although in many cases, I must note, words will never truly be sufficient).

First and foremost, I would like to thank my supervisor and the director of this research, Núria Bel, for giving me the opportunity to develop the work presented in this thesis. Her constant motivation, guidance, encouragement and, most importantly, honest and constructive feedback were fundamental to both its success and my personal development as a researcher.

A special thank you is reserved for a colleague that has also become a friend, Sara Mendes, who constantly and generously provided me with the support, discussions and critical observations that have integrally contributed to this thesis.

I am thankful to Professor Alessandro Lenci for welcoming me into the CoLing Lab at the Università di Pisa and for an enriching, enlightening and enjoyable predoctoral stay. So many of the results, musings, discussions and conclusions formed in this thesis are a direct result of my time spent there.

I would also like to thank all of the members of the IULA and the TRL group for the support that they have given me throughout this adventure. I am grateful for all of the assistance during these past years.

A special mention goes to Silvia Vázquez for being such a good colleague and, more importantly, friend throughout my entire time spent at the IULA. Our enriching travels to conferences both near and far have produced some of my fondest memories of this experience.

I would also like to thank my Ph.D companions (both past and present) who have been accompanying me on the road less traveled these past years: Marco del Tredici, Marina Fomicheva, Jingyi Han, Silvia Necşulescu, and Gianni Zucca. No matter what path we find ourselves in the future, "...we'll always have the Wednesday afternoon seminars".

My appreciation also extends to all of the other people that have helped along the way, especially Francesca Frontini, Fahad Khan, Gianluca Lebani, Héctor

Martínez Alonso and Lucia Passaro. In different moments, at different times and in different locations, you have all had such a great impact on me and this work in this thesis. Thank you for all of the long chats, delicious dinners, and great memories at conferences (especially LREC-2014!), over Skype and at the lab, over a coffee in the back patio at the campus on via Santa Maria and at La Scaletta over some great 'spaghetti al granchio' (which I promise I will learn how to eat next time).

A very special thank you goes to my Barcelona friends, who have been by my side from the early days of CSIM to the present. Thank you for your continuous support and the many shared memories that have made this experience so wonderful.

I would also like to extend a very special thank you to my best friends from the other side of the pond. Although we might be far in distance, thank you for just being there for me (group chat forever!); it means more than anything and has helped me more than you will ever know throughout this journey.

I am deeply and eternally thankful to my parents, Tom and Lori, and my siblings, Marisa, Danielle and Thomas, whose unwavering support and constant curiosity to "What I am doing over there" have been nothing short of pure motivation that has pushed me to reach beyond the stars and to strive to be the absolute best that I can be.

My final thank you goes to Alberto, for being no where else but directly by my side with nothing but love and support from the beginning and throughout this entire journey. These past few years have been a growing experience in more ways than one and he has always seen the best in me and has truly given me the wings to soar proudly through the finish line.

To everyone else that has contributed in one way or another to this thesis, I appreciate it all and, trust me, I am forever indebted.

Finally, all of the above would not have been possible without the financing and the support of the FI-DGR2012 scholarship of the Secretaría de Universidades e Investigación (SUR) (Secretary of Universities and Investigation) of the Departamento de Economía y Conocimiento (DEC) (Department of Economy and Knowledge) of the Generalitat de Catalunya, of the Fons Social Europeu (European Social Fund) and the IULA-UPF scholarship of the Institut Universitari de Lingüística Aplicada (University Institute of Applied Linguistics).

Abstract

Lexical semantic class information for nouns is critical for a broad variety of Natural Language Processing (NLP) tasks including, but not limited to, machine translation, discrimination of referents in tasks such as event detection and tracking, question answering, named entity recognition and classification, automatic construction and extension of ontologies, textual inference, etc.

One approach to solve the costly and time-consuming manual construction and maintenance of large-coverage lexica to feed NLP systems is the Automatic Acquisition of Lexical Information, which involves the induction of a semantic class related to a particular word from distributional data gathered within a corpus. This is precisely why current research on methods for the automatic production of high-quality information-rich class-annotated lexica, such as the work presented here, is expected to have a high impact on the performance of most NLP applications.

In this thesis, we address the automatic acquisition of lexical information as a classification problem. For this reason, we adopt machine learning methods to generate a model representing vectorial distributional data which, grounded on known examples, allows for the predictions of other unknown words.

The main research questions we investigate in this thesis are: (i) whether corpus data provides sufficient distributional information to build efficient word representations that result in accurate and robust classification decisions and (ii) whether automatic acquisition can handle also polysemous nouns.

To tackle these problems, we conducted a number of empirical validations on English nouns. Our results confirmed that the distributional information obtained from corpus data is indeed sufficient to automatically acquire lexical semantic classes, demonstrated by an average overall $F1$-Score of almost 0.80 using diverse count-context models and on different sized corpus data.

Nonetheless, both the State of the Art and the experiments we conducted highlighted a number of challenges of this type of model such as reducing vector sparsity and accounting for nominal polysemy in distributional word representations. In this context, Word Embeddings (WE) models maintain the "semantics" underlying the occurrences of a noun in corpus data by mapping it to a feature vector. With this choice, we were able to overcome the sparse data problem, demonstrated by an average overall $F1$-Score of 0.91 for single-sense lexical semantic noun classes, through a combination of reduced dimensionality and "real" numbers.

In addition, the WE representations obtained a higher performance in handling the asymmetrical occurrences of each sense of regular polysemous complex-type nouns in corpus data. As a result, we were able to directly classify such nouns into their own lexical-semantic class with an average overall $F1$-Score of 0.85.

The main contribution of this dissertation consists of an empirical validation of different distributional representations used for nominal lexical semantic classification along with a subsequent expansion of previous work, which results in novel lexical resources and data sets that have been made freely available for download and use.

Resumen

La información de clase semántica de los nombres es fundamental para una amplia variedad de tareas del procesamiento del lenguaje natural (PLN), como la traducción automática, la discriminación de referentes en tareas como la detección y el seguimiento de eventos, la búsqueda de respuestas, el reconocimiento y la clasificación de nombres de entidades, la construcción y ampliación automática de ontologías, la inferencia textual, etc.

Una aproximación para resolver la construcción y el mantenimiento de los léxicos de gran cobertura que alimentan los sistemas de PNL, una tarea muy costosa y lenta, es la adquisición automática de información léxica, que consiste en la inducción de una clase semántica relacionada con una palabra en concreto a partir de datos de su distribución obtenidos de un corpus. Precisamente, por esta razón, se espera que la investigación actual sobre los métodos para la producción automática de léxicos de alta calidad, con gran cantidad de información y con anotación de clase como el trabajo que aquí presentamos, tenga un gran impacto en el rendimiento de la mayoría de las aplicaciones de PNL.

En esta tesis, tratamos la adquisición automática de información léxica como un problema de clasificación. Con este propósito, adoptamos métodos de aprendizaje automático para generar un modelo que represente los datos de distribución vectorial que, basados en ejemplos conocidos, permitan hacer predicciones de otras palabras desconocidas.

Las principales preguntas de investigación que planteamos en esta tesis son: (i) si los datos de corpus proporcionan suficiente información para construir representaciones de palabras de forma eficiente y que resulten en decisiones de clasificación precisas y sólidas, y (ii) si la adquisición automática puede gestionar, también, los nombres polisémicos.

Para hacer frente a estos problemas, realizamos una serie de validaciones empíricas sobre nombres en inglés. Nuestros resultados confirman que la información obtenida a partir de la distribución de los datos de corpus es suficiente para adquirir automáticamente clases semánticas, como lo demuestra un valor-F global promedio de 0.80 aproximadamente utilizando varios modelos de recuento de contextos y en datos de corpus de distintos tamaños.

No obstante, tanto el estado de la cuestión como los experimentos que realizamos destacaron una serie de retos para este tipo de modelos, que son reducir la escasez

de datos del vector y dar cuenta de la polisemia nominal en las representaciones distribucionales de las palabras. En este contexto, los modelos de *Word embeddings* (WE) mantienen la “semántica” subyacente en las ocurrencias de un nombre en los datos de corpus asignándole un vector. Con esta elección, hemos sido capaces de superar el problema de la escasez de datos, como lo demuestra un valor-F general promedio de 0.91 para las clases semánticas de nombres de sentido único, a través de una combinación de la reducción de la dimensionalidad y de números reales.

Además, las representaciones de WE obtuvieron un rendimiento superior en la gestión de las ocurrencias asimétricas de cada sentido de los nombres de tipo complejo polisémicos regulares en datos de corpus. Como resultado, hemos podido clasificar directamente esos nombres en su propia clase semántica con un valor-F global promedio de 0.85.

La principal aportación de esta tesis consiste en una validación empírica de diferentes representaciones de distribución utilizadas para la clasificación semántica de nombres junto con una posterior expansión del trabajo anterior, lo que se traduce en recursos léxicos y conjuntos de datos innovadores que están disponibles de forma gratuita para su descarga y uso.

Resum

La informació de classe semàntica dels noms és fonamental per a un gran nombre de tasques de processament del llenguatge natural (PLN), com la traducció automàtica, la discriminació dels referents en tasques com la detecció i el seguiment d'esdeveniments, la cerca de respostes, el reconeixement i la classificació de noms d'entitats, la construcció i l'ampliació automàtica d'ontologies, la inferència textual, etc.

Una aproximació per resoldre la construcció i el manteniment manual de lèxics de gran cobertura que alimenten els sistemes PLN, una tasca molt costosa i lenta, és l'adquisició automàtica d'informació lèxica, que implica la inducció d'una classe semàntica relacionada amb una paraula determinada a partir de dades de distribució obtingudes d'un corpus. És precisament per això que s'espera que la investigació actual sobre mètodes per a la producció automàtica de lèxics d'alta qualitat, amb molta informació i amb anotació de classe, com el treball que presentem aquí, tingui un gran impacte en el rendiment de la majoria de les aplicacions de PLN.

En aquesta tesi, tractem l'adquisició automàtica d'informació lèxica com un problema de classificació. Per aquesta raó, adoptem mètodes d'aprenentatge automàtic per generar un model que representi les dades de distribució vectorial que, basades en exemples coneguts, permetin predir més paraules desconegudes.

Les principals preguntes de recerca que plantegem en aquesta tesi són: (i) si les dades de corpus proporcionen suficient informació sobre la distribució per construir representacions de paraules de forma eficient i que tinguin com a resultat decisions de classificació precises i sòlides, i (ii) si l'adquisició automàtica pot gestionar, també, els noms polisèmics.

Per fer front a aquests problemes, hem dut a terme una sèrie de validacions empíriques en noms en anglès. Els nostres resultats confirmen que la informació de distribució obtinguda a partir de dades de corpus és suficient per adquirir automàticament classes semàntiques, demostrat per un valor-F global d'aproximadament 0.80 utilitzant diversos models de recompte de context i en dades de corpus de mides diferents.

No obstant això, tant l'estat de la qüestió com els experiments que vam realitzar destacaven una sèrie de reptes d'aquest tipus de models, com reduir l'escassetat de vectors i donar compte de la polisèmia nominal en les representacions de paraules distribucionals. En aquest context, els models de *Word embeddings* (WE) mantenen la “semàntica” subjacent a les ocurrències d’un nom en les dades de corpus assignant-lo a un vector de característiques. Amb aquesta elecció, hem pogut superar el problema de l’escassetat de dades, com ho demostra un valor-F general de mitjana de 0.91 per a les classes semàntiques de noms de sentit únic, a través d’una combinació de la reducció de la dimensionalitat i de nombres reals.

A més, les representacions de WE van obtenir un rendiment superior en la gestió de les ocurrències asimètriques de cada sentit dels noms de tipus complex polisèmics regulars en dades de corpus. Com a resultat, hem pogut classificar directament aquests noms en la seva pròpia classe semàntica amb un valor-F global de mitjana de 0.85.

La principal aportació d’aquesta tesi consisteix en una validació empírica de diferents representacions de distribució utilitzades per a la classificació semàntica de noms juntament amb una expansió del treball anterior, el que es tradueix en nous recursos lèxics i conjunts de dades que es posan a lliure disposició perquè es puguin descarregar i utilitzar.

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

The automatic acquisition of lexical information involves the induction of a semantic class related to a particular word from distributional data gathered within a corpus. We approach this task by classifying words into previously known nominal lexical-semantic classes. Currently, the information obtained by automatic acquisition is critical for a variety of Natural Language Processing (NLP) tasks, including, but certainly not limited to, machine translation, the discrimination of referents in tasks such as event detection and tracking [Fillmore et al., 2006], question answering [Lee et al., 2001], entity typing in named entity recognition [Ciaramita and Altun, 2005, Fu, 2009], automatic building and extending of ontologies [Buitelaar et al., 2005], textual inference [de Marneffe et al., 2009] and much more. Furthermore, these automatically acquired lexical-semantic classes have been proven to be useful for grammar induction [Agirre et al., 2011], where problems come from the need to generalize over a high-dimensional space.

The significant cost of manually conducting this task hinders, for instance, the production of rich lexica, as well as its creation for different languages. In addition, the domain tuning of lexica is expensive, and the use of an inadequate lexicon is one of the causes of poor performance for many applications. Along this line, good lexical coverage is absolutely crucial to achieve proper performance of any processing component for NLP applications that rely on lexical information. Automatic lexical semantic class acquisition offers a solution for the construction—and, importantly, the maintenance—of large-coverage lexica for feeding these processing components. Thus, current research on methods for the automatic production of high-quality, information-rich, class-annotated lexica, such as what will be presented in the subsequent chapters of this thesis, is expected to have a high impact on the performance of most NLP applications. Moreover, and critically, it will foster significant improvements in their coverage over different domains, as well as languages.

Distributional methods, based on the Distributional Hypothesis [Harris, 1954], build word representations from corpus data that represent a word directly through the contexts in which it has been observed. This hypothesis conveys the oft-referred to statement “similar words tend to occur in similar contexts”. Distributional representations of words consist of a vector of n-components, where each component encodes the frequency of the word occurring with or in a particular context. These vectors model the contexts in which a given word has been observed. However, the definition of what consists of a context is highly dependent on the model considered. For instance, features can consist of co-occurring tokens, phrases, a combination of individual lexical items or the order in which those lexical items appear to name a few. All of the feature information is then represented in feature vectors that are used to classify words into a number of desired classes. Supervised learning methods are then used so that automatic lexical acquisition can be approached by assigning a word to certain classes according to information gathered from its occurrences in texts as represented in the vector.

Yet, there is no consensus on the features that are relevant for the task of classifying nouns; and different features can yield both advantages and disadvantages when constructing a word representation that is both informative and useful to a machine-learning algorithm. For this reason, we must understand what is the required information to be provided for the classifier to learn and to make predictions that successfully construct viable and informative distributional word representations. Herein lies our main object of study in this thesis: the construction of distributional word representations that are both useful and informative to machine learning methods.

The accomplishment of building both viable and informative distributional word representations for machine learning algorithms requires an in-depth study of two main obstacles that have been encountered in the construction of these word representations: mainly, the lack of useful information in feature vectors, i.e., sparsity, and the consideration of the multiple classes to which a word can belong, i.e., polysemy. Along this line, the main goals pillaring this thesis are two-fold:

- to identify and to empirically justify the main challenges confronted in the task of the automatic acquisition of nominal lexical-semantic information;
- to build a lexical-semantic classifier using information obtained from distributional word representations that account for these challenges through

selecting the most informative and distinctive features to build it.

Specifically, we account for the challenges that arise from both the potential sparsity of nouns in corpus data, which negatively affects word representations, and the polysemy of nouns, which can cause a bias for machine learning algorithms through an asymmetry of occurrences as a member of a particular class in corpus data.

The main issue confronted is found in the representation of distributional information of words in corpus data; specifically, in what information is required to achieve the most informative distributional representation that we can build from corpus data. There have been many approaches to the building of distributional semantic representations based on, for instance, the bag-of-words-type model [Bullinaria and Levy, 2007, Bullinaria and Levy, 2012], which uses windows of words for features; linguistically-motivated models [Merlo and Stevenson, 2001, Joanis et al., 2008], which can use subcategorization frames or the exact position occupied by a target word in a given context for features or general purpose distributional semantic models [Baroni and Lenci, 2010, Turney and Pantel, 2010], which use both grammatical and lexical information in a space that encodes networks of semantic information from corpus data that can be adapted to all sorts of tasks. However, depending on the model selected, a sacrifice in either precision or recall must be not only accepted, but assumed.

From these considerations, along with the main goals outlined for this thesis, the general research question arose: what is the most relevant and useful distributional information to include in word representations; and will these representations result in cleaner vectors with more relevant class-indicative information, which in turn permit the construction of more accurate and broad-covering nominal lexical semantic classifiers?

To answer these questions, we first needed to identify strategies to correctly handle language data that has specific characteristics (such as ambiguity or low frequency of use), which are not always easily interpreted with machine learning algorithms. Furthermore, we placed ourselves within the framework of a semantic theory, as it allows for the acquisition of refined semantic features that do not always emerge from a purely corpus-based collocation analysis. We chose the Generative Lexicon theory (GL) [Pustejovsky, 1995] because it offers an alternative to the traditional, sense-enumeration-based understanding of word senses by postulating the concept of so-called complex-type nouns, i.e., nouns that are regular polysemous, meaning that they can be selected for two different classes in one single context and, therefore, present a large challenge for any distributional representation. Placing ourselves in a theoretical framework such as GL allows us to adequately frame the evaluation of automatically acquiring lexical semantic information of these types of polysemous nouns.

We narrowed the focus of this thesis to English nouns, which, besides being less studied in comparison to other parts of speech, such as verbs, are known to be deeply affected by the polysemy problem, which has not typically been dealt with distributionally in the State of the Art and as we will see in Chapter 5. Moreover, we specifically study the grammatical category of nouns because they can behave as arguments, which can be semantically selected for in context. This allows us to use the specific contexts that surround a noun to predict the other nouns that can be selected in that same context. In this way, we are able to automatically acquire the lexical semantic class information for an unknown target noun. We have chosen to conduct our study on a number of lexical semantic classes that are also available in WordNet [Miller et al., 1990], more specifically, COMMUNICATION_OBJECT, EVENT, HUMAN, ORGANIZATION and LOCATION. For our study on the automatic acquisition of nominal regular polysemous information (Chapter 5), we also use a data set built by [Boleda et al., 2012a] that was based on the regular polysemous alternations defined in the CoreLex data base [Buitelaar, 1998].

The first obstacle that we tackled in the work presented in this thesis is sparse data (Chapter 4). As previously described, distributional representations model the contexts in which a given word has been observed, yet, those contexts are highly dependent on the distributional model used. Thus, this sparse data issue is not unique to any distributional model in particular, as each model has its own challenges when it comes to populating vectors with “real” numbers. Sparse vectors are especially problematic to the classifiers because the actual information that is available in the word representation can be undermined by non-, or zero, values due to the fact that machine learning algorithms cannot efficiently differentiate between them. Furthermore, evidence occurring with low-frequency can be disregarded by automatic systems, as demonstrated in the classification experiments of [Bel et al., 2007, Bel, 2010], which results in the word representations providing insufficient class indicative information to a classifier. Thus, we empirically compared distributional models that exploit different types of feature information, which thereby resulted in different representations of nominal behavior in context. The resulting analysis provided insight regarding the effects (both advantages and disadvantages) when considering different types of feature information to build distributional representations for nominal lexical-semantic classification tasks.

In addition to studying the viability for automatic lexical-semantic acquisition between different distributional (count-context) models built from features containing information from different levels of generalization, we also studied the use of Word Embeddings (WE). These representations, rather than counting individual occurrences, map the occurrences of a noun in corpus data into a feature vector, maintaining its underlying "semantics". Our empirical analysis presented evidence that, on the one hand, confirms the ability of certain types of distributional information to provide accurate representations to a machine-learning algorithm, resulting in accurate and robust class membership decisions. On the other hand, it permits the further understanding of the relations between the features used to build distributional word representations, the origins of the obstacles they encounter and, more importantly, how these obstacles can be overcome. Consequently, we validated that the distributional models (both count-context and WE) provide empirical evidence to confirm whether the indicative distributional information available in corpus data is sufficient to be identified, captured and learned by machine learning algorithms, resulting in stronger lexical-semantic class membership predictions.

In Chapter 5, we focused on the challenges that regular polysemy for distributional word representations. Unlike single-sense "simple-type" nouns, complex-type nouns, or nouns that can occur as a member of more than one class in a single context, can also occur in contexts indicative of each of the individual semantic classes that form a regular polysemous alternation. However, these occurrences are not always equal, resulting in characteristic patterns of occurrence in corpus data that differ from simple-type monosemous nouns that cannot and do not instantiate a regular polysemous alternation. Thus, the problem lies mainly in the fact that all of the potential senses of a word that are learned from the contexts become conflated into one representation, which does not necessarily—and, in fact, rarely—equally represent each sense of that word. Three main experiments were conducted to assess and overcome these aforementioned obstacles. The first experiment consisted of a clustering system using representations automatically built from the FORMAL role of the Qualia Structure (QS), a postulation of GL, identified whether there is sufficient distributional evidence to automatically acquire information from more than one related sense of a noun.

The second and third experiments consisted of the design and implementation of a dedicated two-step approach that incorporates the special characteristics of complex-type regular polysemous nouns in an attempt to simultaneously gain more coverage while also increasing precision, which is usually low due to the unique contextual properties of these nouns. We conducted this experiment using both the aforementioned LING model and the more exhaustive and recently exploited WE models in an attempt to evaluate the impact of sense asymmetry on the two models. This experiment accounted for the differences between monosemous and polysemous classes, and highlighted the importance of considering the unique characteristics of these types of classes when building distributional word representations. The results of this experiment led us to propose that complex-type nouns that can instantiate a regular polysemous alternation should be treated as members of their own, individual lexical-semantic class for automatic classification tasks.

This thesis is organized as follows: in Chapter 2 we cover the related work on the use of distributional information to model and build word representations, strategies and methods for the automatic acquisition and classification of lexical-semantic information, as well as the use of semantic theory, more specifically GL, to guide relevant empirical evaluations. Chapter 3 describes each of the data sets used in the subsequent experiments conducted in this thesis and the steps followed to build them, when applicable. Chapter 4 describes the experiments conducted regarding monosemous, or single-sensed, lexical-semantic classes. Along this line, this Chapter also described the identification and implementation of our linguistic model and proposes methods to overcome sparsity in resulting vectorial representations, both through the increase of distributional information and the use of WE representations. Moreover, this Chapter also compares the empirical results of several State-of-the-Art methods to determine the most effective word representations to provide to machine learning algorithms. Chapter 5 describes and identifies the main challenges that machine-learning algorithms encounter with the distributional representations of regular polysemous nouns, mainly sparsity in the vector and asymmetry. This Chapter also describes in detail a two-step strategy proposed to overcome these challenges, as well as how WE representations can be used to acquire lexical-semantic class information of regular polysemous nouns. Finally, Chapter 6 summarizes the main conclusions drawn from the experiment chapters and lists the additional contributions of the dissertation.

Publications

Parts of this thesis (ideas, figures, results and discussions) have appeared previously in the following peer-reviewed publications and have served as the basis for others, which have been marked with (*):

- [Bel et al., 2012]: Bel, Núria, Romeo, Lauren; Padró, Muntsa (2012). "Automatic lexical semantic classification of nouns". In Calzolari, Nicoletta; Choukri, Khalid; Declerck, Thierry (et al.) (Eds.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). Paris: European Language Resources Association (ELRA). p. 1448-1455. ISBN 978-2-9517408-7-7

- [Romeo et al., 2012]: Romeo, Lauren; Mendes, Sara; Bel, Núria (2012). 'Using Qualia Information to Identify Lexical Semantic Classes in an Unsupervised Clustering Task'. In Kay, Martin; Boitet, Christian (eds.) Proceedings of COLING 2012: Posters: 24th International Conference on Computational Linguistics COLING 2012; 2012 December 8-15; Mumbai, India. Mumbai: The COLING 2012 Organizing Committee. p. 1029-1038

- * [Romeo et al., 2013a]: Romeo, Lauren; Martínez Alonso, Héctor; Núria (2013). "Class-based Word Sense Induction for dot-type nominals". In Saurí, Roser; Calzolari, Nicoletta; Huang, Chu-Ren; Lenci, Alessandro; Monachini, Monica; Pustejovsky, James (ed.) Proceedings of the 6th International Conference on Generative Approaches to the Lexicon: Generative Lexicon and Distributional Semantics GL2013: September 24-25 2013 Pisa, Italy. Pisa: Istituto di Linguistica Computazionale Antonio Zampolli. p. 76 -83. ISBN 978-1-937284-98-5

- [Romeo et al., 2013b]: Romeo, Lauren; Mendes, Sara; Bel, Núria (2013). "Towards the automatic classification of complex-type nominals". In Saurí, Roser; Calzolari, Nicoletta; Huang, Chu-Ren; Lenci, Alessandro; Monachini, Monica; Pustejovsky, James (ed.) Proceedings of the 6th International Conference on Generative Approaches to the Lexicon: Generative Lexicon and Distributional Semantics GL2013: September 24-25 2013 Pisa, Italy. Pisa: Istituto di Linguistica Computazionale Antonio Zampolli. p. 21-28. ISBN 978-1-937284-98-5

- [Romeo et al., 2014a]: Romeo, Lauren; Lebani, Gianluca; Bel, Núria; Lenci, Alessandro (2014). "Choosing which to Use? A Study of Distributional Models for Nominal Lexical Semantic Classification". In Calzolari, Nicoletta (Conference Chair), Choukri, Khalid; Declerck, Thierry (et al.) (Eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14): May 26-31, 2014 Reykjavik, Iceland. [s.l.]: ELRA. p. 4366-4373. ISBN 978-2-9517408-8-4

- [Romeo et al., 2014b]: Romeo, Lauren; Mendes, Sara; Bel, Núria (2014). "A Cascade Approach for Complex-type Classification". In Calzolari, Nicoletta (Conference Chair), Choukri, Khalid; Declerck, Thierry (et al.) (Eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14): May 26-31, 2014 Reykjavik, Iceland. [s.l.]: ELRA. p. 4451-4458. ISBN 978-2-9517408-8-4

- [Romeo et al., 2014c]: Romeo, Lauren; Mendes, Sara; Bel, Núria (2014). "Using unmarked contexts in nominal lexical semantic classification". In Tsujii, Junichi ; Hajic, Jan (eds.) Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers: August 23-29, 2014, Dublin Ireland. Dublin, Ireland: Dublin City University and Association for Computational Linguistics. p. 508-519. ISBN 978-1-941643-26-6

- * [Martínez Alonso and Romeo, 2014]: Martínez Alonso, Héctor; Romeo, Lauren (2014). "Crowdsourcing as a Preprocessing for Complex Semantic Annotation Tasks". In Calzolari, Nicoletta (Conference Chair), Choukri, Khalid; Declerck, Thierry (et al.) (Eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14): May 26-31, 2014 Reykjavik, Iceland. [s.l.]: ELRA. p. 229-234. ISBN 978-2-9517408-8-4

Chapter 2

STATE OF THE ART

Before evaluating the State of the Art through a literature review of the automatic acquisition of lexical-semantic information, we revised some of the main concepts upon which the State of Art is based to have a better overview of the implications and obstacles that encountered in our analysis.

The Distributional Hypothesis [Harris, 1954] forms the basis upon which this thesis is built. The Distributional Hypothesis conveys the oft referred to statement "similar words tend to occur in similar contexts" [Rubenstein and Goodenough, 1965, Schütze and Pedersen, 1995, Landauer and Dumais, 1997, Pantel, 2005]. [Sahlgren, 2008] further postulates that there is a correlation between distributional similarity and meaning similarity, which therefore allows us to utilize the former in order to estimate the latter. Along this line, distributional approaches have been adapted to use distributional properties of linguistic entities as the building blocks of semantics, especially for meaning acquisition tasks. Recent research has used the Distributional Hypothesis to build word representations in vectorial spaces to use for tasks in the field of Natural Language Processing (NLP), such as tasks that automatically acquire lexical semantic information [Brent, 1993, Merlo and Stevenson, 2001, Stevenson and Joanis, 2003, Baldwin and Bond, 2003, Baldwin, 2005, Joanis et al., 2008, Bullinaria and Levy, 2007, Bullinaria, 2008, Bullinaria and Levy, 2012, Turney and Pantel, 2010, Baroni and Lenci, 2010].

These distributional approaches typically treat a word as an n-dimensional vector that encodes the patterns of co-occurrence of that word with other expressions in a large corpus of language [Sahlgren, 2008, Turney and Pantel, 2010, Baroni and Zamparelli, 2010]. These representations can also include the lexical and syntactic constraints related to semantic categories that can be identified, captured, represented and learned for semantic prediction tasks. [Copestake and

Herbelot, 2012]. The distributional representations of words can be provided to machine learning algorithms that use the available information to make predictions regarding the semantics of an unknown word. For these reasons, distributional representations are especially useful because they provide a general purpose representation of natural language meaning [Erk, 2013] that can be learned and used to acquire further information.

2.1 Building distributional word representations

The Distributional Hypothesis postulates that words occurring in a similar context tend to have similar distributional representations. Distributional models directly approach the meaning of words from their occurrences in corpus data through the information that is considered to be an indicative property of that word, or of the lexical-semantic class of which that word is a member. However, distributional representations can vary greatly, depending on the specific aspects of meaning they are designed to model. Because of this, the selection of the most useful and/or indicative features is one of the most important tasks to build distributional models because it directly affects the how a word is represented and, consequently, the classification decision made using its representation.

The extraction of distributional properties from corpus data can result in different types of so-called *count-context* representations, including linguistically-motivated distributional representations, structured distributional representations, unstructured distributional representations [Baroni and Lenci, 2010] and, more recently, Word Embeddings WE representations [Mikolov et al., 2013, Levy and Goldberg, 2014b, Levy and Goldberg, 2014a].

On the one hand, count-context models are distributional representations that count individual occurrences in corpus data to keep track of those contexts where of a given word is found [Clark, 2014, Erk, 2012, Turney and Pantel, 2010, Baroni and Lenci, 2010]. Each context becomes a dimension in a feature vector and the frequency information of a given word occurring with that context is represented the value of that feature. with its co-occurrence information as a value. On the other hand, WE representations frame the vector estimation problem directly as a supervised task, where the weights in a word vector are set to maximize the probability of the contexts in which the word is observed in the corpus, rather than collecting context vectors [Bengio et al., 2003, Collobert and Weston, 2008, Turian et al., 2010, Collobert et al., 2011, Huang et al., 2012, Mikolov et al., 2013, Baroni et al., 2014]. In this Section, we explore both types of representations.

2.1.1 Count-Context models

As described above, count-context models are distributional semantic spaces built by counting each of the individual contexts implied by the parameters that define a given feature set. There are several types of count-context models, which are distinguished by different types of features. Due to the nature of count-context models, the features that characterize these representations have different levels of granularity, such as words, individual word occurrences or phrases [Erk, 2012]. For this reason, the selection of features is highly relevant, as they have a direct impact on the resulting word representations, and therefore the classification predictions that they obtain. In the following Sections, we describe in detail each of the different types of count-context models that we will study in this thesis.

Unstructured distributional models

Unstructured distributional models consist of simple word co-occurrence statistics and they consider pre-defined windows around a target word or pre-selected contexts in which a target word occurs as descriptive feature information. One of the most commonly used unstructured models is the so-called "bag-of-word" (BOW) approach, which represents a text as a "bag" of its words, disregarding grammar and word order, yet maintaining the multiplicity of each individual instance. As it is one of the simpler models that studied, it incorporates no further linguistic information into its representations, which as we will see, can exclude valuable information.

[Lund and Burgess, 1996] and [Landauer and Dumais, 1997] built unstructured distributional models by inducing knowledge directly from local co-occurrence data in large corpora. Their models captured the substitutability and the semantic similarity of word relations. However, these model were highly affected by the parameters selected. For instance, [Landauer and Dumais, 1997] reported that their model performed poorly when it relied only on local co-occurrence count (too many dimensions) and when it it tried to represent all its word knowledge in less than 100 dimensions; yet, a strong performance was achieved with around 300 dimensions. Thus, the definition the number of dimensions for the word representations is not a trivial task, especially when constructing unstructured models. [Kiela and Clark, 2014] further confirmed this idea in their systematic study of parameters used in the construction of semantic vector space models.

They also concluded that larger vectors did not always lead to better performance, as their results indicated that performance tends to stabilize as vector size increased. Furthermore, their model became compromised with smaller corpora, demonstrating sensitivity to a reduction in corpus data. This result indicated that corpus size was also a factor in determining the predictive power of the model, as unstructured models have been known to be more predictive when built with larger corpora.

[Bullinaria and Levy, 2007]; [Bullinaria and Levy, 2012] and [Bullinaria, 2008] built unstructured models to induce aspects of word meaning using simple word co-occurrence counts from corpus data. Their unstructured distributional models studied semantic word categorization as a function of window type and size, semantic vector distribution, as well as corpus size. [Bullinaria and Levy, 2012] reported the best performance using their models for semantic categorization at approximately 80%.

In our classification experiments, presented in Section 4.3 of this thesis, we, too, observed such an effect with the performance of our LINE model on smaller corpus data. However, we will continue to see throughout the work presented that sparsity is not an obstacle exclusive to unstructured models, as more structured models can also be subject to challenges concerning sparse data [Turney and Pantel, 2010].

Linguistically-motivated distributional models

Linguistically-motivated models use features that use linguistic knowledge to identify lexico-syntactic patterns identified to capture different contexts where a number of words belonging to a class tend to occur, going beyond the simple lexical co-occurrence information used by the “bag-of-words”-type approaches. These models assume that linguistic information can be provided by the distribution of occurrence to motivate lexical classes [Grimshaw, 1990].

Besides the inclusion of lexical information (e.g. a set of verbs of a similar semantic-type that recurrently select for a target noun in the subject position), linguistic models also take into account the crucial role that syntax can have in defining the distributional properties of classes by specifying patterns made of a combination of lemmas and part of speech. Because the lexico-syntactic patterns are linguistically-motivated, they are directly based on linguistic knowledge, and thus can provide a more precise representation than simple co-occurrence counts or bag-of-word models. Moreover, the use of patterns based on linguistically-motivated information can predict other words that can occur in that same position occupied by the target noun. In this way, these patterns are considered to cue a semantic property that a set of words, or class, may have in common. This information is then used as an indicator for members of that class.

[Hearst, 1992] conducted one of the first studies using lexico-syntactic patterns to automatically acquire lexical information, more specifically, the hyponymy lexical relation, from unrestricted text. The use of lexico-syntactic patterns to construct a model highlighted the benefits of a method that does not require pre-encoded knowledge (from a syntactic dependency parse, for instance). Furthermore, it demonstrated its applicability across wide ranges of text, especially because they defined and used easily recognizable and frequently occurring patterns that capture the particular lexical-semantic relation of hyponymy, such as: *"X is-a Y"* and *"X and other Y"*. Although this strategy obtained promising results, the scope of relations that the patterns were able to find were small in comparison to the size of the corpus.

Recently, [Panchenko et al., 2012] has expanded upon the original [Hearst, 1992] approach to gain coverage and recall, improving upon the limitations of the scope that the original patterns were able to reach. Extending a set of the 6 original [Hearst, 1992] patterns with 12 more linguistically-motived patterns, for a total of 18 patterns, both positive and negative contexts were also included to exclude meaningless information and find further correlations between related words. However, recall increased only with the increase of corpus data, rather than the inclusion of more cues. Thus, the improvement of recall was attributed to an increase in corpus size. In this way, although corpus size was increased, the word representations were still reliant on the limited information that the patterns were able to extract. To further overcome the sparsity resulting from manually-identified cues, [Snow et al., 2006] used known hypernym/hyponym pairs to generate training data for a machine-learning system, which then learned indicative lexico-syntactic patterns. However, the detection of these patterns required syntactic information from an English-language dependency parser; hence, this approach was dependent on external resources and the independence of external resources was one of the advantages of the manually-identified approaches.

[Merlo and Stevenson, 2001] built a linguistic model by selecting very specific ad-hoc linguistically-motivated cues for classifying verbs undergoing different types of diathesis alternations, along the line of the study conducted in [Brent, 1993]. They selected linguistic cues to classify English verbs into three classes: unaccusative, unergative and object-drop. For instance, animacy of the subject is a significant cue for the class of object-dropping verbs, in contrast to verbs in unergative and unaccusative classes. In contrast, [Joanis et al., 2008] built a linguistically-motivated model using general linguistic information, such the frequency of filled syntactic positions or slots, tense and voice of occurring verbs, etc. to classify English verbs into a number of [Levin, 1993] classes. [Jones and Mewhort, 2007] built a different type of linguistically-motivated model, based on semantic norms, that addresses word order phenomena, that considered the lexical probability that a word like “Hong” is highly likely to be followed by “Kong” but unlikely to be preceded by it. Finally, [Moon and Erk, 2013] represented the meaning of a word as a probability distribution over potential paraphrases. The results obtained by their model confirmed the importance of linguistic models, as they concluded that the consideration of syntactic order along with collocation information is crucial to performance.

Considering the above, linguistically-motivated models can, on the one hand, offer a higher degree of generalization to the word representation in comparison to unstructured models, through the use of pre-defined, manually-crafted or automatically-extracted linguistically-motivated lexico-syntactic patterns indicative of a specific semantic property or relation of a class. Furthermore, these linguistically-motivated models can models can often be highly predictive even when there is not a large amount of corpus data available, in contrast to the unstructured models, we also observe this result in Chapter 4. On the other hand, the fact that these patterns tend to be handcrafted, learned or manually-selected was a limitation to the full representations of words in corpus data. Moreover, sometimes these patterns can also unintentionally introduce noise into the model, especially in the case that the pattern is to general, such as the case of [Joanis et al., 2008].

Structured distributional models

Structured distributional representations are built by collecting corpus derived information in the form of word pairs and dependency relations [Grefenstette, 1994, Padó and Lapata, 2007, Baroni and Lenci, 2010]. These representations can organized in the form of tuples, including the word pairs and parser-extracted syntactic relations or lexico-syntactic patterns linking the pair [Grefenstette, 1994, Lin, 1998b, Lin et al., 2003, Poesio and Almuhareb, 2004, Erk and Padó, 2008, Padó and Lapata, 2007].

One of the major advantages of the structured models is that they tend to include information from a syntactic-dependency parser. Thus, these models also take into account the crucial role of syntactic structures in the distributional behavior of words. However, on the other side of the coin, these extremely features are also very fine-grained, and thus it also tends to be very sparse. The use of fine-grained features makes it more difficult to generalize upon.

The work presented in this thesis focuses specifically on Distributional Memory model (DM: [Baroni and Lenci, 2010]). The DM model was proposed as a general-purpose resource for semantic modeling. It consists of work-link-word tuples, which are extracted with different levels of lexicalization. The framework of DM was designed to exploit corpus data to its full extent for any type of semantic task and, furthermore, the tuple structure that it uses for features attempts to overcome the limitations of ad-hoc or manually-constructed patterns. In this way, this model can exploit different views of extracted data and different algorithms to tackle various tasks by collecting just one set of statistics from the corpus data.

Extensive and systematic studies have been conducted with the DM model, including but not limited to similarity judgments, synonym detection, noun categorization, detection of selectional preferences, etc., which have demonstrated that it is both general, yet comprehensive enough to address a variety of semantic tasks. Overall, in the large battery of experiments considered in their seminal work [Baroni and Lenci, 2010] report that in nearly all of the considered test sets, their best implementation of the DM model is at least as good as other models reported in the State of the Art. In related work, [Blacoe and Lapata, 2012] used the DM model to address to problem of modeling compositional meaning for phrases and sentences, while the DM model was used by [Lenci, 2011] to represent the expectations of the subjects about the most likely words co-occurring in given syntactic role in order to address the problem of how thematic fit is dynamically updated depending on the way other arguments are filled. Furthermore, [Lenci and Benotto, 2012] successfully model hypernyms in English using DM, while its usefulness in modeling semantics has also, importantly, proved to be multilingual, as demonstrated by a DM model built for Croatian [Šnajder et al., 2013] .

Considering the above, we consider that the structured DM model is a versatile model that can be used to address many semantic phenomena, including but not limited to classification and lexical acquisition, as demonstrated in [Baroni and Lenci, 2010]. In the work presented in this these, we refer to DM as the TYPE$_{\text{DM}}$ instance of DM, which is readily available for download and use.[1]. We provide further details on this instance of DM in Section 4.3

[1] http://clic.cimec.unitn.it/dm/

2.1.2 Word-Embedding models

Proposed and recently adapted for use in NLP tasks by various authors [Bengio et al., 2003, Collobert and Weston, 2008], Word-Embedding (WE) models, based on neural network approaches, map words into a low-dimensional spaces, in contrast to count-context models. Unlike the distributional models and vector spaces that count co-occurrence information, as previously described, WE representations do not directly encode frequency information into a vector to represent lexical items; rather, these spaces offer a mapping from raw corpus data to a vectorial space that represents the similarity of lexical items in similar contexts. By mapping word occurrences into dense feature vectors, WE representations learn or assign similar vectors to words occurring in similar contexts. Practically-speaking, these models represent each word as a n-dimensional vector of real numbers. [Baroni et al., 2014] eloquently defined the difference of WE models from count-context models, which we quote below:

> "Instead of first collecting context vectors and then re-weighting these vectors based on various criteria, the vector weights are directly set to optimally predict the contexts in which the corresponding words tend to appear. Since similar words occur in similar contexts, the system naturally learns to assign similar vectors to similar words."

WE representations introduce the novel idea of directly capturing the similarity between words by assigning them similar vectors according to the contexts that they are observed in. Thus far, the count-context models we have reviewed are limited by certain obstacles: mainly vector sparsity and issues with ambiguity due to selectional preferences in context. Furthermore, WE representations offer an additional abstraction step, by both mapping and tuning the information encoded into the vector that dramatically improves accuracy. Along this line, WE provides a cleaner representations to a classifier that do not rely heavily on the ability of the classifier to select for the most distinctive class-indicative features.

Considering the above, WE assign a similar vector to two (or more) similar words in terms of co-occurrence. Moreover, feature selection is no longer necessary because the actual representation of the word is produced in the learning phase and the number of features is externally defined. These distributed word representations are usually learned by means of the gradient descent back-propagation algorithm in order to minimize the differences between training samples. Intuitively, the neural networks that WE representations are based on take into account the observed word-context pairs and induce latent parameters on the basis that words that appear in the same contexts have similar parameters.

Furthermore, note that WE representations optimize the global probability distribution in order to meet the condition that all word-context pairs observed indeed came from the corpus data compared to a corrupted corpus, which contains the "negative" samples used for training neural networks. [Mikolov et al., 2013] provides a detailed explanation of this process, which is not reported here as it goes beyond the scope of this thesis. Thus, it has been demonstrated that the learned vectors indeed capture syntactic and semantic similarities [Mikolov et al., 2013].

Additionally, the learned vectors have proved to be very useful for different NLP tasks, and perform better than count-context models (see [Baroni et al., 2014] for a comparison). The use of these word representations have proved to be very advantageous in the task of "semantic similarity evaluation" due to the fact that these representations, in a continuous dimensional space, permit the discovery of semantically similar words with Euclidean methods, such as the cosine distance, for instance. One of the most referred to work in this area is the WORD2VEC representations developed by [Mikolov et al., 2013], whose code and data experiments are freely available for use and download[2].

[Levy and Goldberg, 2014a] extended the [Mikolov et al., 2013] system that uses only raw corpus data to build representations to also include the syntactic information from a dependency-parser. Again, the objective was to reduce the scope of "co-occurrence" (or context windows) to words that are indeed in a dependency relation with the word in question. Their results in similarity evaluation tasks are also very encouraging, reporting no issues due to data sparsity, which led us to adapt their system for our own classification tasks.

The use of WE representations in this thesis is a direct result of the temporal implication that they had on the State of the Art. In the summer of 2014, the main experiments to be included in this thesis were completed. Yet, in parallel to the completion of the original programmed work, there was an explosion of WE to build distributional word representations that was obtaining results that were completely outperforming all of the other models that we had studied up until that point.

[2] https://code.google.com/p/word2vec/

Because the main objectives of this thesis contemplate improving the performance of lexical-semantic classes by handling obstacles related to this task such as sparse data and polysemy, the dense and compact, yet informative word representations built with WE provided a very logical and viable alternative to overcome these issues, especially encountered with the count-context models. For these reasons, in the summer of 2014, we implemented classification experiments using WE representations to train our distributional models. After our preliminary experiments reported results with an increase in $F1$-Score of approximately 8 points, we conducted all of our experiments using WE representations. Thus, due to this temporal change in the construction of distributional models for machine-learning tasks, the last Sections of both Chapters 4 and 5 report upon our results and reflect upon the implications of these method to build word representations to automatically acquire lexical semantic information.

2.2 Automatic acquisition of general lexical information

In the work presented in this thesis, we approached the automatic acquisition of lexical-semantic information by classifying nouns into known lexical semantic classes. In addition to the classical distributional hypothesis [Harris, 1954], we considered that lexical semantic classes are products of emergent properties of a number of words that recurrently co-occur in a number of particular contexts, following [Bybee and Hopper, 2001] and [Bybee, 2010]. Furthermore, [Korhonen, 2010] proposed that class-indicative properties can be manifested in statistical differences over uses of different features, which permits the collection of this information to be used as a more cost-effective solution to easily—and automatically—acquire lexical-semantic class information.

In the framework of usage based construction grammar theories [Goldberg, 2006] and supported by psycholinguistic and cross-lingual evidence, a lexical-semantic class is a generalization that comes about when there is a systematic co-distribution for a number of words and a number of contexts in the broad sense. Thus, different contexts where a number of words tend to occur become overt linguistic cues of a particular semantic property that a set of words has in common and, therefore, upon which members of that class can be recognized. Simply put, words that belong to the same lexical-semantic class will tend to share a number of particular contexts.

Construction-based grammar hypotheses allow us to predict that there are a set of word occurrences, not in one or another discriminating context, but a number of them what constitutes a class mark. The structuralist notion of markedness [Jakobson, 1971, Bybee, 2010] allows for principled predictions about the probability of observing these contexts if understood as class marks. Furthermore, [Grimshaw, 1990] proposed that linguistic information can be provided by occurrence distribution, as is usually done in linguistic theory to motivate lexical classes. Moreover, this markedness notion would allow us to predict that members of the class will appear in marked contexts, as well as in unmarked contexts, although the unmarked contexts can be interpreted either as an instance of a non-member, or a situation where the distinction is irrelevant [Jakobson, 1971]. Hence, the selection of features to build word representations is one of the most important tasks to build distributional models to automatically acquire lexical semantic information.

Considering what types of distributional information to use as a feature, or a characteristic, of a lexical semantic class is a critical methodological decision when building word representations. This is because it directly affects what types of information are (or are not) used, and therefore it directly affects the representation provided to a classifier. Currently, count-context models define the distributional properties used to construct representations in terms of *documents* [Landauer and Dumais, 1997,Griffiths et al., 2007], which captures information available from entire document; *lexical collocates* [Lund and Burgess, 1996, Schütze, 1998, Bullinaria and Levy, 2007, Bullinaria and Levy, 2012], which capture information from a defined context window, *PoS tags* [Joanis et al., 2008], which use syntactic category information for representation; *syntactic structures* [Dorr and Jones, 1996], which construct representation with syntactic information from a dependency parser; *lexico-syntactic patterns* [Hearst, 1992, Grefenstette, 1994, Merlo and Stevenson, 2001], which use pre-defined patterns including lexical information and PoS tags considered linguistically relevant; and *tuples* [Kilgarriff, 2003, Erk and Padó, 2008, Padó and Lapata, 2007,Baroni and Lenci, 2010], which consider word pairs and the parse-extracted syntactic relation or the lexico-syntactic patterns linking them.

Currently, there is no consensus on what features to use for general acquisition tasks, and in many cases, the feature sets are constructed ad-hoc to address the objectives of that specific task. There have been some attempts to standardize the feature selection process for distributional semantics, such as the proposal of the DM model [Baroni and Lenci, 2010] that we study in Chapter 4 for classification. More recently, WE strategies [Mikolov et al., 2013, Levy and Goldberg, 2014a], use a uniform process that exploits the information from the entire corpus to map target noun occurrences into a dense feature space, thus eliminating the need for a specific feature selection, as required for count-context models.

2.2.1 Information of verbs and other parts of speech

One of the first studies that identified and counted topical cues for classification was conducted by [Brent, 1993], who hypothesized that in language acquisition it is possible to approximate cues to determine syntactic structure, by stringing local-surface cues together rather than global constraints. Furthermore, the results obtained confirm that it is possible to discover relevant syntactic structures in an utterance without prior knowledge of all of the words. A possible set of cues was proposed to identify information in English subcategorization frames, considering for instance function morphemes, utterance boundaries and knowledge of proper names, combined with inference mechanisms. On the one hand, they concluded that syntactic frames can, in fact, be identified with relatively simple surface cues, yet, on the other hand, this simplicity of some of these features was also a limitation because they did not achieve very high accuracy, primarily due to the ambiguous nature of many words.

[Levin, 1993] manually categorized verbs based on their diathesis alternations. This pioneering study served to justify the hypothesis that there was a direct link between the syntax of a word and the semantic arguments that are able to constrain it. Furthermore, this manual classification of verbs provided the baseline for almost the entire State of the Art in automatic verb classification. [Dorr and Jones, 1996] used automatic methods to further the correlation between the semantic classes of verbs and patterns of grammar codes. They verified the central thesis of [Levin, 1993] by automatically extracting syntactic information from machine-readable resources (MRDs). Two different experiments were conducted; one that considered polysemy of verbs and one that did not. In the case of the former, information for all of the different classes of the verb were encoded in one representation, while in the case of the later; individual representations were constructed for each class of the verb. Unsurprisingly, the results reported for the latter were much clearer, as they were based on representations that were already disambiguated. Furthermore, these results highlighted the critical need for any lexical acquisition system to accurately handle ambiguity.

[Lapata, 1999, McCarthy, 2000] both further built upon the work presented by [Dorr and Jones, 1996], proposing the use of corpus data to extract feature information, as a way of by-passing the reliance on the availability and adequacy of MRDs. Additionally, the use of corpus frequency information was proposed to estimate the probability of a given alternation. Following [Dorr and Jones, 1996], they both further confirmed a significant relationship between the similarity of selectional preferences at the target slot and the grammatical restrictions that can be both identified and learned. However, the use of corpus data to extract information actually increased the effects of data sparsity, although it did positively increase coverage and, consequently, recall. As we will see in the forthcoming Chapters, although the increase of data should logically decrease the sparsity of information used, we will see that it is not necessarily the case, as more data does not imply higher quality representations; rather, it is the information within the word representation that is be the key to overcoming the sparse data problem hindering most corpus-based studies.

[Lapata, 1999, McCarthy, 2000, Lapata and Brew, 1999] further strived to overcome data sparsity using probabilistic models that combined linguistic knowledge via Levin's classifications and frame frequencies acquired from the BNC. Notably, [Lapata and Brew, 1999] achieved an accuracy of 91.8% with Levin classes and 83.9% with class ambiguous verbs using the information available in subcategorization frames (SCF) to disambiguate verbs. However, their system was heavily reliant on the verb class information provided by Levin—any verb that is not a part of the Levin list, would be represented by a zero. Furthermore, there was also a strong emphasis placed on the importance of frequent classes that did not take into account how individual verbs can be distributed across classes. This can be problematic, as we also saw in Chapter 4 of this thesis, because most ambiguous nouns do not occur symmetrically in all of their potential classes in corpus data. Thus, when considering only the most frequent class, the representation can be biased or skewed incorrectly because the verb may occur in several classes with a different frequency, which can lead any classifier to make an incorrect classification system.

[Schulte im Walde, 2000] automatically obtained the semantic classes of verbs using probability distributions over verb SCFs. The verb frame types used for as syntactic descriptors contained at most three arguments, including nominative, dative, accusative, noun phrases, reflexive pronouns, prepositional phrases, expletives, non-finite clauses, finite clauses, and copula constructions. Following [Lapata and Brew, 1999], [Schulte im Walde, 2000] empirically demonstrated that semantic classification for German verbs is largely recoverable from the patterns of verb-frame co-occurrences. [Lenci, 2014] also proposed the information in SCFs as appropriate features to semantically classify verbs in Italian, while also emphasizing the importance of further understanding the meaning components, i.e. the semantic features, that are relevant to analyze verb meaning.

[Stevenson and Merlo, 1999] began to move away from subcategorization frames, expanding on the idea of classifying verbs according to linguistically-motivated grammatical features extracted automatically from corpus data. They expected that the semantic role assignments of verb classes to be reflected in their syntactic behavior, and consequently in the distributional data they collected from corpus data. Moreover, the extraction of features directly from corpus data further reduced reliance on external resources, such as MRDs. More specifically, in their seminal work they propose to automatically classify verbs based on argument-structure properties. However, [Stevenson and Merlo, 1999] attempted to generalize upon the very fine-grained syntactic restrictions reported in [Levin, 1993] that [Dorr and Jones, 1996] based their experiments upon. Again, in parallel to [Dorr and Jones, 1996], they concluded that there is a significant relationship between classes of verbs and the syntax, in their case, argument structure.

[Schulte im Walde, 2006] further expanded on a series of experiments for the semantic classification of German verbs, parting from the idea that semantic verb classes can be generalized according to their semantic properties. This was done by capturing large amounts of verb meaning without defining the idiosyncratic details for each verb. Following the current trend, [Schulte im Walde, 2006] used a combination of SCFs, prepositional information and selectional preferences as features. Classification results coincided with those from a manual classification exercise, although manual correction and completion was still necessary. Furthermore, [Sun et al., 2008, Sun and Korhonen, 2009] also used SCFs, instead of syntactic slots, as features for classification, demonstrating that considerable additional improvement can be obtained also with semantic features in automatic classification. Differing from [Schulte im Walde, 2006], their feature sets included automatically acquired SCFs, along with (statistical) information related to the PoS tags, GRs (subject, object, indirect object associated with verb), argument heads, and adjuncts of verbs, as well as both shallow and deep syntactic and semantic features. Their results further justified the [Levin, 1993] hypothesis that verb classification relies not only on syntactic but also on semantic features.

[Merlo and Stevenson, 2001, Li and Brew, 2008, Joanis et al., 2008] considered a wider range of information, including also the semantic preferences of verbs, which consequently helped to alleviate some of the complications caused by sparse data problem widely reported in the representations derived from previous work. [Merlo and Stevenson, 2001] addressed the generalization of distinctions in argument structure by identifying linguistically distinctive features that exhibit distributional differences across the verb classes directly in corpus data. Moreover, they considered that the statistical distributions of these features contributes to the learning of the classification of the verbs. [Stevenson and Joanis, 2003] clustered verbs into lexical semantic classes, using a set of noisy features to capture broader syntactic and semantic properties of verbs, thus increasing coverage. They explored both manual, unsupervised and semi-supervised methods for feature selection, concluding that a manual selection of a subset of features based on the known classification performs better than using a full set of noisy features.

[Joanis et al., 2008] further developed the idea of [Stevenson and Joanis, 2003] to expand upon a general feature set. Results obtained demonstrated that a general feature space can achieve a rate of error reduction ranging from 48% to 88% over a chance baseline and across classification tasks of varying difficulty. However, their general feature space, including features such as syntactic slots, slot overlaps, tense, voice and aspect, and animacy did not generally improve the classification accuracy over SCFs. [Li and Brew, 2008] built upon the methodology reported in [Joanis et al., 2008] to explore an even wider range of features, focusing on mixing syntactic information with information from lexicalized slots, information derived from dependency relations, lexicalized co-occurrence information and adapted co-occurrence information. Furthermore, it was proposed to keep all prepositions and to replace all verbs in neighboring contexts of each target verb with their part-of-speech tags, and a combination of SCFs and co-occurrences. Empirical evidence indicated that both syntactic and lexical information are useful for verb classification.

Finally, [Merlo and Stevenson, 2001] demonstrated that a small number of linguistically-motivated lexical features are sufficient to achieve an acceptable accuracy rate (in their case 69.8) and that relevant semantic properties of verb classes (such as causativity or animacy of subject) may be successfully approximated through countable syntactic features. However, the use of patterns again limited the amount of information extracted, resulting in vector sparsity.

Furthermore, there has also been some work done on the automatic acquisition of information for other grammatical categories, such as adjectives and prepositions. [Celli and Nissim, 2009, Girju, 2009], for instance, studied the semantic classification of prepositions through their experiments to identify the semantic relations in complex nominals, while [Bannard and Baldwin, 2003] used distributional similarity to analyze prepositional semantics. In regards to adjectives, [Carvalho and Ranchhod, 2003] automatically disambiguated adjectives in Portuguese from nominal headers for PoS taggers, while [Bohnet et al., 2002] aimed to use automatic methods to classify adjectives in German. More recently, [Boleda et al., 2012b] automatically induced lexical-semantic classes of adjectives in Catalan, using both theoretically-motivated features that cue properties of each class, as described in literature, and PoS features Today, much of the work on the classification of adjectives is focused on Sentiment Analysis, which goes beyond the scope of this thesis, yet we note the seminal works of [Hatzivassiloglou and McKeown, 1997] and [de Marneffe et al., 2010], which used similar semantic methods to those reviewed here, based on distributional techniques, to successfully automatically obtain the subjective adjectives and their orientation from corpus data.

2.2.2 Information of nouns

The acquisition of lexical semantic information of nouns parts from the same hypothesis that frames the acquisition of any part of speech: there is a strong correlation between syntax and semantic meaning, i.e. the meaning of a noun is represented from observed or inferred contexts in which it is found. The work of [Hindle, 1990] represents one of the first attempts to exploit syntax for the acquisition of semantic information of nouns. They demonstrated the plausibility of deriving semantic relatedness from the distribution of syntactic forms, such as the distribution of subjects, verbs, and objects in a corpus of English text. The results obtained demonstrated modest success, yet they encountered a number of challenges that remained to be solved in future work, such as the consideration of polysemy, the use of non-content words and the need for very large corpus data, as well as a syntactic-dependency parser. [Lin, 1998a] also relied on the use of dependency relationships as word features for automatic thesaurus creation, more specifically distributional patterns of words from a parsed corpus were used to infer the meaning of an unknown word. However, polysemy was not accounted for as their representations were learned from all of contexts the word, the different senses of the word were all conflated into one representation, which can cause uncertainty in a class membership decision. [Hearst, 1992] attempted to avoid the need for pre-encoded knowledge for the similar task of hyponymy acquisition by identifying a set of lexico-syntactic patterns that are easily recognizable and that occurred frequently and across text genre boundaries and that indisputably indicate the lexical relation of interest. On the one hand, the use of these patterns eliminated the need for an external parser, while on the other hand, it further restricted the number of hyponyms recovered because not all cases occur within the specific patterns used.

Other attempts to automatically acquire lexical-semantic nominal information include [Light, 1996], which used only information from derivational affixes to classify nouns. Following [Hearst, 1992], these morphological cues were considered to be good surface cues at they were easy to identify, abundant and correspond to the needed lexical semantic information. However, one limitation of this desideratum is that there are many words in English that may not have a derivational cue, and the reliability of it on its own may be too low for many NLP tasks. [Gillon, 1992, Baldwin and Bond, 2003, Baldwin, 2005] considered different types of cues as features for nominal classification. For instance, [Gillon, 1992] used surface cues such as quantifiers, such as numerals, articles, modifying determiners, etc. to distinguish between count and mass nouns. [Baldwin and Bond, 2003, Baldwin, 2005], however, induced mass/count information from a parsed English corpus using parallel supervised classifiers that took into account morpho-syntactic information, such as head number, modifier number, subject-verb agreement, occurrence in *"N of N"* constructions, etc. In their experiments with nominal classification, [Bel et al., 2007] considered (among other lexical features, such as subcategorized complements and bounded prepositions) the local contexts in a PoS tagged corpus as features to classify Spanish mass nouns.

Finally, [Bel et al., 2010] used lexico-syntactic patterns to develop class-based lexica by automatic means, focusing on non-deverbal event nouns for both English and Spanish. Lexico-syntactic patterns were identified to characterize contexts in Spanish, where members of a given lexical-semantic class tended to occur, such as: nominal suffixes, prepositional phrases, nouns occurring as external or internal arguments of verbs, present of temporal quantifying expressions, such as *"two weeks of"*, the fact that non-deverbal event nouns will not be in prepositional phrases headed by locative prepositions, that non-deverbal event nouns have an external argument that can also be realized by an adjective. Furthermore, lexico-syntactic patterns were also identified to characterize contexts in English, where members of a given lexical-semantic class tended to occur, such as: process nominals and non-deverbal event nouns can be identified by appearing as complements of aspectual PPs, non-deverbal nouns may occur as external or internal arguments of occurrence verbs or time-related verbs, intention to register event nouns whose external argument, although optional, is realized as a genitive complement, etc. Although achieving an accuracy of almost 80% in English, like the other approaches for linguistically-motivated models, they came across two main obstacles: (i) noise, where nouns sometimes occur in contexts that were not aimed at, and (ii) sparsity of information in feature vectors, which is affected by the low frequency of some nouns, as well as the low frequency of the occurrences of some indicative contexts.

Two related tasks, Named-Entity Recognition and Classification (NERC) and Word Sense Disambiguation (WSD), also deal with assigning lexical items into categories or senses. However, they differ from lexical-semantic acquisition because they work on token occurrences and tend to adopt an enumerative approach as their main goal is to separate each individual sense of a given word. In the case of WSD, for instance, the system determines the singular sense of a word in a particular context, while in the case of NERC, the system identifies and classifies particular occurrences of proper names into an already predefined set of categories. These types of systems define word meaning by an enumerable, static set of senses per word. Yet, they do not consider that a word can be more than one sense in a single context. This does not always constitute the most accurate representation because it ignores cases of regular polysemy, which we address in the following section. Although these tasks go beyond the scope of the work presented in this thesis, for an overview of relevant and noteworthy NERC work see, for instance, [Nadeau and Sekine, 2007, Tkachenko and Simanovsky, 2012, Ritter et al., 2011], and for WSD see [Rigau et al., 1997, Atserias et al., 2005, Cuadros and Rigau, 2006, Navigli, 2009, Toral et al., 2009, Navigli, 2012, Agirre et al., 2014, Azpeitia et al., 2014].

2.3 Distributional representations and the Generative Lexicon

As we have identified throughout this literature review, most previous approaches to the automatic acquisition of semantic classes have mostly disregarded the challenge of polysemy by considering only monosemous or already disambiguated words or classes, by simply ignoring it or by discussing it only in the context of analyzing results obtained. Polysemy is a challenge for distributional models mainly because the word representations are learned from all of the contexts of a word in corpus data, and therefore the different senses of a word are conflated into one single representation. Dealing with the multiple distinct senses of a word in a distributional representation can be considered a research line in itself.

Typically, in external lexical resources, such as a dictionary or WordNet, sense distinctions are made by considering each sense independently. However, and critically, this does not take into account the (systematic) relations that may occur between the multiple senses of a word. This is particularly problematic when words allow for multiple selection, i.e. when different senses of the same lexical item can both be selected for in one context (see Example 1). Known as logical, or regular, polysemy this type of ambiguity has been shown to have well-defined properties [Apresjan, 1974, Pustejovsky, 1995, Buitelaar, 1998, Martínez Alonso et al., 2013] and has been consistently reported as a factor in lexical semantic acquisition tasks. Example (1) illustrates how the word *bank* is selected for both as a LOCATION noun, where the target noun is the modified nominal (i.e. constructed *bank*), and an ORGANIZATION noun, where the target noun has an agentive subject (i.e. *bank* offers).

(1) The newly constructed (LOCATION) bank offers special conditions (ORGANIZATION) to new clients.

Along this line, besides the logical theoretical implications, the acquisition of information regarding regular polysemy in distributional word representations can also reduce redundancy in lexical resources, as well as the need for many fine-grained sense distinctions, which is one of the major criticisms of WordNet. Yet, the distinction of nouns that represents this phenomenon are semantic. Thus, we place ourselves within the framework of a semantic theory since the acquisition of refined semantic features do not always emerge from a purely corpus-based collocation analysis. Along this line, we place ourselves specifically in the framework of the Generative Lexicon Theory (GL) [Pustejovsky, 1995], as it models the phenomenon of regular polysemy by internally and logically structuring the semantic composition of lexical items [Pustejovsky, 1995]. GL postulates various levels of representation to semantically represent words while allowing for the computation of meaning in context. The Qualia Structure (QS) is one of these levels, consisting of four roles (FORMAL: what an object is; CONSTITUTIVE: what it is composed of; TELIC: its purpose; AGENTIVE: its origin), which model the predicative potential of lexical items.

More specifically, the QS also models phenomena, such as lexical items inherently complex in their meaning. These complex-type nouns are defined by a logical pairing of senses denoted by their individual types [Pustejovsky, 1995, Pustejovsky, 2005]. Thus they are characterized by the properties of more than one class [Pustejovsky, 1995, Pustejovsky, 2013] and, critically, they exhibit characteristics properties of both classes in corpus data (see again Example (1)).

In this way, according to the GL, differing from simple types, complex types are composed of more than one constituent sense that can be recovered both individually and simultaneously in context. In other words, complex types are words of a semantic type made up to two classes ($x \cdot y$). [Ježek and Lenci, 2007] presented an analysis of a verb that has been well-characterized as a complex-type in GL literature (the Italian for *leggere* "to read") to determine if the selecting environments of internal arguments can be validated and refined using corpus data. Likewise, [Rumshisky et al., 2007] presented one of the first empirical regular polysemy models that explicitly and specifically addressed the study of complex-types nouns. They proposed a method for the automatic detection of selector contexts specific to the components of a complex-type noun. They built upon the work of [Pustejovsky et al., 2004], which used GL concepts to argue and demonstrate that word senses are not directly encoded in the lexicon of the language, but rather are a product of the so-called "selection contexts", which can be categorized into the QS.

Yet, even considering the above, there has still been very little empirical work regarding the modeling of complex-type nouns and their regular polysemous alternations [Copestake, 2013]. [Buitelaar, 1998] used WordNet as a basis to empirically identify nouns that regularly alternate between at least two WordNet senses to define the complex types described in the CoreLex data set. [Utt and Padó, 2011] built an empirical model, also based on WordNet, to make an ontological distinction between homonymous and polysemous nouns. [Boleda et al., 2012a] modeled the alternations defined in the CoreLex database. In this work, a general framework was designed to ground sense alternations in corpus data, rather than in WordNet. It generalized each alternation at the type level, above individual instances, to predict of alternations of an unseen word. [Martínez Alonso et al., 2013] defined a scheme to provide reliable human annotations for complex-type nouns in English, Spanish and Danish. Moreover, [Martínez Alonso and Romeo, 2014] outlined a methodology to reduce the workload of experts for complex semantic tasks, such as the annotation of complex-type nouns with all of their sense information. Furthermore, [Martínez Alonso, 2013] built a sense-prediction system to automatically find empirical evidence to justify the incorporation of a third underspecified sense for complex-type nouns in sense inventories. [Romeo et al., 2013a] using Word Sense Induction techniques to automatically induce the sense alternation of complex-type nominals in corpus data. [Ježek and Vieu, 2014] assessed the possibility to empirically distinguish between complex-type nouns and simple-type nouns through an analysis of co-predictability contexts, which have been postulated to be characteristic contexts of complex-type nouns.

There has been other work conducted that tried to model the different senses of a word, such as that of [Schütze, 1998], which represented words, contexts and senses in order to assign to each target word to its most similar semantic sense cluster. Although this method obtained promising results, it did not also consider the relations between words that are representative of multiple senses. Therefore, it does not accurately handle regular polysemy as it does not consider the possible relations that there are between the different senses of a word, which is critical to accurately represent of complex-type nouns.

2.3.1 Automatic acquisition of lexical information and the GL

The Generative Lexicon [Pustejovsky, 1995] was first introduced as a knowledge representation framework offering a rich and expressive vocabulary for lexical information. As one of the most difficult problems facing theoretical and computational semantics is defining the representational interface between linguistic and non-linguistic knowledge, GL was initially developed as a theoretical framework for encoding selectional knowledge in natural language. Along this line, GL differs from more traditional lexical organization as it does not assume that word meaning can be exhaustively defined by an enumerable set of senses per word. This has a two-fold benefit:

- it does not require the pre-encoded knowledge of all of the possible senses of a word, which can result in incomplete coverage;
- it can also incorporate the creative uses of words in novel contexts because it accounts for meaning generated in context, which can be crucial for the accurate treatment of lexical items that are found in slots selected for by a regular polysemous alternation.

Furthermore, [Pustejovsky and Ježek, 2008] argued that lexical representations built from evidence of distributional behavior alone are unable to fully explain the rich variation in linguistic meaning in language. [Pustejovsky and Rumshisky, 2008] further justified this point by exploring the so-called "tensions" between corpus data and the linguistic theory, such as GL, that models it. The main conclusion was that both corpus-based and model-based linguistics have roles in constructing an adequate characterization of language usage, and therefore both must be considered in the design and identification of features used to build our distributional models.

One of the most direct applications of GL in the field of distributional lexical semantics was the automatic acquisition of QS information. As previously described, the QS represents the entire semantic composition of a word, which ultimately determines both the semantic meaning of that word, as well as its constraints in context. Moreover, the main goal of automatically acquiring QS information has been described as a method to automatically acquire deep semantic lexical knowledge from corpus data.

[Yamada et al., 2007] proposed a method for automatically extracting the TELIC and AGENTIVE roles of nouns from corpus data. Their experiments were based on the identification of syntactic constructions that are indicative of verbs constituting the TELIC or AGENTIVE roles of a given noun. In parallel, [Cimiano and Wenderoth, 2007] identified lexical patterns that identify all of the noun properties as defined by the entire QS of a word. The main difference of using patterns that identify relations corresponding to Qualia roles from those defined by [Hearst, 1992], for instance, are that the semantic information is related to the entire semantic composition of a word, which is also indicative of the different relations of a word. [Baroni and Lenci, 2010] attempted to improve upon the results reported in [Cimiano and Wenderoth, 2007] using the DM model. Instead of manually-crafting patterns, they exploited the information already available in the tuples of the DM model. They were able to approximate the patterns proposed by [Cimiano and Wenderoth, 2007] to automatically extract QS information and the results obtained were slightly above the best reported by [Cimiano and Wenderoth, 2007], which served to further demonstrate the transferability and adaptability of the DM model for a variety of semantic tasks.

[Katrenko and Adriaans, 2008] expanded upon the method of [Cimiano and Wenderoth, 2007] to automatically acquire QS to investigate the use of this automatically acquired information and impact on a noun categorization task. Demonstrating the effects on classification when using the information provided by different levels of their automatically acquired QS, the FORMAL role was concluded to be sufficient for discrimination between the semantic classes of nouns, while the addition of information of other roles such as TELIC and AGENTIVE did not improve results. However, focusing solely on the automatic acquisition of information from Qualia role, again, did not consider all of the available co-occurrence information, as it was limited to the use of a small portion of corpus data.

2.3.2 Other applications of GL

The GL has also been adapted for use in many other applications. Large scale applications of the GL can be found in the European Union-funded SIMPLE project [Lenci et al., 2000], which used GL as a basis to build multi-lingual semantic lexica, due to the relations that can be detected from the QS. The lexicon built with the SIMPLE project provides QS information to its lexical entries and, more importantly, regular polysemous classes represented by a complex type, which established a link between the systematically related senses in the lexicon. [Pustejovsky et al., 2006] also developed a large ontology and dictionary to allow for more widespread access to GL-based lexical resources.

Finally, the common thread among the large variety of work presented in this Section is the deep relationship between the GL theory and corpus analysis.

Moreover, the work reviewed demonstrates that the theoretical postulates of the GL are sufficiently adequate to frame the empirical evaluation of our task to automatically acquire lexical-semantic information by considering the effects of this phenomenon on our word representations.

2.4 Thesis overview

Based on the literature review conducted in this Section, we identified the main obstacles to the automatic acquisition of lexical-semantic information that justify the work conducted in this thesis:

- sparsity, which negatively affected precision and/or recall in the results obtained;
- polysemy, which negatively affected precision due to the fact that some words can be a member of more than one class. This issue was typically bypassed as not being considered for the methodology or the authors tend to use already disambiguated words for training and evaluation purposes.

First, we consider the limitations that data sparsity presents to distributional word representations. Along this line, we focus on the features used to build different distributional representations. Yet, this is not a trivial task because the generalization of features can provide noisy information into feature vectors, while bag-of-word-type approaches do not have the advantage of syntactic information that more complex models contain. Thus, an empirical evaluation on different models must be conducted to identify the distributional information crucial to built efficient word representations. We explore this objective in detail in Chapter 4

Second, distributional word representations conflate all of the senses of a word, which results in the obstacle to identify the relation of each individual sense. Furthermore, all of the senses of a word do not occur equally in corpus data, resulting in a frequency bias toward one sense. This causes an imbalance of information between the senses in the feature vector. Due to this obstacle, machine-learning algorithms are not able to handle the lower-frequency senses due to their smaller amount (or lack) of information. Thus, a method that can accurately handle the distributional word representations of these types of nouns is critical. We explore this objective in detail in Chapter 5.

Chapter 3

DATA DESCRIPTION

The main focus of this thesis is the automatic classification of nouns using distributional word representations into lexical-semantic classes. This Chapter explains in detail the data that is used to both train and to evaluate our lexico-semantic classifiers in each experiments conducted. Furthermore, we also describe the corpora used to extract the distributional data used to build these supervised classifiers.

Each experiment described in this thesis, extracts distributional information from corpus data to build word representations. This information is extracted either with Regular Expressions, pre-defined context windows or using tools that extract particular and relevant information in the form of tuples or other types of pre-defined structures. Frequency information for each extracted feature is stored in an n-dimensional vector. Machine-learning classification algorithms then use these vectors to obtain probability scores regarding class membership of a given noun. The selection, construction and use of the data sets and corpora used to build the word representations that form the basis of each experiment conducted in this thesis, are explained in detail in the following Sections.

3.1 Data sets

In this Section, we describe the compilation, extraction and construction of the different data sets used in this thesis.

Our first step was to define the specific lexical semantic classes that we would study. The selection of lexical-semantic classes was based on the criteria that each class was lexically relevant, meaning that words of this class occur relatively frequently in context, and grammatically salient, meaning that the words of a given class demonstrated definitive grammatical and lexical tendencies. In our case, we selected five (5) predominant classes for the English language: EVENT, HUMAN, ORGANIZATION, LOCATION and COMMUNICATION_OBJECT. Henceforth, these classes will be referred to as: EVT, HUM, ORG, LOC, and COM, respectively. Table 3.1 contains examples of lexical items that pertain to each individual class.

To obtain lexical items that represent each of these classes, we consulted WordNet for its extensive coverage in English and for its overwhelming acceptance in the field of NLP as a sort of gold-standard indicator of lexical semantic classes [Miller et al., 1990]. Primarily designed as a computational account of the human capacity of linguistic categorization, WordNet is a lexical database that covers an extensive set of lexical-semantic categories that organize lexical meaning, or senses, through representative lexical items, in contrast from more traditional sense-enumerated dictionaries. The senses that categorize the lexical items of each different grammatical category in WordNet are assigned by humans. The nominal database of WordNet is the source from which we extracted our data sets. It is organized as a sense hierarchy, with *unique beginners*, or categories that are not subsumed by any other category, at the top.

class	Examples
EVT	*malfunction*, *accident*, *schism*
COM	*book*, *letter*, *summary*
HUM	*comedian*, *instructor*, *nurse*
LOC	*campus*, *ghetto*, *playground*
ORG	*administration*, *crew*, *staff*

Table 3.1: Examples of nouns that pertain to each selected lexical-semantic class studied

The objectives of this thesis require the exploration of distributional representation of all types of nouns, including lexically ambiguous words. Along this line, on the one hand, we considered many words that are clear cut "monosemous" or single-sense members of a given nominal lexical-semantic class, such as *girl*, which is a clear and exclusive member of the HUM class. On the other hand, we also considered polysemous or multi-sensed words, such as the noun *newspaper*, which provided a clear example of a noun that is lexically ambiguous.

(2) He was the editor of the newspaper (COM) at that time.

(3) The journalist was curious as to whether the newspaper (ORG) would pay her to report on the riots.

To further illustrate the lexical ambiguity of the noun Example (2) shows how the noun *newspaper* is selected for as a member of the COM class, referring to a physical object printed on paper and contains informative articles while Example (3) shows how it is selected for as a member of the ORG class, describing the administration that produces the physical object referred to in Example (2). We address the representation and handling of this phenomenon in detail in Chapter 5.

3.1.1 Monosemous data sets

One of the main goals of this thesis is to study different distributional word representations and their effect on nominal lexical-semantic classification. In order to obtain these representations and evaluate our models, we need data sets of words containing lexical-semantic class information. In order to obtain these data sets, we followed the methodology described below. All of these data sets will be available on-line for download and use in the final version of this thesis.

Each noun for the monosemous, or single-sense, data sets, was selected with a relatively simple procedure that was conducted for each noun class considered. For each lexical semantic class, we extracted all of the words from WordNet that contained a corresponding sense. In other words, we extracted all of the items that were tagged in WordNet as *people* for the HUM class; as *location* for the LOC class; as *group* for the ORG class; as *event* for the EVT class; and as *communication* for the COM class.

Class	Targets (Class Members)	Targets (Not Class Members)
EVT	260	260
HUM	246	246
ORG	138	135
LOC	157	156
COM	262	259

Table 3.2: Number of target nouns (including the distribution between members and non-members) used per lexical-semantic class

After extracting a lists of nouns for each classes, we then filtered each list using simple heuristics. We filtered out all compound words, multi-word expressions and words containing non-alpha-numeric characters. We then removed all proper nouns still included in the lists. Finally, we filtered out all nouns that had less than 3 characters, in order to remove any cases of acronyms, for instance. In this way, we ensured that our conclusions strictly concern common nouns. The figures reported in Table 3.2 reflect the final distribution of nouns that were used for training and evaluation purposes in the work presented in this thesis.

3.1.2 Polysemous data sets

Another main goal of this thesis is to determine how to handle distributional word representations that are affected by phenomena such as regular polysemy. Thus, besides the monosemous data sets described above, we also needed to encode information regarding the ability of certain nouns to also be selected for as a member of another class. To obtain this information, we conducted a human annotation task to annotate the data sets described above with information regarding the potential of each individual noun to instantiate a class pertaining to a regular polysemous alternation.

Human-annotated polysemous data sets

In Chapter 5, we primarily focus on the word representations and the classification of regular polysemous nouns. Thus, we needed to obtain a data set that also contains information regarding the potential of a noun to be systematically selected for as each sense that corresponds to the sense components that form a complex type. This information is usually not included in language resources, and it is specifically not included in WordNet [Boleda et al., 2012a]. For this reason, we conducted a human annotation task to manually build polysemous data sets from our automatically extracted "monosemous" data.

We enlisted three experts to annotate each individual noun from the data sets described in Section 3.1. Each expert was either native or highly proficient English speakers and, moreover, familiar with the phenomenon of regular polysemy. The annotators were asked to indicate whether each noun can be selected for in context as a member of a specific class (y) that pertains to a regular polysemous alternation ($x \cdot y$) and which was different from its original class (x) in the monosemous data sets.

Specifically, and for the sake of brevity, we only focused on the manual annotation of two specific regular polysemous classes: LOC·ORG and EVT·COM. Appendix A.2 provides more information and detailed examples regarding the exact annotation task conducted, as well as the scheme and results obtained. A noun was considered to be regular polysemous if a majority of the human annotators indicated that the noun can be considered a member of both classes that form the specified regular polysemous alternation. In this way, we used a voting scheme to select the nouns from which we built the polysemous data set. We included those nouns considered to be members of more than one class by at least two annotators. Thus, in the case of LOC·ORG, a LOC noun must also be marked that it can be selected for as an ORG noun and vice versa; in the case of EVT·COM, a EVT noun must also be marked that it can be selected for as a COM noun and vice versa.

To further illustrate this, consider the following examples that we observed in our data sets. In regards to the EVT/COM data sets, on the one hand the EVT noun *campaign* was marked by all three annotators to be able to be selected for as a COM noun in context. Thus in our data sets, this noun was tagged as regular polysemous. On the other hand, the EVT nouns *malfunction* and *disappearance* were not marked as able to be selected for as COM nouns. Therefore, these nouns remained marked as monosemous EVT nouns in our data sets. This was also the case for the COM nouns *portfolio* and *memo*, which were marked as not able to be selected for as an *evt* noun and remained marked as monosemous COM nouns in our data sets. In regards to the LOC/ORG data sets, the LOC noun *institute* was marked by all three annotators to be able to be selected for as a ORG noun in context. Thus in our data sets, this noun was tagged as regular polysemous. However, the ORG nouns *workforce* and *league* were not marked as able to be selected for as LOC nouns, therefore, they remained marked as monosemous ORG nouns in our data sets. Finally, the LOC nouns, such as *frontier* and *coastline*, that were not marked as able to be selected for as ORG nouns remained marked as monosemous LOC nouns in our data sets.

Table 3.3 presents the final number of nouns considered per polysemous alternation. Tables A.7 and A.6, for in Appendix A.2 provide the full lists of nouns and their final annotations.

	Complex types	Simple types
ORG·LOC	79	184
EVT·COM	99	381

Table 3.3: Number of complex-type and simple-type nouns in the data sets obtained by human annotation

Automatically-extracted polysemous data sets

The [Boleda et al., 2012a] data set[1] is a large and automatically constructed data set of nouns that belong to specific regular polysemous alternations. Like our data sets described above, the [Boleda et al., 2012a] data set was extracted automatically from WordNet and consists of a data set for each of the disemous alternations (or combination of two alternating classes) defined in the CoreLex database [Buitelaar, 1998].

As described in Chapter 2, the CoreLex database is a lexical resource designed specifically to study regular polysemy. It identifies the lexical items that share alternating senses in WordNet. [Buitelaar, 1998] built this database with a frequency criterion to filter out those combinations of WordNet classes that have only one member; this criterion is compliant with the GL guideline that postulates regular polysemy to be a recurrent phenomenon [Pustejovsky, 1995] and therefore, the alternation must be represented by more than one noun. [Buitelaar, 1998] identified a total of 529 polysemous classes that met this criterion; 60 of which are disemous and, along the lines of [Boleda et al., 2012a], are focused on in Chapter 5.

The 60 disemous classes in CoreLex are used as gold standards. Each of these 60 gold standards contain 40 lexical items, 10 of which defined as target lexical items (m, n), or true members of that regular polysemous alternation. The other 75% of the gold standard consists of what [Boleda et al., 2012a] defines as *distractors* or lemmas that do not instantiate the regular polysemous alternation. They defined three different types of distractors, which equally compose of the remainder of the data set (i.e. each composes 25% of the remainder of the data set):

- distractors that share m with the target but not n;
- distractors that share n with the distractor but not m;

[1] [Boleda et al., 2012a] made this data set fully available for download and use at `http://www.nlpado.de/?sebastian/data.shtml`.

- distractors that share neither m nor n with the distractor.

Essentially, [Boleda et al., 2012a] built these data sets to avoid the critique that their classifier was conducting coarse word sense disambiguations. By providing distractors, these data sets more effectively demonstrate whether a system is able to make finer-grained semantic distinctions regarding regular polysemous nouns, instead of broad coarse-grained classifications. A broad coarse-grained classification would simply include nouns representative of m, n, as well as m or n, which does not make any distinction between regular polysemous nouns and monosemous nouns.

3.1.3 Further details

Although each of the data sets that we built specifically for the experiments presented in this thesis were encoded with different types of semantic information (i.e. monosemous vs. polysemous information), they were used with the same methods as target nouns to extract information to build distributional word representations.

Each noun in the data sets was not contrasted with the actual occurrences of the nouns in corpus data because we used several different corpora, which we explain in detail in Section 3.2. Thus, the number of nouns available varied with the corpus used. Likewise, the lack of contrast more accurately mirrored a production-level system which does not control for what nouns are available in the data it is provided. Therefore, each noun appears x times in each corpus. In this way, by using the information of each individual corpus, our methodology takes into consideration the "messiness" that is always encountered when working with any type of raw language data.

For experimental and evaluation purposes, the data sets were balanced with respect to class members and elements not belonging to the class. The elements considered to not belong to a class were randomly selected from the set of nouns that did not contain a sense in WordNet that corresponded to the target class being classified.

3.2 Corpora

We used several different corpora in order to extract a wide variety of distributional data for our experiments and to ensure the transferability of our methods, as well as the validity of the results obtained. Table 3.4 presents and compares each of the corpora used in the experiments presented in the subsequent Chapters of this thesis.

Each corpora was a general domain corpus and we did not tune our methodology to any specific domain in the classification experiments. Furthermore, we selected corpora of various sizes to determine the effects of the amount of corpus data in our experiments. Consistent results were obtained from the experiments reported in this dissertation using different sized corpus, as will be explained in detail in the following Sections and Chapters.

Corpus ID	Sources	Token Number	Domain(s)	Experiments	References
IULA3M	Texts are selected and classified according to topics proposed by specialists in each area in English	3.2 Million	Law, Economics, Environmental sciences, Medicine, Computer science and Linguistic sciences	Section 4.1	[Castellví Cabré et al., 2012]
IULA21M	Texts are selected and classified according to topics proposed by specialists in each area in Spanish	21 Million	Law, Economics, Environmental sciences, Medicine, Computer science and Linguistic sciences	Section 4.1	[Castellví Cabré et al., 2012]
CRAWL30M	Texts were crawled using specific URLs as indicator.	30 Million	general, web-crawled	Section 4.2	[Pecina et al., 2011]
UKWAC60M	Texts were crawled using URLS ending with `.co.uk` as indicators	60 Million	general, web-crawled	Sections 4.2; 5.1 and 5.2	[Baroni et al., 2009]
BNC90M	Texts include extracts from regional, and national newspapers, specialist periodicals and journals for all ages and interests, academic books and popular fiction, published and unpublished letters and memoranda, school and university essays, among many other kinds of text.	90 Million	general	Sections 4.3 and 4.4	[Burnard, 2007]

LARGE3BN	Concatenation of UkWaC corpus (see above), BNC corpus and a mid-2009 dump of the English Wikipedia	2.83 Billion	general, parts web-crawled	Sections 4.3; 4.4 and 5.3	[Burnard, 2007];
					[Baroni et al., 2009];
					also following [Baroni and Lenci, 2010]

Table 3.4: Description and comparison of the different corpora used for the experiments described in this thesis

Each of the above corpora were tokenized and PoS tagged using the Penn Treebank Tagset for English. The BNC90M, UKWAC60M and the LARGE3BN corpora were also parsed, tokenized, PoS tagged, lemmatized with the TreeTagger5; they were dependency-parsed with the MaltParser, as described and detailed in [Baroni et al., 2009] and used also in [Baroni and Lenci, 2010].

Chapter 4

DISTRIBUTIONAL WORD REPRESENTATIONS FOR CLASSIFICATION

Distributional representations model the contexts in which a given word has been observed. However, what is considered to be a context is highly dependent on the distributional model used. For instance, some models can use co-occurring tokens as features, while other models consider combinations or lexical items as features and still others use PoS tags, syntactic information, etc. Nonetheless, and regardless of the type of context required by a distributional model, the extraction of feature information from corpus data can result in *sparse* word representations that, due to the heterogeneity of occurrences of words in context, contain a high amount of zero values in the vectorial representation.

Sparse vectors are especially problematic to the classifiers because the "real" information available in the word representation is undermined by the non-, or zero, values, which machine learning algorithms can not efficiently differentiate between. Furthermore, evidence occurring with low-frequency is typically disregarded by automatic systems, as demonstrated in the classification experiments of [Bel et al., 2007, Bel, 2010], which results in the word representations providing insufficient class-indicative information to a machine learning classifier. Thus, sparse data can affect word representations either because of the low frequency of many of the words to classify, which does not provide sufficient information for the classifier to make an accurate classification decision, or due to the low frequency of particular representative contexts that are needed to produce an accurate classification, which results in missing values being more "informative" than actual data.

The sparse data issue is not unique to any distributional model in particular, as each model has its own challenges when it comes to populating vectors with "real" numbers. For instance, on the one hand, the LING model, which we first explore in Section 4.1, uses manually-identified linguistically-motivated lexico-syntactic patterns to extract distributional information, which are dependent on both the occurrence of these patterns in corpus data and the occurrence of target nouns with these patterns in corpus data, which is problematic for many low-frequency nouns. On the other hand, some distributional models, such as the LINE model, use co-occurring tokens in a context window as features that are difficult to generalize upon, resulting in the need for large amounts of corpus data to achieve a sufficient predictive model; while more complex distributional semantic model, such as the Distributional Memory: DM [Baroni and Lenci, 2010], use features in the form of tuples containing information from a syntactic-dependency parser, thus containing very specific information, which result in high dimensional vectors with few "real" values.

Finally, we study Word Embedding WE in Section 4.4. WE representations attempt to reduce the large, and typically very sparse, dimensions of the more traditional *count-context* distributional representations. This is done by producing representations that directly map the occurrences of a word in corpus data, which has been said to be able to increase the expressive power of the representations [Bengio et al., 2003, Chen et al., 2013, Mikolov et al., 2013, Levy and Goldberg, 2014a, Levy and Goldberg, 2014b]. The advantage of these representations is that they have been said to reduce the dimensionality of the vector space, while they increase in predictive power. Nonetheless, WE representations cannot be inspected, and they are useful when a large (it is important to emphasize *large*, here, as these representations work best on corpora sized above 1 billion tokens) corpus is available.

Each distributional model has both advantages and disadvantages when building word representations. This Chapter primarily focuses on the use of distributional representations of words from corpus data and the impact of their different types of features used to classify nouns into lexical-semantic classes. In the work presented in this Chapter, we deal with how to accurately represent words so that classifiers can learn the relevant aspects of these representations to correctly predict the classifications of unknown words.

In addition, we also address the related problem of *noise*, or instances of nouns that are not members of a class but appear in indicative contexts of that particular class. This is because for many features there is not an $1 - 1$ association with a specific class, which can cause many of the surface patterns to be ambiguous.

Moreover, noise is problematic to word representations because: (i) the typical application of filters to remove unwanted information cannot be easily applied as we cannot afford to eliminate any available data and (ii) not all of the cases of noise are representative of this obstacle, as handling better the polysemy of nouns can play a role in overcoming this obstacle. Furthermore, we must consider a priori the limitations of the use of low-level tools, such as Regular Expressions over PoS tagged corpora, that can also introduce noise as they sometimes capture unwanted information.

4.1 Building a linguistically-motivated distributional model

In this Section, we focus on the identification of lexical-syntactic patterns that are indicative of nouns belonging to specific lexical-semantic classes. The work presented in this Section had a two-fold objective:

- to define the linguistic patterns that are indicative of members of each nominal lexical-semantic class;
- to assess to what extent data sparsity is an issue in the vectorial spaces constructed with linguistically-motivated information.

Our hypothesis is that lexical-semantic classes are bound by certain linguistic information, which is recurrent to members of a given class sufficiently enough to be both recognizable and discriminatory. More concretely, members of a given nominal lexical semantic classes can be identified by their occurrence in indicative properties that appear to be linguistically significant for a number of linguistic phenomena that characterize the class. We specifically adapted this hypothesis to identify particular linguistic contexts that represent distributional characteristics of a specific lexical class in corpus data and which also support the building of specialized classifiers. Moreover, rather than very indicative, exclusive cues to identify members of a given nominal lexical semantic class, we identified a set of indicative though not exclusive cues, for each lexical-semantic class. We did this because some features can still be useful, although they are not class-exclusive, because they occur more frequently in data and, furthermore, can also extract a larger amount of distributional information to provide to the classifier.

Finally, the work presented in this Section addresses lexical-semantic classes in both English and Spanish. We studied the lexical-semantic classification of nouns in two languages to demonstrate the validity and transferability of our methodology to different languages.

4.1.1 Identification of class-indicative lexico-syntactic patterns

In what follows, we defined the criteria used for the identification of the linguistically-motivated lexico-syntactic cues through a linguistic study of possible class-indicative, or marked, contexts. The results of the following experiments serve to verify that there is sufficient linguistic information from which we can extract distributional data that is sufficient to classify nouns into nominal lexical semantic classes. To begin, we first defined the major linguistic categories from which we part to identify the linguistically-motivated indicative features for each class. The linguistic categories include: *predicate selectional restrictions*, *grammatical marks*, *prepositional information* and *affixes*.

(i) **predicate selectional restrictions**: Predicate selectional restrictions are useful to identify class-indicative contexts because most verbs impose particular semantic restrictions to their subjects and objects. For instance, verbs like *happen* and *cause* are said to select different types of nouns as subjects, and these differences can be generalized under the lexical-semantic class concept. *Happen* selects for EVT nouns as subjects, whereas *cause* selects for agentive entities, among which HUM nouns, see for instance [Rumshisky et al., 2007].

HUM nouns in both English and Spanish can be identified as subjects of particular agentive verbs, and those that denote an intelligent act, such as *admire*, *talk*, *think*, etc. [Levin, 1993] exhaustively categorized the semantic roles of verbs in English and [Vázquez et al., 2000, Ferrer, 2004] conducted a similar task for verbs in Spanish, which provides useful information regarding the semantic class of the argument that the verb selects for. Selectional restrictions also apply to complements other than the subject and object. In the case of LOC nouns, verbs imposing certain selectional restrictions also impose subcategorization frame constraints in the form of prepositional complements [Jackendoff, 1983]. Thus, for English, verbs such as *come*, *go* and *arrive* are used as cues with different prepositions, while for Spanish this holds true with the verbs *venir*, *ir* and *llegar*. For English we have also identified some motion verbs that do not require prepositions: *enter*, *leave*, etc., which are also indicative of LOC nouns [Levin, 1993].

Selectional restrictions are also imposed by non-verbal predicative elements like adjectives that can restrict the nouns they combine with. While the strongest case is for collocations, there are also classes of adjectives that impose certain constraints on the classes of nouns that they modify. For instance, [Dixon, 1982] identified that *human propensity* adjectives tend to modify HUM nouns. Thus, the modification of a noun by this type of adjective is indicative of nouns belonging to the HUM class. Furthermore, the modification of a noun by particular adjectives can be indicative to nouns belonging to other lexical-semantic classes. For instance, geographical provenance adjectives that indicate nationality or religion, etc. are indicative modifiers of HUM nouns, while adjectives such as such as *far*, *remote*, etc. are modifiers indicative of LOC nouns.

(ii) **grammatical functions**: There are particular grammatical functions that require nouns from a specific lexical-semantic class. While the class of the subject is largely determined by the selectional restrictions of the predicate, as we have just exemplified, we can say that Indirect Objects, both in English and Spanish, preferably select for HUM nouns. Furthermore, to a certain extent, *by*-Objects in passive constructions are also occupied by HUM nouns. HUM nouns are also related to the dative alternation phenomena in English. In addition, Direct Objects in Spanish that are marked with the preposition *a* are mostly indicative of HUM nouns [Leonetti, 2004], as seen in Example 5.

(4) Ayer la chica conocía **a** su nueva profesora. // *Yesterday girl met her new professor.*

(5) La hermana ayuda **a** su hermano menor // *The sister helps her younger brother*

Certain LOC nouns can also have marks of grammaticalization, such as subject complements. For instance, consider the phrase "*A **school** is a place of learning.*", where *place* is also referring to the *school*, or a LOC where people can learn.

Adjuncts and modifiers of nouns to be classified are also indicative of certain lexical-semantic classes, especially when they also co-occur with particular particles. Clear cases of modifiers that describe the semantic characteristics of the noun that they modify are relative clauses headed with certain marked relative pronouns, such as *who* and *whom* (*quién* is the Spanish correlate). For instance, these types of pronouns clearly refer to a HUM antecedent, while *where* (or the Spanish equivalent *donde*, which is more restricted to LOC nouns than its English counterpart) are indicative of nouns belonging to the LOC class.

For English in particular, genitive complements (*my brother's book*, for example) tend to be filled with nouns belonging to the HUM class. Furthermore, possessive determiners are known to modify HUM nouns, as for instance in *his colleagues*. However, and as expected, these cues are not necessarily "exclusive" to nouns belonging to just one lexical-semantic class. Along this line, these cues represent instances of features that are only indicative in correlation with other cues as they cannot be considered exclusive indicators of the classes. This point is further justified by usage grammar theories, which postulate that emergent classes can be based on a number of marked correlations.

(iii) **prepositions**: Prepositions, especially those said to be *content* prepositions, can be very indicative of nouns that belong to a given lexical-semantic class. [Tseng, 2001, Rauh, 1993, Jackendoff, 1973, Taylor, 1993]. On the one hand, there are prepositions, such as *during* and the corresponding Spanish preposition *durante*, that are key indicative features to identify members of the EVT class. While, on the other hand, there are prepositions, such as *at*, *within*, *across* or *under*, that are good hints of LOC nouns in English.

Other informative marks prepositions such as: *en* and *según* (*in* and *according to*) are indicative of Spanish LOC and HUM nouns, respectively. Furthermore, nouns themselves can also combine with complements and modifiers that are selected by the semantics of the noun. Depending on the language, they can also appear as noun-compounds or as PPs, which can help to indicate the class of the noun that it is heading [Celli and Nissim, 2009].

(iv) **affixes**: The final category that we consider is affixes. Affixes provide crucial indicative information regarding the lexical-semantic class of a noun. Along the lines of [Bybee, 1985, Bybee, 2007, Bybee, 2010], we consider it to be an important distinguishing feature, even for only moderately inflected languages, such as English, because it provides evidence regarding the semantic preferences of the root. This was further confirmed with empirical evidence by [Light, 1996] who demonstrated that particular derivational affixes are good indicators of HUMAN nouns in some languages.

For English nouns, suffixes, such as *-er*, *-or*, *-ist*, etc., effectively identify HUM nouns, while for Spanish nouns, suffixes, such as *-aco*, *-ano*, *-dor*, etc., are good indicators of nouns belonging to the HUM class. Indicative suffixes for LOC nouns in Spanish, such as *-ería*, *-al*, *-dero*, etc., tend to be much more frequent than those discriminatory suffixes for LOC nouns in English (*-dom*, *-eria*, *-place*, etc.). Thus, the predicative power of a morphological cue is highly dependent on the language and the class for which it is discriminating.

Because we were primarily concerned with looking for features that were indicative of a given lexical semantic class, although not necessarily exclusive of that class, we also considered what we called "*negative*" features, or cues. Negative markers are features that **do not** occur with class members. Thus, the fact that a target is **not** seen with a negative feature and **is** seen with a positive feature actually provides useful information to our classifiers.

To further illustrate the concept of the negative features, consider the positive cue of a noun phrase headed with the preposition *durante* ("during") for the EVT class in Spanish. Grammatically, this preposition can never head a noun phrase that is not an EVT noun. See Examples (6), (7), and (8) to see the use of *durante* in several different noun phrases.

(6) *Durante* la fiesta - EVT. (*During* the party).

(7) #*Durante* el medico - HUM. (#*During* the doctor)

(8) #*Durante* la foto - COM. (#*During* the photograph)

As Examples (6), (7), and (8) demonstrate, a preposition like *during* cannot be used to identify nouns of classes other than EVENT. However, when we consider the correlation between several different features in order to arrive at a classification decision, the inclusion of a feature that does not pertain to a class can provide us with additional information about the distributional characteristics and behavior that these nouns should not have in context. Thus, the negative cues consist of information that is indicative of any other class.

In summary, each class was characterized by a number of different cues for each language that were manually identified following the guidelines mentioned before. Not all of them have the same distribution varying in sparseness (low frequency) and noise (also occurring with non-members of the class). Table 4.1 contains some examples of patterns used for each class, while Appendix A.1 contains a list of all of the cues defined under the categories elaborated above.

4.1.2 Experiments

Our experiments have covered English and Spanish nouns for the following classes: EVT, HUM, ORG, COM, and LOC. To elaborate on the results and conduct a multilingual comparison, we provide details on the experiments for LOC, EVT and HUM classes both for English and Spanish. For our experiments, we used the

Class	Examples of lexico-syntactic patterns
ORG	x-NN $(found\|establish\|organize)$-VBD
LOC	$(inside\|outside)$-IN $(the\|a\|an)$-$(DT\|Z)$ x-NN
COM	$(submit\|publish\|report)$-V* $(the\|a\|an)$-$(DT\|Z)$ x-NN
EVT	during-IN $(the\|a\|an)$-$(DT\|Z)$ x-NN
HUM	x-$(-er\|-or\|-man)$-NN

Table 4.1: Examples of lexico-syntactic patterns indicative of 5 different lexico-semantic classes, which we refer to as marked contexts

3-million token IULA3M corpus for English and the 21-million token IULA21M corpus for Spanish, as described in Section 3.2.

Each of the identified lexico-syntactic patterns have been formalized in a regular expression, which was then used to directly extract information from corpus data. The target nouns for each class were defined in the monosemous data set, described in Section 3.1 and each data set was balanced with respect to class members and elements not belonging to the class. The relative frequency for the occurrence of each noun with a defined context populated a n-dimensional vector provided to a classifier for each word.

For classification, we used a Decision Tree classifier in the WEKA [Witten and Frank, 2005] implementation of pruned C4.5 DT [Quinlan, 1986] and evaluated our representations in a 10-fold cross-validation testing environment.

The C4.5 (J48) [Quinlan, 1986] Decision Tree (DT) classifier is the first classifier that we used for the work presented in this thesis[1]. We selected this classifier, in particular, because it is fast to train and because it has been demonstrated to work well in a binary class prediction task [Kotsiantis, 2007, Wu et al., 2008]. However, the J48 classifier does not have a high tolerance to noise or to handle missing values. This can be especially problematic when a vector contains a high number of missing values for attributes that register the occurrence of a lexical item in particular contexts. This problem is especially exacerbated when a target lexical item occurs with low frequency in corpus data.

[1] In subsequent Sections, as a consequence of the analysis of the results obtained at each step, we refined both our method and the type of machine learning classifier used. Each classifier used is clearly described in each Section where it is merited. We explore and discuss the impact of different classifiers throughout the work presented in this thesis.

Note that a missing, or zero, value can indicate one of two scenarios: (i) that the missing value is caused by the token not being observed with a particular context, although it could in other data collections or (ii) that the token does not belong to the class and thus cannot occur in that particular context. This uncertainty between lack of useful information and an indicative non-inclusion of information creates uncertainty for this type of classifier because the zero values provide incompatible learning examples that make the word representations lose their predicative capacity, and therefore are not taken into account, or the missing values actually are considered by the classifier to be handled as informative features due to their large quantity and, thus, render ineffective the class membership decisions of the classifier [Bel, 2010].

In plain terms, there is no way to distinguish between the large amount of zeros that indicate a lack of occurrence of a target noun with a class-indicative feature and the similarity to two items that is due to their large amount of zeros in each respective feature vector. For this reason, we also place a special emphasis on the origins of false negatives in our results because they tend to caused by the high sparsity of distributional word representations.

4.1.3 Results

Table 4.2 presents the results obtained in our experiments for both English and Spanish in terms of accuracy, False Positives (FP), or those items incorrectly classified as members of the target class, and False Negatives (FN), or those items incorrectly classified as not belonging to the target class.

	English			Spanish		
Class	Acc (%)	FP(%)	FN(%)	Acc (%)	FP(%)	FN(%)
HUM	79.01	5.52	15.47	77.29	9.67	13.04
LOC	66.21	11.64	22.15	77.55	9.84	12.61
EVT	73.05	8.38	12.56	80.90	6.53	12.56

Table 4.2: DT classification results for English and Spanish, including accuracy, percentage of false positives and percentage of false negatives

With an average accuracy of 72.75% for English nouns and 78.58% for Spanish nouns, the overall results confidently demonstrate that the selected cues are informative in distinguishing the addressed lexical semantic noun classes.

Moreover, our results also demonstrate that it is possible to exploit the correlation between syntactic, morphological and lexical co-occurrences to identify members of a lexical-semantic class. Finally, there is no statistically significant difference between the average accuracy obtained for Spanish and that the average accuracy obtained for English, verifying that our approach is valid for different languages.

4.1.4 Discussion

Table 4.2 presents the results obtained from our experiments using the linguistically-motivated class-indicative cues described above. An average accuracy of 72.75% for English and an average accuracy of 78.58% for Spanish, allowed us to confidently confirm our hypothesis that nouns of lexical semantic classes can be automatically classified using the distributional information extracted from linguistically-motivated class-indicative cues. Furthermore, the results indicated that not all classes are equally identifiable using surface cues, as demonstrated by a decrease in accuracy of more than 7 points for the LOC nouns in English. This reduction of accuracy for the LOC class in English signaled differences in terms of the degree of grammaticalization. Not all nouns occurred with the same frequency, and likewise, not all classes were easily identified with surface marks. One of the clearest examples, for instance, of a high and frequent grammatical mark was for Spanish HUM nouns that tended to be marked as direct objects headed by the preposition *a*.

Moreover, the results presented in Table 4.2 demonstrated that FPs are a result of noisy instances in feature vectors, as we will explain below. Some examples of what we found to be noise are, for instance, the noun *pancarta* (*banner*), which was found after the prepositional expression *después de* (*after*), referring to the temporal sequence of a demonstration headed by it, is a clear case of coercion. Another example is the noun *cárcel* (*prison*), for which there are some occurrences of *años de cárcel* (*years of prison*) in the corpus. This would lead us to consider that *prison* or *banner* can either be interpreted also as an EVT or that the cue produced some undesired matching.

Furthermore and as expected, FNs show that the main problem is indeed the sparsity in vector representations. This was confirmed when we realized that there are 68 English HUM nouns (almost 13% of the total) that were not found in any of the contexts that were taken as cues. To further expand upon the issue of sparse data, we highlight that the sparsity in feature vectors was not necessarily an effect caused by the amount of corpus data. The English corpus is approximately 14% smaller than the size of the Spanish corpus and, yet, the results are not statistically significantly different, indicating the size of corpus data is not a significant factor when words are accurately represented with linguistically-motivated class-indicative distributional information.

Finally, the morphological cues for each class yielded, though applicable for both languages, have different results depending on the language. For example, in English, derivational suffixes were strong marks for the HUM class, as many HUM nouns are nominalizations. However, this did not hold true for the LOC class in English. In this case, the derivational affixes are quite noisy, in comparison to Spanish. This could be attributed to the fact that in English, the LOC class relies heavily on compounding such as *rice field* and *rose garden*, while in Spanish affixation is the preferred strategy, as illustrated in by the translations of *arrozal* “rice field” and *rosaleda* “rose garden”.

4.1.5 Final remarks

Overall, the results of the experiments support our hypothesis that linguistic information can be used to build comprehensive and accurate word representations. Moreover, the results confirmed that cue correlations, more than particular, exclusive cues, provided a strong predicative power. This was observed because none of the cues used prove to be exclusive of the class. Yet, our classifier was able to assign the correct class to a word by identifying the correlations between a series of our linguistically-motivated indicative marks.

The results presented in this Section also demonstrated the ubiquity of the sparse data problem in this type of approach. On one hand, the use of linguistically motivated cues as indicators toward a particular lexical semantic class can provide very precise, though potentially infrequent, information as the occurrence of a target noun with a cue is dependent on the corpus data available. On the other hand, a lack of occurrences of indicative contexts (or of a target noun with indicative contexts) will result in the classifier not to use the indicative information toward a particular class as such, rendering this evidence ineffective for nominal classification. Finally, the methodology presented in the Section has been proven to be language-independent, although the linguistically-motivated class-indicative cues, themselves, are not.

Our main conclusion from the work presented in this Section is that representations built with linguistically-motivated class-indicative patterns can effectively build distributional word representations but still require more distributional information to overcome issues of sparsity that we observed. Therefore, in the following Section, we design a strategy to effectively and efficiently utilize more of the available distributional information in corpus data that encodes specific behavioral characteristics of a lexical-semantic class.

4.2 Overcoming data sparseness with unmarked information

In order to further reduce the problems of sparsity observed in the results in Section 4.1, we turn to what we called “unmarked” contexts in order to provide additional and useful distributional information to classifiers, which we demonstrate in this Section.

We defined **unmarked contexts** to be very general contexts that are typically disregarded in distributional models because they have been thought to be too general to contribute any relevant class-indicative information to a classifier, as they tend to be observed with nouns of all lexical-semantic classes. Furthermore, the basic claim leading most authors to neglect this kind of context is that it presents a challenge for classifiers to accurately use this type of information in class membership decisions and, therefore, is bound to negatively affect results (see [Cooke and Gillam, 2008, Turney and Pantel, 2010, Bullinaria and Levy, 2012], among many others). In this way, unmarked contexts directly contrast with the class-indicative **marked contexts** that we previously defined in Section 4.1. At the same time, however, they corresponded to a large amount of corpus data was not being previously used, as discussed in Section 4.1.

Despite all this, we hypothesized that there are distributional differences in occurrences in this type of context, which can be identified and captured and learned from the data. As corpus data have a finite number of occurrences, we expected that nouns of a given lexical-semantic class have a different distribution than occurrences of nouns in those contexts that are not members of that class. Therefore, our objective in this Section was to define a strategy that accurately included the information extracted from unmarked contexts in the distributional word representations, resulting in more accurate class membership decisions.

According to [Jakobson, 1971], unmarked contexts occur frequently because they consist of all information not considered to be marked or indicative toward a particular class. Furthermore, [Bybee, 2010] claims that general contexts, not exclusive to a particular class (i.e. unmarked contexts, as defined above), are more frequent than contexts marked toward a particular class, as they occur with nouns of all classes. In view of this, it further became apparent that a large part of available distributional data was not being taken into consideration when these very general co-occurrences (e.g. co-occurrence with an article are not taken into account) observed with nouns of all classes were not considered at all in lexical semantic classification tasks. This means that while there is sometimes not enough information for classification, we are also not considering a large part of the information available. Although these contexts are not necessarily discriminative, we hypothesized that our distribution of the information extracted with unmarked contexts among members of a given class differs from its distribution among lexical items that are not members of the class, thus becoming an indicative characteristic of the class.

In Section 4.1, we saw the impact of the of sparse data problem, especially due to the fact that the classifiers cannot properly distinguish between zeros and missing values, affecting their ability to learn and to make accurate classification decisions. On one hand, the use of linguistically-motivated cues as indicators toward a particular lexical semantic class can provide very precise, although potentially infrequent, information as the occurrence of a target noun with a cue is highly dependent on the corpus data available. On the other hand, the infrequency of indicative information is problematic to classifiers, as a lack of occurrences can result in the classifier to not consider the available indicative information toward a particular class as such, rendering this evidence ineffective for nominal classification. It became apparent that we needed to design a method in order to capture and incorporate this type of information in feature vectors because it is a source of information that is not typically affected by low frequency and, therefore, it is always available to use.

Following previous work on the relation between these types of contexts and sense selection [Rumshisky et al., 2007], we hypothesized that the distribution of members of a class with respect to their occurrence in particular unmarked contexts is consistent, meaning that class members occur similarly with unmarked contexts and this behavior, which is divergent from the general occurrences of all other nouns in the corpus with the same context can be captured and used to inform classifiers and to improve results.

[Rumshisky et al., 2007] was one of the first to empirically provide evidence toward an asymmetry in the way certain word senses are used in language, preferably or rarely occurring in certain very general contexts (e.g. subject position, occurrence with an adjectival modifier, etc.). The generic asymmetry of use can occur across all argument positions and there are even some syntactic characteristics that can be strong indicators of a likely semantic interpretation for that noun, which would indicate the class that it is being selected for in that context, such as the difference in occurrence as a plural or a singular noun or a noun phrase headed with a definite or an indefinite article.

This type of asymmetry refers to a difference in distribution, more specifically, how semantically *neutral* contexts either co-occur with a target lexical item with either more or less frequency, depending on the sense in which a word is used, meaning that the contexts that a target item co-occurs with change according to the sense that they are being selected for in that context. To illustrate this concept, as argued in [Rumshisky et al., 2007], consider the EVENT noun *invention*, which tends to be selected for more often as a RESULT-EVENT modified with the determiner *the* in the argument position selected for by verbs such as *produce*, *explain*, *protect*, *develop*, *combine*, etc. than as a PROCESS-EVENT modified with the determiner *an* in the argument position selected for by verbs such as *welcome*, *avoid*, *stimulate*, etc., although the noun can be selected for as both.

Parting from the conclusions of [Joanis et al., 2008], we considered that there are class tendencies that can observed when using general contexts, such as unmarked contexts. To the best of our knowledge, the use of unmarked contexts in cue-based lexical semantic classification has not been previously explored, as these contexts are considered to introduce noise, or non-discriminative information, into the classifier, due to its claimed undifferentiated co-occurrence with nouns of all classes.

In this Section, we proposed a strategy that informatively includes unmarked contexts in word representations. We encoded the deviation of the behavior of each target noun, as observed in unmarked contexts, with respect to the average behavior of nouns in the corpus, based on our idea that using information regarding occurrences in unmarked contexts will provide additional relevant information to the classifier; especially with regards to those nouns of a given lexical semantic class for which their occurrences in marked contexts do not provide sufficient information for classification.

Considering such distributional evidence can increase the amount of information made available to classifiers, our main claim is that our strategy informatively includes this type of distributional information in classification tasks by taking advantage of a bigger portion of corpus data, thus, improving the accuracy of classifiers.

4.2.1 Identifying unmarked contexts

In contrast with mainstream approaches to cue-based lexical-semantic classification, we argued for the inclusion of a type of distributional information typically not considered to be indicative of class membership, and thus not informative to automatic classification systems. These very general contexts of occurrence typically disregarded, as they are thought to be too general because they occur with nouns of all lexical-semantic classes, and therefore thought to not contribute any relevant information. At the same time, they correspond to a large amount of corpus data that is a priori not considered due to the assumption that it does not provide any class-indicative information.

Following the conclusions of [Rumshisky et al., 2007] regarding asymmetries in the distribution of word senses in general contexts, our hypothesis is that the distribution of members of a class with respect to their occurrence in particular unmarked contexts is consistent and thus can be captured and used to inform classifiers and improve results when considered along with other indicative, or marked, contexts. Furthermore, the inclusion of unmarked contexts alleviates problems caused by data sparsity in classification tasks by providing additional information to classifiers. To assess to what extent this information can be used in classification tasks, we had to identify such contexts and verify whether our hypothesis was confirmed, i.e. if different lexical classes showed significant variations in terms of distribution that might be explored to augment the amount of information made available to classifiers.

Considering the characteristics of the contexts discussed above, we identified 32 unmarked contexts under a frequency criterion (see Table 4.3 for a description of the different contexts identified)[2] for English nouns from the HUM, LOC, ORG, EVT and COM lexical-semantic classes. To identify these contexts, we considered that the more frequent contexts will combine with more nouns in the corpus and thus should not be marked for any restricted set. However, although they are not considered to be class marks, we expected these contexts to be asymmetrically distributed between lexical semantic classes, in an analogous way to what was observed by [Rumshisky et al., 2007], with regard to the distributional behavior of different word senses in language use.

We first studied the distribution of these contexts in a web-crawled corpus (see CRAWL30M in Section 3.2). We compared the distribution of each context

[2] The full list of formalized regular expressions used to extract distributional information for unmarked contexts in corpus data is freely available for download and use at `http://repositori.upf.edu/handle/10230/24562`

over all the nouns in the corpus and over nouns defined as part of a specific lexical semantic class, according to our data sets (see Section 3.1 for detailed descriptions of the construction of the data sets used). We calculated the average of occurrence of each noun that pertained to a particular lexical-semantic class in a specific unmarked context, as well as the average of occurrence of all the nouns in the corpus with that same context; we then determined that if there was a statistically significant difference between the behavior of nouns from specific classes and the behavior of nouns in general with regard to the contexts identified as unmarked.

Feature Type	Description	Examples
article	target noun preceded by a(n) (in)definite article	$(a\|an)$-$(DT\|Z)$ x-NN or (the)-$(DT\|Z)$ x-NN
number	target noun in plural/singular form	x-NNS or x-NN
copula	target noun as subject/object of verb to be	x-NN be-VBZ,or be-VBZ,x-NN
modifiers	adjective or nominal modifier preceding target noun	x-JJ x-NN or x-NN x-NN
preposition of	target noun preceding/following the preposition *of*	x-NN of-IN or of-IN x-NN
subject of V	target noun as subject of each of the 20 most frequent verbs in the corpus	x-NN$(have\|get\|make)$-VB(Z D)

Table 4.3: Description of unmarked contexts identified and used in our experiments

The results showed there were, in fact, statistically significant differences $(p < 0.05)$[3] in the behavior of nouns in particular classes with regard to certain unmarked contexts. For instance, the occurrence of COM, ORG, LOC, and HUM nouns with a definite article (*the*-DT) showed to be divergent from the average. The occurrence with an indefinite article (*a*|*an*-DT) proved to be significantly different for LOC nouns, while the co-occurrence with an adjective (x-JJ) was significantly different for COM nouns. Thus, the empirical evidence obtained confirmed that there are differences in the behavior of particular lexical semantic classes with regard to their occurrence in unmarked contexts. Thus, the next step consisted in determining the best way to make this information available to classifiers.

To mirror the specificity of the distribution of each noun with regard to each context considered, we subtracted the mean of occurrence of nouns in each context from the actual occurrences of the target noun represented by the vector in that same context to obtain each feature f, as defined in Equation 4.1, where c_i represents a given context, t a target noun, n any noun belonging to N, the set of

[3] The statistical significance was calculated using Student's t-test (cf. [Krenn and Samuelsson, 1997]).

all nouns in the corpus, and *freq*, the frequency of occurrence (e.g. = frequency of occurrence of the target noun t in context c_i).

$$f = \frac{freq(t \mid C_i)}{freq(t)} - \frac{1}{|N|} \sum_{n \in N} \left[\frac{freq(n \mid c_i)}{freq(n)} \right] \tag{4.1}$$

We can encode the deviation of the behavior of that noun with regard to the general behavior of all nouns in the corpus using the difference between the number of occurrences of a given noun and the average occurrence of all nouns in a specific context. Under the hypothesis that nouns of the same class display similar tendencies in terms of deviant behavior in the contexts considered, our strategy, therefore, provides relevant information to the classifier. We apply our strategy to two different corpora making apparent its robustness.

4.2.2 Experiments

In order to evaluate the impact of adding unmarked contexts to the previously defined linguistically marked contexts, first, we had to extract distributional information regarding the unmarked contexts identified (see Table 4.3), as well as distributional information regarding class-indicative marked contexts. In our experiments, we used the class-indicative marked contexts that have been previously identified and described in Section 4.1 (see Table 4.1 for examples or Appendix A.1 for a complete list of cues for each class). Our experiments covered English nouns of the classes: COM, ORG, HUM, EVT and LOC from the data sets introduced in Section 3.1. Each data set was balanced with respect to class members and elements not belonging to the class. The final numbers of nouns considered for our experiment is presented in Table 4.4.

Class	ORG	LOC	EVT	COM	HUM
Class Members	138	157	260	262	246
Elements not belonging to the class	135	156	260	259	246

Table 4.4: Number of nouns included in data sets per class

For the purpose of the work presented here, we experimented with two corpora to determine the transferability and robustness of our method, independently of specific corpus data: the 30 million token CRAWL30M corpus to identify unmarked contexts and to train our classifiers and the 60 million token UKWAC60M corpus (see Section 3.2 for a detailed description of each corpus used). The use of two corpora ensured that our approach and classifiers were not over-fitted to any specific corpus data, instead confirming that the method we proposed can be generalized and the results obtained are replicable given any data set.

We extracted the information using Regular Expressions to identify occurrences of nouns in marked and unmarked contexts. For marked contexts, we used the linguistic patterns defined in Section 4.1. The relative frequency of each particular noun seen with a marked context was stored in an n-dimensional vector that corresponds to the number of features used. The occurrences of a noun in unmarked contexts were encoded in the same vectors following the strategy outlined above (see Equation 4.1). Once all of the information was compiled, the vectors were provided to classifiers.

For classification, due to the limitations discussed in Section 4.1 and the error analysis conducted, we selected the Logistic Model Trees (LMT) [Landwehr et al., 2005] Decision Tree (DT) for classification in this Section. We used the LMT classifier in WEKA [Witten and Frank, 2005] implementation in a 10-fold cross-validation setting for evaluation.

The LMT classifier essentially builds "logistic model trees", which are classification trees with logistic regression functions at the leaves, instead of selecting only the most informative attributes for a tree structure. Thus, it partitions the feature space into classes of observations to assemble into a tree. Because the C4.5 algorithm selects for only the minimum number of attributes, many informative attributes are not taken into consideration for classification, thus, in many cases, sufficient information is not provided to the classifier, resulting in inaccurate classification decisions.

Furthermore, the LMT has been shown to better handle *information sparse* vectors that are marked with a high amount of "zero" values, such as ours, by relying on simple regression models if only little and/or noisy data is available. It also adds a more complex tree structure if there is enough information to warrant such a structure, thus arriving at even more accurate class membership decisions when there is robust information available. Thus, the LMT classifier produces a tree that contains linear regression functions at the leaves. Each function represents the weight that a given cue contributes toward classification. Moreover, based on logistic regression, this type of model makes no assumption regarding the normality of the distribution of its variables. In our case, this is important given the Zipfian distribution of corpus data [Zipf, 1935], which is an important feature considering the type of data that we are working with.

We conducted a binary classification for each semantic class considered with the word representations that also included information regarding unmarked contexts. For a fair comparison, we obtained what we consider to be the baseline results from word representations that only include information regarding marked contexts with the LMT classifier over CRAWL30M. This baseline allows us to directly compare and assess the impact of unmarked contexts in nominal lexical semantic classification.

4.2.3 Results

Tables 4.5 and 4.6 present the results obtained in our experiments in terms of Precision (P), Recall (R) and $F1$-Score ($F1$). The overall accuracy of all classifiers for each experiment is also provided. The baseline classifiers achieve an average accuracy of 70.84%. By including unmarked contexts in the vectors provided to the classifiers, the average accuracy of the classifiers rises to 75.16%, representing an error reduction of 4.32 points. We tested the statistical significance ($p < 0.1$) of this increase in the accuracy of classification and, for all classes except for HUM, the increase in accuracy between the baseline results and those obtained when including unmarked contexts is significant.

Class	original marked contexts			original marked + unmarked contexts			marked contexts			marked + unmarked contexts		
	P	R	$F1$	P	R	$F1$	P	R	$F1$	P	R	$F1$
ORG	0.64	0.62	0.60	0.70	0.68	0.68	0.76	0.74	0.74	0.75	0.74	0.74
LOC	0.72	0.70	0.70	0.73	0.73	0.73	0.70	0.70	0.70	0.77	0.79	0.77
EVT	0.70	0.68	0.67	0.74	0.73	0.72	0.73	0.72	0.64	0.73	0.72	0.69
COM	0.67	0.66	0.65	0.74	0.73	0.73	0.71	0.70	0.69	0.71	0.71	0.71
HUM	0.86	0.84	0.86	0.87	0.86	0.86	0.87	0.87	0.87	0.85	0.84	0.84
Acc.	70.84%			75.16%			75.05%			76.35%		

Table 4.5: Precision (P), Recall (R), and $F1$-Score ($F1$) of classifiers over CRAWL30M

Knowing that one of the potential downsides of using unmarked contexts in classification tasks is an increase in noise, which will be elaborated upon later in this Section, we conducted an error analysis of the results obtained using the original marked contexts. This analysis made apparent that most of the noise was due to imprecise information extracted with our regular expressions, leading us to revise them as a result of this observation. In this process, there was no definition of new marked contexts. This revision resulted in the 4 different experiments of which have results presented in Table 4.5. The revisions resulted in more accurate and better defined regular expressions of our marked contexts.

As indicated by the results, these revisions in combination with the unmarked contexts further raised the average accuracy of the classifiers to 76.35% (see Table 4.5), representing an error reduction of 5.51 points with regard to the baseline. Having obtained these promising results over the data in the corpus used to develop our approach (CRAWL30M), it was crucial to also verify the replicability of our method using a different and completely independent corpus, as described above. Moreover, replicating the original experiments over a different corpus was also important to assure that the revisions made to the regular expressions did not result in any over-fitting between the extraction of distributional information and the corpus being used. We also note that the UKWAC60M corpus is 50% larger than the corpus used to identify the unmarked contexts. The results obtained using this corpus data are presented in Table 4.6.

Class	marked contexts			marked + unmarked contexts		
	P	R	$F1$	P	R	$F1$
ORG	0.72	0.69	0.69	0.76	0.76	0.76
LOC	0.74	0.71	0.71	0.75	0.75	0.75
EVT	0.68	0.67	0.67	0.73	0.73	0.73
COM	0.69	0.69	0.68	0.70	0.70	0.70
HUM	0.86	0.86	0.86	0.84	0.84	0.84
Acc.	72.69%			76.03%		

Table 4.6: Precision (P), Recall (R), and $F1$-Score ($F1$) of classifiers over UKWAC60M

The classifiers that included unmarked contexts yielded an average accuracy of 76.03% over UKWAC60M, representing an error reduction of 3.34 points with regard to the classifier including only marked contexts (using the revised version of the regular expressions used to extract the lexical-syntactic patterns from Section 4.1), which is a statistically significant improvement ($p < 0.1$). Moreover, these results represented an improvement of accuracy by 5.19 points with regard to the baseline. This demonstrates, on the one hand, that the identification of relevant contexts based on CRAWL30M data did not result in an over-fitted approach; and, on the other hand, that the method presented here is robust, as we used our classifiers with a completely different corpus and still yielded comparable results. Because they were conducted on a corpus that was previously unseen, the results demonstrated the viability, as well as the transferability of our method, below we detail only the results obtained on UKWAC60M data, as these results are independent of all the preliminary studies conducted and thus demonstrated the potential applicability of our approach to any corpus.

Class	marked contexts				marked + unmarked contexts			
	members		non-members		members		non-members	
	P	R	P	R	P	R	P	R
ORG	0.79	0.52	0.65	0.86	0.78	0.72	0.75	0.80
LOC	0.82	0.55	0.66	0.73	0.78	0.70	0.73	0.80
EVT	0.73	0.57	0.63	0.78	0.74	0.72	0.72	0.73
COM	0.72	0.62	0.66	0.75	0.72	0.65	0.68	0.74
HUM	0.87	0.84	0.84	0.87	0.86	0.82	0.82	0.86

Table 4.7: Precision (P) and Recall (R) of classification of members and non-members of different lexical classes over UKWAC60M

Table 4.7 presents the precision and the recall of each individual classifier over UKWAC60M both with regard to the members of a given class, and those nouns that are not members of that class. This table allows us to identify more precisely gain insight regarding the contribution of unmarked contexts to the error reduction in classification. According to our results, unmarked contexts allow us to gain an average of 10.2 points in recall for class members, demonstrating that they provide useful information to classifiers, which allows them to cover cases which marked contexts alone were not able, most likely due to data sparsity. However, the impact on precision varies between classes, as the inclusion of very frequent information in the vectors representing target nouns may provide additional noise to the classifier.

The precision of classification of class members decreases slightly with the inclusion of unmarked contexts, although the differences are not statistically significant ($p < 0.1$). However, the precision of the classification of nouns not belonging to the classes considered significantly increases ($p < 0.1$) with the inclusion of unmarked contexts in all cases except for the HUM class. This shows that although unmarked contexts do not contribute to a better definition of the characteristics of individual classes (see Table 4.7), they do allow for a cleaner discrimination of members and non-members of a class, contributing to a better partition of the classification space.

Class	marked contexts		marked + unmarked contexts	
	FN (%)	FP (%)	FN (%)	FP (%)
ORG	23.32	6.71	13.43	9.98
LOC	22.30	5.75	14.74	9.71
EVT	21.91	10.42	13.82	12.97
COM	18.94	12.00	17.26	12.63
HUM	7.79	6.01	8.90	6.45

Table 4.8: Percentage of False Negatives (FN) and False Positives (FP) in classifiers over UKWAC60M with and without unmarked contexts

Table 4.8 presents the percentage of False Positives (FP), i.e. nouns incorrectly marked as members of the class, and False Negatives (FN), i.e. nouns incorrectly marked as not belonging to a class, in the results of each classifier both with and without the inclusion of unmarked contexts. Again, for each of the classes, except HUM, the inclusion of unmarked contexts decreases the percentage of FN, mirroring a reduction in sparsity and further indicating an increase of the amount of relevant information. Yet, there was an increase of FP across all classes, which signified an increase of the noise provided to the classifier, discussed in detail in the next Section.

4.2.4 Discussion

In Section 4.2.3, we presented the results obtained in our experiments using distributional information regarding both marked and unmarked contexts for the classification of English nouns. Each unmarked context was selected based on two criteria:

- Indicative of a grammatical mark where nouns of any class can co-occur in context;
- High amount of occurrences in corpus data with nouns of any class.

Based on these two criteria, we objectively selected the contexts that we defined as unmarked contexts. Along this line, we were rigorous in selecting contexts that met both criteria; especially as our goal was to effectively increase the information included in the word representation. This way, we did make sure that the contexts selected were frequent; yet, we did not discount the possibility that there are other unmarked contexts that exists. However, we leave their further identification and investigation to future work.

Overall, our results show that unmarked contexts either improve accuracy or do not affect classification results. Specifically, the improvements in accuracy are particularly significant for those classes for which there were difficulties to find enough occurrences in marked contexts in previous experiments, i.e. those classes with a higher level of FN when classified without using unmarked contexts.

Table 4.9 presents data regarding the range of absolute frequency of occurrence in corpus data of nouns and how the average absolute frequency of each word compares to the average absolute frequency of occurrence in marked contexts and in unmarked contexts per class. Although, each class contains words with a minimum frequency of 1, the range of maximum values per class differs. The COM noun *information*, for instance, is the most frequently occurring nouns for that class with an average absolute frequency of $55,218$ in the corpus ($65,899$ occurrences with marked contexts and $89,254$ occurrences with unmarked contexts), while the HUM noun *author*, the most frequently occurring noun for that class, occurs with an absolute frequency of $5,542$ in corpus data (5635 occurrences with marked contexts and 9917 occurrences with unmarked contexts).

Unmarked contexts occurrences analysis							
Class	Number of In-class Nouns	Ave. absolute frequency of occurrences of nouns in corpus data	Av. absolute frequency of occurrence of nouns in marked contexts	Av. absolute frequency of occurrence of nouns in unmarked contexts	Minimum absolute frequency of class nouns	Maximum absolute frequency of class nouns	Most Frequent Unmarked Contexts
COM	240	$1,484$	746	$2,532$	1 (*newscast*, *playbill*)	$55,218$ (*information*)	1. headed with definite article 2. modified by JJ
EVT	240	675	186	$2,544$	1 (*flashing*, *pileup*)	$15,932$ (*experience*)	1. occurs as singular 2. headed with definite article
HUM	227	451	285	864	1 (*collegiate*, *defeatist*)	5542 (*author*)	1. header of an "of" PP 2. Modified by a nominal
LOC	138	562	50	$1,089$	1 (*crawlspace*)	9708 (*property*)	1. headed by an "of" PP phrase 2. selected for by the verb "to have"
ORG	125	$1,283$	167	$2,453$	1 (*rabbinate*, *matriarchy*)	$19,831$ (*company*)	1. modified by a,nominal 2. Singular 3. selected for by the verb "to get"

Table 4.9: Average absolute frequency and range of noun occurrences per class in marked and in unmarked contexts

Theses observations further indicate that the target nouns consistently occur with a high frequency with unmarked contexts than with marked contexts. Thus they tend to provide more available information that results in being useful, especially in the cases where marked contexts do not provide an explicitly clear representation to the classifier, as we will detail below. Yet, as we also mentioned, there are some words that occur with an absolute frequency of only 1 in corpus data, such as the EVT noun *flashing*. This noun, for instance, does not occur in any marked contexts, yet it occurs 4 times in unmarked contexts, further confirming the availability of unmarked contexts, especially in the cases where a noun occurs with extremely low frequency in corpus data. In these cases, the unmarked contexts provide us with the distributional information necessary for a class membership decision.

In general, the results confirm our hypothesis that the distribution of words in unmarked contexts, when considered along with linguistically-motivated class-indicative marked contexts, provide useful information to improve classifiers, particularly when not enough class-specific information is available. In the subsequent discussion, we further make apparent the main advantages of our strategy.

A trade-off between silence and noise

An important result of our experiments is the overall reduction in the negative effect of sparsity caused by silence, or the low frequency of particular representative contexts that are needed to produce an accurate classification, which decreased by an average of 5.21% (see the difference in terms of FN in Table 4.8). This is attributed to an increase in accuracy (see Table 4.6): as more information is supplied to the classifier, the additional information permits more accurate membership decisions. To illustrate this, we consider examples from the COM, ORG and EVT classes, for which there was not enough information for classification when unmarked contexts were not considered. The inclusion of unmarked contexts provided information resulting in correct classifications.

The COM noun *theorem theorem* occurred 118 times in the corpus, though only 8 times in one marked context, most specifically in a PP headed by the preposition “to”, which was not enough to accurately classify it as a member of the COM class. As this noun does occur in class-marked contexts, the information provided is not sufficient for the classifier to make an accurate prediction regarding its class membership. Thus, the lack of enough information provided to the classifier is responsible for its misclassification. However, after the inclusion of information regarding the behavior of this noun in unmarked contexts, the classifier was able to accurately decide for its inclusion as a member of the COM class.

This was also observed in the case of the ORG noun *secretariat* and the EVT noun *impulse*, which occurred 190 and 154 times, respectively, in the corpus, yet only 8 and 12 times in marked contexts, including a PP complement headed with the preposition “for” and in a PP headed with the preposition “by” without an article. Yet, again, this was not enough for an accurate classification and the subsequent inclusion of information regarding the distribution of these nouns in unmarked contexts allowed for their correct classification.

One of the main concerns regarding the use of unmarked contexts was the introduction of extra noise as a side effect, and the way this affects classification results. The impact of noise is further made apparent by the amount of FP observed in classification results, see Table 4.8. In our experiment, we did identify some cases of nouns correctly ruled out as members of a class when using only marked contexts, which were incorrectly classified as class members after the inclusion of unmarked contexts. The slight increase of FP in our results (see Table 4.8) shows this method uses an approximation in order to represent the distribution, which at times, as indicated above, can fail to provide a clear-cut distinction to the classifier. However, in the overall results, this limitation is compensated by the larger amount of nouns that were correctly classified after the inclusion of unmarked distributional information (see Tables 4.5 and 4.6).

Analyzing the additional FP observed, we identify two different cases:

1. nouns correctly classified using only marked contexts as not belonging to a class based on a borderline probability, which were incorrectly classified as members of that class when unmarked contexts were also considered, again based on a borderline probability;

2. nouns correctly classified as not belonging to a class as they hardly or never occurred in class-marked contexts, but whose behavior in unmarked contexts was similar to that of members of the class being classified, thus providing contradictory information to the classifier and resulting in incorrect classification.

The first case is illustrated by a noun like *biography*, which occurs 598 times in marked contexts and was correctly predicted not to be a member of the LOC class with a borderline probability score (0.47) when just included marked contexts in its representation. The inclusion of unmarked contexts provided information to the classifier, which slightly changed this probability (0.56), and resulted in an incorrect classification. The noun *megalopolis* illustrates the other case. Occurring only 3 times in class-marked contexts of the COM class, this LOC noun had been correctly classified as not belonging to the COM class. However, its behavior in unmarked contexts showed more similarities with members of the COM class than with non-members, which resulted in its incorrect classification.

Due to the general low frequency of many of the words used, we observed a bias toward the second case, in which the unmarked contexts were used to “fill in the blanks" where the missing values in the word representations of low-frequent words were not sufficient for the classifier to make an accurate decision. This was noted by the consistently low probabilities reported for these nouns. For the purpose of this work, we consider low frequency to be an issue for words that occur generally with an absolute frequent of less than 100, which actually affects 54.78% of our data set. In the UKWAC60M corpus this signifies that the total absolute frequency of each of those nouns corresponds to only 0.00000167% of corpus data.

Illustrating two paradigmatic cases of noise in the results of the classifiers, these examples make apparent how unmarked contexts are sometimes responsible for incorrect class membership decisions, and how further improving their use in classification tasks, particularly in the case of *borderline* classification decisions, remains a promising line of research to explore in the future.

More robust classification decisions

Besides the reduction of the impact of the lack of information in the features vectors, that resulted in the consequent improvements in accuracy, as discussed in the previous section, we also noticed that the introduction of unmarked contexts provided additional information regarding the distribution of nouns that were classified by chance (i.e. correctly classified nouns, with a borderline probability score), resulting in more robust classification decisions.

We saw this with the EVT noun *consolidation* and the LOC noun *coalfield*, for instance. Each of these nouns was correctly classified using only marked contexts, yet with borderline probability scores: 0.52 and 0.53, respectively. Upon providing information regarding unmarked contexts to the classifier, these nouns continued to be correctly classified but with much higher probability scores, and thus achieved more reliable classification decisions of 0.75 and 0.76, respectively.

These examples are considerably different from those discussed earlier in this Section, as they are far from being cases of sparsity. In fact, the EVT noun *consolidation* occurs with an absolute frequency of 312 times in the corpus and 317 times in marked contexts while the LOC noun *coalfield* occurs 52 times in the corpus and 53 times in marked contexts. We also acknowledge the difference between occurrences and marked contexts [4]

In both of the above cases, almost all of the occurrences in marked contexts were found to be with only one context, more specifically with morphological affixes. The high frequency of nouns in just one cue is attributed to their occurrence with a morphological marker, as explained in detail in Section 4.1. Morphological markers provide a large amount of information in the feature vectors because they correspond to an affix of the target noun, thus the occurrence of this cue is directly related to the number of times that the noun is seen in corpus data. However, although this context provides a large amount of indicative information for very frequent nouns, this is not the case for nouns that occur with low frequency in corpus data. Moreover, some nouns are polysemous, and thus, their class membership can change based on the selectional preference of the context, which would render null the effectiveness of this cue if the lexicalized context is not indicative of the target class. In this way, we must rely also on the context of its occurrences for the classifier to make an accurate decision.

Furthermore, the use of only morphological information can result in few correlations between the evidence available due to the fact that occurrence with only one marked context was observed, which causing a lower probability score. This is problematic because morphological cues are the most frequently occurring indicative cues, as they correspond directly to *each* occurrence of a word. Thus, they are extremely indicative for highly inflectional languages; however, in less or moderately inflectional languages, such as English, where there are not nominal inflectional suffixes for all words or for all classes, this type of context is neither sufficient nor available for all words.

The results obtained in the experiments using unmarked contexts also demonstrate that classification results are unevenly affected by unmarked contexts. As made apparent by the results, the contribution of unmarked contexts to the classification of different semantic classes is not always the same. For example, we observed that classes that demonstrated more disperse linguistic behavior of their members, such as the ORG, LOC or EVT classes, improve more with the inclusion of unmarked distributional information than classes with a more homogeneous distributional behavior, such as the HUM class.

[4] Although there is a finite number of occurrences of a word in corpus data, there is not a finite number of times that a noun can occur with a marked context or unmarked contexts. Due to the nature of the patterns, different patterns may capture one occurrence of the word in a variety of ways. To illustrate this concept, a noun can be captured in a pattern of a target noun modified by an adjective and again in that same context it can be captured if it also contains an indicative affix. It can even be captured a third time in that same context if the NP is the subject, for instance.

To make our statement clearer, we claim that some nominal classes are composed of nouns that tend to occur in a wider range of contexts, thus displaying a more heterogeneous and disperse distributional behavior. This heterogeneity is made apparent by an analysis of the overall distribution of the marked contexts between the members of each lexical semantic class. In contrast with more heterogeneously behaving noun classes, other classes are composed of members that display a more homogeneous collective behavior that can be more easily captured by distributional approaches.

Analyzing the distribution of cues between class members in UKWAC60M, we identified, in each class, a set of cues that occurred with the majority of nouns of the class, and which we will consider to represent the core linguistic behavior of each specific class. We also observed the amount of cues included in this set differed considerably from class to class (see Figure 4.1). Thus, the larger the amount of marked contexts shared by the majority of the members of a class, the more homogeneous we can claim their behavior to be. In the specific case of the classes considered in this section, 30.7% of the cues for the HUM class are shared by the majority of HUM nouns, while 26.6%, 13.3%, 9.5% and 9.1% of the cues for the COM, ORG, EVT and LOC classes, respectively, are shared by the majority of the nouns of each class, respectively as represented in Figure 4.1.

An effect of a class collectively having a more heterogeneous linguistic behavior is that the evidence regarding each of its marks will typically be more disperse and, as a result, often not strong enough to be considered by classifiers, which explains the improvement introduced by unmarked contexts. In contrast, classes like HUM, that are composed of nouns that generally occur in a common set of prototypical contexts of that class can, on the one hand, identify contexts that mirror the prototypical behavior of that class more straightforwardly and, on the other hand, the class members almost always show enough occurrences in such contexts to be merit an accurate classification decision. There are also strong marks based on suffixes and degree of grammaticalization for the HUM class (as demonstrated in Section 4.1), which can be more readily captured. For instance, on the one hand, suffixes, such as: *-er* and *-or* are indicative of many HUM type nouns (e.g. *doctor*, *painter*, *officer*, etc.) while the preposition *during*, when preceding a nominal phrase, is very indicative of occurrences of EVT nouns.

These examples provide instances of features that can be easily identified for inclusion in a feature vector, readily providing a large amount of class-indicative information. On the other hand, there are other types of features that although indicative, result in a much sparser feature vector because of their reliance of

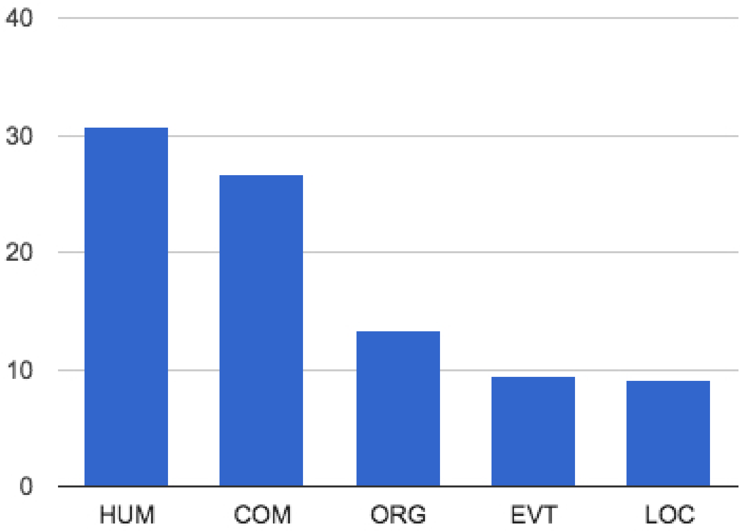

Figure 4.1: Percentage of cues occurring with the majority of class members, per class

occurrence within corpus data. For instance the occurrence as the subject of an agentive verb, which is considered an indicative feature for the ORG class, does not necessarily occur readily with all members of the class, thus making marked contexts that provide a homogeneous representation of the class more difficult to capture.

In this way, when the more readily available and frequent marked contexts occur with members of a class, the inclusion of extra contexts (e.g. unmarked contexts) are rendered ineffective, as class membership decisions are already accurately made to a great extent (in our case 86.19% of the times) based on the information provided by marked contexts. This is consistent with the stability of the results reported for the HUM class in the different experiments performed, which did not demonstrate any significant changes with the inclusion of unmarked contexts.

4.2.5 Final remarks

Our main goal in this Section was to evaluate whether unmarked contexts improved accuracy in our lexical semantic classification task by reducing sparsity in vectors. Departing from the hypothesis that these contexts can provide additional information to classifiers when there is not enough distinctive co-occurrence information available, the results demonstrated that the use of unmarked contexts, which are typically discarded as non-discriminatory, can significantly improve the results of lexical semantic classification when considered along with marked contexts. Our results show that by using both types of distributional information (i.e. marked and unmarked), we reduced the sparse data problem and subsequently improved classification, and indicated by the increase in classification accuracy observed in Tables 4.5 and 4.6).

Yet, we also considered whether the combination of shallow distributional models will provide the extra information necessary to make classification decisions, where one or both of the individual models does not provide sufficient information to the classifier. Thus, we also considered the combination of linguistic information with linear "bag-of-words"-type features in an attempt to further explore the potential of models that use surface information. In combining the features from two models, as proposed in this Section, we can further determine whether the distributional information of one model can be compensated with the distributional information of the other, especially in the case that the information provided by one of the models is insufficient for classification. In the following Section, we further explore and discuss these combinatory strategies).

4.3 Comparing lexical semantic classification models

Following the results obtained in Section 4.2, our next logical step was therefore to work on further reducing sparsity without introducing noise. To do this, we compared our model built with linguistically-motivated class-indicative cues, as described in Section 4.1, with other distributional models, including more sophisticated and also simpler methodologies. The resulting analysis provides insight to how the different features used to build these distributional models that represent various levels of generalization (i.e. with contrasting levels of complexity in terms of linguistic information, ranging from pre-defined tuples to simple linear token information) can affect classification decisions.

Furthermore, the analysis of results from the classification decisions obtained with each model also provides information regarding the origin of the obstacles that have been identified to affect this task, as well as solutions to overcome them. Along this line, we conducted an empirical study of the classifications that each of these different distributional representations produce, which allowed us to determine the effects (both advantages and disadvantages) of considering one type of information over the other.

As discussed in detail in Chapter 2, distributional models can vary greatly by exploiting different representations of features. In the experiments presented in this Section, we studied a structured distributional semantic resource, an unstructured linear model and a linguistic model, as described below.

4.3.1 Distributional (semantic) models

The goal of the experiments presented in this Section is to empirically evaluate the performance of different distributional models in a nominal lexical semantic classification task, departing from the experiments explained in previous Sections. We studied three models that exploit different types of distributional features, thereby providing different representations of nominal behavior in context:

- The structured Distributional Memory model (henceforth DM: [Baroni and Lenci, 2010]), introduced in Chapter 2, is a generalized framework for distributional semantics that uses word-link-word tuples from a dependency parse of a corpus as features. This is the only model that we have used, thus far[5], that incorporates the syntactic information provided by a dependency parser.

  We note that the DM consists of three different variations that are each representative of different levels of lexicalization. For instance, the LEX$_{\text{DM}}$ variation is the most heavily lexicalized of the three and considers each token; while the DEP$_{\text{DM}}$ variation has a minimum degree of lexicalization, basing itself on the dependency paths between words. Finally, the TYPE$_{\text{DM}}$ model represents a sort of middle level in regards to lexicalization. Based on the idea motivated by [Baroni et al., 2010] that what matters is not the frequency of the link between two words, but the variety of the surface forms that express the link, this variation represents the types of contextual realizations, not the tokens. Furthermore, TYPE$_{\text{DM}}$ model is representative of the type level of generalization that we wish to achieve to classify nouns into given lexical semantic classes. For these reasons, the work using DM presented in this thesis is based on the TYPE$_{\text{DM}}$ model.

[5] In Section 4.4, we study Word Embeddings WE representations that also use information from a syntactic dependency parser.

- The linear model (henceforth: LINE), built by extracting tokens in context windows of a target noun, is based on a bag-of-words-type model [Bullinaria and Levy, 2012]. In this model, features consist of tokens extracted from a standard 5-word context window [Evert, 2008], to the right and to the left of each target word.

- We also continue our study using our linguistically-motivated model (henceforth: LING), which we built using the linguistically-motivated class-indicative features of a given lexical semantic class, as described in Section 4.1.

The description of features that comprise each model further indicates the differences of distributional information used to build the resulting word representations. Table 4.3.1 presents examples of features from each of the models considered. Furthermore, in Table 4.10 we can directly observe how the different levels of generalization of feature information affects the number of cues required by each model.

	Targets	$type_{\text{DM}}$	LING	LINE
COM	208	$775,747$	16	$27,095$
EVT	211	$687,019$	20	$27,086$
HUM	208	$656,023$	17	$27,078$
LOC	114	$572,191$	22	$27,073$
ORG	111	$535,675$	16	$27,042$

Table 4.10: Number of target nouns per class and number of features per class for each model considered

	$type_{\text{DM}}$	LING	LINE
accident-N	sub-int-happen-V	x-NN when-WRB	car
	sub-int-occur-V	until-IN the-DT x-NN	injury
	obj-cause-V	since-IN the-DT x-NN	road

Table 4.11: Example of features used for each model for the EVENT noun *accident*. In DM, features represent the syntactic position of a target noun as a combination of dependency (*sub-int*) or its dependent head (*happen-V*); in LING, features represent linguistically-motivated class indicative lexico-syntactic contexts, such as a target noun (*x-NN*) preceding a specific adverb (*when-WRB*); in LINE, features represent simple co-occurring words in a 5-word context window

Finally, and as briefly mentioned in Section 4.2, in this Section we further explored the potential of models that use surface information through the combination of features from certain different models studied. More specifically, from the LING model and from the LINE model, which resulted in a fourth model (henceforth: LINGLINE) that uses both linguistically-motivated information as well as linear context as features. In combining the features from these two models, we further confirmed that the distributional information of one model can be compensated with the distributional information of the other, especially in the case that one of the models provides insufficient data for classification.

In line with the unmarked context approach proposed in Section 4.2, the combinatory LINGLINE model, on the one hand, the LINGLINE model combines the distributional information extracted from the carefully-constructed linguistic patterns defined in Section 4.1 and the distributional information from pre-defined context windows, following the bag-of-words-type models. On the other hand, our unmarked contexts strategy combines the same linguistically-motivated class-indicative information with the unmarked context information encoded from the deviation of the occurrence of a target noun with that context from the average occurrence of all nouns with that same unmarked context. The comparison of these models indicates whether harnessing data from more than one model in a robust and informative way, can benefit classifiers, especially to overcome the problem with sparse data.

4.3.2 Experiments

Each of the aforementioned models was trained on two different corpora: the 90 million token BNC90M corpus and the 3 billion token LARGE3BN corpus (Section 3.2 provides a detailed description of each corpus). We conduct these experiments specifically on these two corpora because of their difference in size. In this way, the results obtained provide empirical evidence that determines whether corpus size has an affect on classification decisions for any specific model.

Although the same corpora was used to extract each model, the DM model was the only model that also incorporated information from a full syntactic annotation (tokens, PoS tags and syntactic dependency information) into its features. Each feature for the DM mode consists of a type with its links generalized as patterns inside the tuple. Each feature was extracted from corpus data using the $type_{\mathrm{DM}}$ methodology to extract tuples, as defined by [Baroni and Lenci, 2010].

The LINE model uses only lexicalized tokens as features. To extract the features for this model, all PoS tags and punctuation were removed from the corpus data and all tokens of at least 3 characters were extracted from a 5-word context window to the right and to the left of each target word defined in our data sets.

Finally, the LING model uses the linguistically-motivated class-indicative lexical-syntactic patterns, following the method to extract this marked context information defined in Sections 4.1 and 4.2. The LING model required tokens and corresponding PoS tags for feature extraction. Each lexico-syntactic pattern was formalized in a Regular Expression to extract information regarding the occurrences of nouns with each context. Table 4.10 provides the final number of features considered per model and class.

All of the extracted feature information was used to build a word representation for each class with each model by populating an n-dimensional vector for each noun with the positive Local Mutual Information (pLMI: [Evert, 2008]) for each feature. In the experiments presented in this Section, we use positive Local Mutual Information to weight occurrence values because it is an approximation of the log-likelihood ratio measure that has been shown to be a very effective weighting scheme, especially in the case of sparse frequency counts (see [Baroni and Lenci, 2010] for more details), which have been negatively affecting our vectors in all of the work presented thus far.

The pLMI was calculated using the DISSECT toolkit [Dinu et al., 2013]. Following standard practice [Bullinaria and Levy, 2007], all negative weights were raised to 0 and the information for each class and model was compiled into a sparse matrix, consisting of four elements: target word, feature, weight and class membership information that was provided to the classifier for classification.

Target nouns for each class were all obtained from the data sets described in Section 3.1. We only considered those nouns that occurred both in BNC90M and LARGE3BN. To ensure a direct comparison between the results obtained on the small and large corpus data, we only classified nouns that occurred in both corpora. As with the previous experiments, the data sets were also balanced with respect to class members and elements not belonging to the class (see targets in Table 4.10 for the final distribution of target nouns, which presents the number of class members, each appearing n times in both corpora). A binary classification was conducted for each semantic class in each model studied.

In this Section, each binary classification experiment was performed with a CART: Classification and Regression Trees algorithm [Breiman et al., 1984] in the Sci-kit learn [Pedregosa et al., 2011] implementation. This classification algorithm was selected for experiments in this Section due to the need to construct a more powerful classifier that can handle high-dimensional vectors. From a technical perspective, we encountered computational obstacles when handling large data sets in the WEKA implementations of the classifiers that we have used thus far. Thus, we needed to build a classifier that can efficiently handle our data sets, mainly due to the inclusion of the DM model, which yielded hundreds of thousands of features. The Sci-kit learn toolkit provided the necessary framework to build a classifier that fit our needs. Within the classifiers available, the CART classifier most closely resembled the classifiers that we have previously experimented with, hence we selected it for the work presented in this Section.

The CART algorithm is very similar to the C4.5 DT, but differs in the fact that it constructs binary trees using the feature and threshold that yield the largest information gain at each node. Maximizing the information gain at each node refers to the idea that the algorithm needs to choose a split among all those possible at each node so that the resulting child nodes are the "purest". Thus, it minimizes the uncertainty of that particular split as the best selection for a given attribute. For our classification tasks, we understand Information Gain to define the preferred selection or sequence of features required to most rapidly and efficiently arrives at an accurate classification decision. The mathematical definition of Information Gain is illustrated in Equation 4.2, where S is an number of training examples, $Entropy(S)$ measures the impurity of S and A is an attribute [Mitchell, 1997].

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|Sv|}{|S|} EntropySv \tag{4.2}$$

4.3.3 Results

The results in Table 4.12 were obtained using the 3 billion token LARGE3BN corpus. The results show that overall the $F1$-Score of each model demonstrates a statistically significant improvement ($p < 0.05$) over a random baseline when large amounts of data are considered. In regards to the performance of the individual models, we observed that TYPE$_{\text{DM}}$ obtained the highest overall results, with its $F1$-Score demonstrating a statistically significant improvement ($p < 0.05$) over the $F1$-Score of both LING and LINE. We attribute this to the inclusion of syntactic information provided by a dependency parse in the model, which is one of the main differences between the TYPE$_{\text{DM}}$ model and the LINE model. We reflect on this point in more detail in Section 4.3.4.

	$type_{\text{DM}}$			LING			LINE			LINGLINE		
	P	R	$F1$	P	R	$F1$	P	R	$F1$	P	R	$F1$
COM	0.87	0.88	0.88	0.68	0.67	0.67	0.79	0.78	0.78	0.81	0.80	0.80
EVT	0.85	0.81	0.83	0.81	0.79	0.79	0.83	0.85	0.84	0.80	0.85	0.81
HUM	0.92	0.91	0.91	0.88	0.90	0.88	0.76	0.78	0.76	0.89	0.84	0.86
LOC	0.83	0.81	0.81	0.73	0.77	0.74	0.80	0.77	0.78	0.84	0.84	0.83
ORG	0.84	0.82	0.83	0.72	0.74	0.72	0.72	0.76	0.73	0.79	0.77	0.77
MacroAvg	0.86	0.84	0.84	0.76	0.77	0.76	0.78	0.78	0.77	0.82	0.82	0.81

Table 4.12: Precision (P), Recall (R), and $F1$-Score of classification using each model of each class with a 3 billion token corpus (LARGE3BN)

Interestingly, the LING and the LINE models, which both consider shallower features, achieve an $F1$-Score of 0.76 and 0.77, respectively, can already be considered successful for use in NLP tasks. However, there is no statistical significance between the $F1$-Scores of theses models, although there is a slight difference in their recall and precision, especially when considering individual classes. This implies that each model has different advantages in regards to the lexical semantic classification of nouns, which we further investigate in our Error Analysis in Section 4.3.3.

With respect to the individual classes, we observed that HUM and ORG classes obtained stronger classification from the LING model while the COM, EVT and LOC classes obtain stronger results with the LINE model, indicating that the distributional model selected for classification should consider the indicative properties of the class being classified, as demonstrated in Section 4.1. This result further confirms one of our conclusions from Section 4.1, mainly that not all classes are equally identifiable with specific surface cues, due to more heterogeneous occurrence behavior in corpus data. For instance, the LING model benefits classes, such as ORG and HUM, that have readily identifiable class-specific features, such as morphological or grammatical marks while the LINE model benefits classes in which the features considered to be indicative of a class in linguistically-motivated models may fail to handle the heterogeneity of members as they occur in actual language use. For these types of classes, the information provided by the LING model may be too disperse in feature vectors to be accurately captured by classifiers, while the linear features that are used in the LINE model are much more numerous and thus contribute to a larger internal variation in the vectors.

Furthermore, in Table 4.12, we also observed that the combined LINGLINE model demonstrates a statistically significant improvement ($p < 0.05$) over both the LING and the LINE models, respectively. As there is no statistical difference between the LING and the LINE models, individually, these results confirm that there is a benefit to simultaneously use the features of both models; underlining the compensatory effect of using information provided by the combination of features from LING and LINE. For instance, LING includes indicative yet potentially sparse and/or noisy features while LINE includes a simply large amount of co-occurrence information. In this way, where the distributional information provided by the features of the LING model is not sufficient for the classifier to make a decision regarding class membership, the LINE model can provide extra information to the classifier to arrive at a generally more reliable decision and vice versa. However, we also acknowledge that the LINGLINE model still does not outperform TYPE$_{\text{DM}}$, which again emphasizes the added value provided by the richer syntactic information available in TYPE$_{\text{DM}}$. This result also confirms the success of more structured DSMs to identify paradigmatically similar words, which essentially form the basis for one of the basic criteria of semantic classification.

As previously stated, we used two corpora for the work presented in this Section to also study the effect of corpus size on classification decisions made with different distributional models. The results in Table 4.13 were obtained using the 90 million token BNC90M corpus, which is approximately 20 times smaller than the LARGE3BN corpus, to train each model.

	LING			LINE			LINGLINE		
	P	R	$F1$	P	R	$F1$	P	R	$F1$
COM	0.66	0.68	0.66	0.74	0.73	0.73	0.72	0.73	0.72
EVT	0.71	0.71	0.71	0.66	0.64	0.65	0.74	0.71	0.71
HUM	0.87	0.86	0.86	0.70	0.70	0.69	0.91	0.87	0.89
LOC	0.69	0.64	0.64	0.64	0.62	0.61	0.74	0.74	0.73
ORG	0.81	0.76	0.78	0.70	0.70	0.69	0.77	0.77	0.76
MacroAvg	0.74	0.73	0.73	0.68	0.67	0.67	0.77	0.76	0.76

Table 4.13: Precision (P), Recall (R), and $F1$-Score of classification using each model of each class with a 90 million token corpus (BNC90M)

Even when trained on smaller corpus data, the $F1$-Score of each of the models still demonstrates a statistically significant improvement ($p < 0.05$) over a random baseline. In regards to the performance of the individual models, we observed that LINGLINE, obtains the highest overall results, with its $F1$-Score demonstrating a statistically significant improvement ($p < 0.05$) over the $F1$-Score of both once more affirms the compensatory benefit of combining the features of both models, especially in the case where one model lacks sufficient information to make an accurate classification decision.

This result is further confirmed by the different of results between the LING and the LINE models, where the LING model obtained a statistically significant increase in $F1$-Score over the LINE model. Moreover, this result indicates that the reduction of corpus data negatively effects the LINE model and effectively reduces the amount of distributional data available, which hinders its ability to accurately predict class membership.

Error Analysis

We conducted an error analysis based on the confusion matrices that resulted from each classification experiment. In this way, we were able to identify that the bottleneck of each model is a function of its resulting False Positives (FP) and False Negatives (FN). Roughly speaking, we categorized FP to be interpreted as a consequence of "noisy" feature vectors, while FN were interpreted as a consequence of sparsity, or lack of evidence in the feature vectors, as previously described in Section 4.1. In what follows, we summarize the observations that can be drawn from the error patterns showed by each model in the different corpus settings.

LARGE3BN (2.83 billion tokens):

We first look at the types of features used by each model. As we previously described, the LING model consists of manually-identified linguistically-motivated features considered to be indicative of the semantic properties of a given lexical-semantic class. However, these features are not always exclusively indicative of one class, as their predictive power can also arise through correlations between a set of these features. Thus, there is a possibility that some of the features used by this model are noisy and, thus, can hinder the ability of the classifier to make an accurate decision.

For a further inspection, we constructed a confusion matrix that contains information regarding the semantic class to which a given FP belongs. The binary setting of the classification task did not allow for an analogous analysis to be conducted on FN. Table 4.14 presents the overall results of this analysis. We observed a large amount of EVT nouns to be classified as COM nouns and vice-versa. For example, the COM noun *reservation* was incorrectly classified as an EVT noun, while the EVT noun *discrepancy* was incorrectly classified as a COM noun. This trend was also observed with ORG and HUM nouns. For instance, we saw that a large part of the FP of ORG are members of the HUM class (such as: *comedian* and *graduate*) and a large part of FP of HUM are members of the ORG class (such as: *choir* and *regime*). A high amount of confusion between FP of the LOC and EVT classes was also observed.

	COM	EVT	HUM	LOC	ORG
COM	0	36	43	29	26
EVT	49	0	27	49	26
HUM	20	20	0	22	52
LOC	23	37	16	0	18
ORG	21	21	33	27	0

Table 4.14: Confusion matrix of FPs from the LARGE3BN corpus

On the one hand, we can again attribute these FP to the fact that HUM nouns, for instance, are explicitly marked, either grammatically or morphologically (i.e. suffixes such as *"-er"*, *"-or"*, *"-ir"* or the subject of psychological-type verbs), while ORG nouns can be considered collective HUM nouns, or a subset of this class (see Section 4.1 for a detailed discussion of this phenomenon).

On the other hand, these misclassifications are also related to very particular cases of lexical ambiguity. For instance, COM and EVT nouns, as well as LOC and EVT nouns, have been considered in literature as examples of regular polysemy [Pustejovsky, 1995], as discussed in Chapter 2, in which a lemma can be selected for in more than one sense. Under this assumption, some misclassifications can be caused by the fact that a lemma is also a member of another (potentially related) semantic class. It is important to note, however, that there is a systematicity in the misclassification of the nouns observed in the confusion matrix that is attributed to specific cases of lexical ambiguity. Nonetheless, a discussion of polysemy goes beyond the scope of this Section, although it is revisited as the focus of Chapter 5).

BNC90M (90 million tokens):

When training with smaller corpus data, we observed that the results of the LING model were consistent with results obtained with the 3 billion token LARGE3BN corpus data. The most significant difference, however, was observed in the results of the LINE model. Using a smaller corpus, we can directly see where the unstructured information of the LINE model is compensated with the linguistic information of the LING model. This further highlights the value of combining the distributional information from both models in the LINGLINE model(see differences between models in precision and recall in Table 4.13), especially in those cases where one model does not obtain sufficient distributional data for an accurate classification decision. Hence, our results indicated that the LINGLINE model effectively reduced the overall ratio of FP and FN that occur with each model, individually, and furthermore, that it resulted in more accurate classifications, as well as a broader coverage.

Finally, in regards to the semantic classes of the obtained FP, we observed trends similar to those discussed with regard to the results obtained with the larger corpus data. Along this line, we can say that although the amount of distributional information is reduced, the tendencies of the behavior of the nouns remained consistent.

4.3.4 Discussion

The work presented in this Section empirically evaluated the performances of different distributional models in a nominal lexical semantic classification task. Overall, the TYPE_{DM} model consistently obtains the strongest performance, demonstrated by a statistically significant difference between its results and the results obtained with the other models. On the one hand, this can be attributed to the inclusion of syntactic information provided by a dependency parse that can provide more structure to lexicalized features. Furthermore, it can filter additional noise incurred by extraction of the tuples of the DM model by reducing noise in the vectors provided to the classifier.

Because the TYPE_{DM} model is the only model that uses also the information syntactic dependency parser, we directly attributed the statistically significant increase in precision, recall and $F1$-Score to the reduction of noise that it provides due to the inclusion of this information. Moreover, we also attribute this result to the fact that the DM model a priori reduces very infrequent, potentially noisy occurrences, such as parsing errors, for instance, because its tuples contain some generalized link and semantic information, which can excluded noisy through generalization. Along this line, we consider that general structural information (e.g. syntactic parse patterns, copulative structures, position with relation to a verb link, attribute nouns, prepositional phrases, etc.) provided by the TYPE_{DM} model become indicative of a given lexico-semantic class. Thus, the results obtained by the DM model indicate that the quality of classification tasks increases with the inclusion of syntactic annotation.

In regards to the other distributional models that we considered, the results of the LING model further indicate its dependence on the availability of specific lexico-syntactic information in corpus data and the accuracy of a classifier to correlate the relations between a set of individual features that together are indicative of a given semantic class. However, as observed in Table 4.13, the results indicated that the LING model does not necessarily require a large amount of corpus data, in contrast to the LINE model. Hence, when available, the linguistically-motivated cues of the LING model do provide sufficient information to the classifier allowing it to make accurate class membership decisions.

As previously described, the results obtained with the LINE model indicated its dependence on the availability of a large amount of corpus data to ensure a sufficient amount of surface information. Consequently, when a large amount of data is not available, the LINE model looses its predictive capacity. Moreover, as the LINE model uses of many shallow features, there is a higher risk that many of those features are uninformative and thus provide no useful or indicative information to the classifier. Although we did not test the DM model on smaller corpus data, our intuition is that it would also behave like the LINE model with respect to data sparsity. In this light, these results confirm that sensitivity to data sparseness is a general problem of count-context models, independently of their being structured or not. Yet, an increase of features does not necessarily reduce sparse data nor does it decrease noise. A common solution to improve distributional models is to increase the numbers of features used in an attempt to capture more distributional information. A larger amount of features would actually cause further dispersion to the information available in the feature vector, further increasing the amount of zeros for both positive and negative features, which we have already confirmed that the classifier cannot distinguish between and is one of the causes of unreliable classification decisions.

Moreover, as seen in Section 4.1, not all lexical classes may be equally identifiable through surface features (see also the differences in the $F1$-Scores of each individual class in Table 4.12 and 4.13). In this way, the availability of contextual distributional information, such as that considered in LINE, can help to overcome the limitations assumed when manually-identifying linguistically-motivated class-indicative features, such as low frequency of target occurrences or simply a sheer lack of class-indicative marks. For these reasons, we can consider the LING model to be ideal when trained on smaller corpora while the LINE model can have more predictive power when trained on large corpus data.

Furthermore, we observed that this combination of information produced a compensatory effect in which each of the models provides information in the LINGLINE model that may be lacking when considering the distributional information provided by only one model, especially in the case of the LINE model when trained on smaller corpus data. Thus, we also considered that the combined LINGLINE model not only obtained State-of-the-Art results, but it does on different-sized corpora. As a large, robust, syntactic dependency-parsed (large) corpus is not always available for all languages, domains and/or tasks, the joint exploitation of linguistically-motivated cues and linear co-occurrence features, as demonstrated by LINGLINE, is also a viable alternative for classification.

Finally, we also compared the results of the LINGLINE model to the results obtained using also unmarked contexts, as described in 4.2. Table 4.15 provides a comparison of the results obtained from both experiments. The results firmly confirm the benefits of the compensatory effect of combining distributional information from more than one model. Furthermore, we observed that the results from each experiment are consistent, and remain consistent on different corpus data, which further validates the transferability of the combinatory strategy.

This is important to note, especially because one criticism of distributional approaches is that sufficient corpus data was not used for training and testing. These results further demonstrate that it is not necessarily the size of corpus data that affects the quality of the results obtained, but rather results are effected by the distributional representation used for classification.

Class	marked + unmarked contexts: UKWAC60M			LINGLINE: BNC90M			LINGLINE: LARGE3BN		
	P	R	$F1$	P	R	$F1$	P	R	$F1$
ORG	0.76	0.76	0.76	0.77	0.77	0.76	0.79	0.77	0.77
LOC	0.75	0.75	0.75	0.74	0.74	0.73	0.84	0.84	0.83
EVT	0.73	0.73	0.73	0.74	0.71	0.71	0.80	0.85	0.81
COM	0.70	0.70	0.70	0.72	0.73	0.72	0.81	0.80	0.80
HUM	0.84	0.84	0.84	0.91	0.87	0.89	0.89	0.84	0.86
Acc.	76.03%			80.39%			77.06%		

Table 4.15: Comparing the results from the LINGLINE models and the Unmarked Contexts approach

The results in Table 4.15 also demonstrate how the increase of information in the feature vectors can affect both the precision and the recall of the classifiers. This is true for all classes using the combined models when compared to models containing just one type of information. This is important to note because it further confirms our hypothesis that the effective inclusion of a larger part of previously excluded available corpus data, in addition to the information provided by a pre-defined models (such as the linguistic information of the LING model or the linear information of the LINE model) is crucial to the increase of accuracy of the classification decisions. With the useful inclusion of this extra distributional information, we are able to improve the acquisition of lexical-semantic class information from distributional data in such a way that it provides more sufficient information to the classifier, to further improve class membership decisions.

4.3.5 Final remarks

Overall, the work presented in this Section provides an empirical evaluation of classifiers using word representations produced with different distributional models. The results obtained underline the advantages and disadvantages of each model. Moreover, the results in this Section consistently indicated that the inclusion of syntactic information is an effective filter of noise. In this way, retrieving information that appears related not only uses the parser to provide more structured information to the classifier, but takes into account those features known to be both grammatically and lexically relevant. In this way, we can conclude that the quality of classification increases with the complexity of the syntactic information included in the features of distributional models.

Furthermore, the analysis conducted in this work resulted in a strategy that is capable of leveraging the bottlenecks of each model, by combining distributional linguistic and linear features, especially when large robust data is not available. Along with the results obtained in Section 4.2 using unmarked contexts, the results obtained with the LINGLINE model serve to increase the reliability of automatically constructed resources that require nominal lexical semantic class information.

One limitation of this work is the assumption that the lexical semantic classes considered are monosemous. As we observed in Section 4.3.3, this assumption had a negative effect on some of the results obtained, especially because the data sets used were not disambiguated for any specific task. In this sense, lexical ambiguity is a more or less ubiquitous problem in most NLP tasks, at least in all of those tasks that involve access to the contents of utterances. This phenomena will be addressed in detail in Chapter 5.

Finally, thus far in Chapter 4, we have addressed different distributional models, and moreover, different strategies to build distributional word representations. Besides observing the increase of predictive power with the inclusion of syntactic dependency information. The DM model, for instance, obtained an $F1$-Score that was an average of 6 points higher than the other models. We also observed the decrease in predictive power of certain models when corpus data was reduced (recall the decrease of $F1$-Score of 10 points with the LINE model when using the BNC90M). Furthermore, the empirical evidence confirmed that supplementary information is critical when there is not sufficient data for an accurate class membership decision. For these reasons, we conclude that we still need strategies to identify how to include more indicative distributional information, which do not affect the sparsity in vectorial spaces. These strategies must ensure that these spaces are filled with descriptive and relevant information, resulting in a decrease of noise as well as a reduction in sparsity.

Thus far, we have based the acquisition of lexical information on the increment and refinement of the features used to construct the distributional model. Each model studied had used the frequency of occurrences of each of these features with target lexical items in corpus data to build word representations. In this way, these models assume that the more frequently a feature occurs with members of a given class, the more predictive that feature is, resulting in more predicative power in the word representation.

In this next Section, we use Word Embedding (WE) models [Mikolov et al., 2013], based on Neural Networks to map the similarity of instances observed directly from corpus data to build word representations. These representations have been demonstrated to provide more similar word representations to the classifier, which reduces the reliance on the selection of relevant features. We explain in detail the experiments conducted with these WE models in Section 4.4.

4.4 Using WE representations for nominal lexical semantic classification

Recently, Word Embedding (WE) models [Mikolov et al., 2013] have been adapted and used more and more to build word representations for NLP tasks. such as Named Entity Recognition, Chunking and Semantic Role Labeling [Turian et al., 2010, Collobert et al., 2011, Socher et al., 2011]. Additionally, theses learned vectors have proved to perform better than frequency-based vectors (see [Baroni et al., 2014] for a thorough comparison) in certain NLP tasks. The salient novelty of this approach is that the features of word representations are parameters to be learned. Moreover, the number of dimensions is defined a priori, and the representation of corpus data is then condensed into the pre-defined number of features specified. In this Section, we evaluate how WE can be used for and how they effect our nominal lexical semantic classification tasks.

In contrast to the models studied in Section 4.3, (WE) models are learned using neural networks by means of the back-propagation algorithm that minimizes the differences between training samples. Intuitively, neural networks take into account observed word-context pairs and induce latent parameters on the basis that words that appear in the same contexts have similar parameters. It has been demonstrated that these learned vectors indeed capture syntactic and semantic similarities [Mikolov et al., 2013].

[Levy and Goldberg, 2014a] have extended the approach proposed by [Mikolov et al., 2013] that uses only the raw text of corpus data, to also include syntactic information provided by dependency-parsed corpora in the representations. The objective is to reduce the scope of "co-occurrence" (or context windows) to words that occur in a dependency relation with a target word. This results in WE representations that assign similar vectors to words that occur in similar contexts. Essentially, this provides word vectors that are already similar to each other to a machine-learning classifier, which removes much of the decision-making process, as the representations are already indicative of the classes that the words belong to due to their similarity.

4.4.1 Using word embeddings for lexical-semantic classification

In previous Sections, we have identified that one of the crucial issues in lexical classification is word representation as a problem of feature design and selection due to the sparsity of many of the frequency-based features. If we recall from the previous Section, different models require different numbers of features, directly affecting the resulting word representations. The most exhaustive models, such as the DM model, also required the highest number of features, which although effective, does result in sparsity in the vectors.

For our specific task, we build our word representations using WE derived from dependency-parsed information for our task of lexical semantic classification. Our intuition was the dense WE vectors, with no zero values, that maintain the distributional information of corpus data, will indeed be useful to a classifier that tries to find similar semantic components between different words. Moreover, we considered this particular instance of WE specifically because it also includes dependency information from a syntactic parse, which previous work has demonstrated to have a crucial role in obtaining stronger classification results, as discussed in detail in Section 4.3. In this Section, we further build on the work presented in Section 4.3 by further comparing the different distributional word representations already studied with the word representations that we obtain from WE models.

One of the most salient differences between the count-context models used for nominal lexical semantic classification in Section 4.3 and the work described in this Section is the number of features considered (see Table 4.10). While the number of features considered for LING, LINE and $type_{\mathrm{DM}}$ could vary depending on the corpus (due to the length of its vocabulary) or the number of selected features, for the WE model the number of features used is constant. This is because the number of features considered per word is actually a parameter to be learned and is defined a priori as such. For the purpose of the work in this Section, we used a standard dimension size of 200 [Levy and Goldberg, 2014a] to build our word representations. Table 4.10 provides a comparison of the number of features considered for each model.

As described above, WE do not look to find the most indicative features of the lexical-semantic class, such as LING attempts to do, nor does it consider all of the frequency information found in the context window, such as LINE. Rather, WE tune the values of a set of features to assign similar words similar vectorial representations.

4.4.2 Experiments

For our experiments, we used dependency-based word embeddings to build our word representations, as detailed in [Levy and Goldberg, 2014a]. The dependency-based word embeddings are actually a modified version of the skip-gram approach[6] [Mikolov et al., 2013] that also incorporates the syntactic information from a dependency parse into its representation. Therefore, this

[6] As mentioned earlier, the WE built with the skip-gram approach [Mikolov et al., 2013], better known as WORD2VEC, are available for download and use at https://code.google.com/p/word2vec/.

approach also considers contexts based on the syntactic relations the word participates in, which is in contrast to the simple raw words of the windows of size k around the target word of WORD2VEC.

We compared the performance of new classifiers against the three models previously used for the same task and using distributional vectors, as described and presented in Section 4.3. To ensure a direct comparison to directly compare the results obtained in this Section, we trained the WE on both the 90 million token BNC90M corpus and the 3 billion token LARGE3BN corpus, following the methodology defined in Section 4.3. All evaluations were conducted in a 10-fold cross validation setting.

4.4.3 Results

Table 4.16 presents the results obtained for the LINGLINE and TYPE$_{\text{DM}}$ model, as well as the results obtained using WE representations. The results obtained with WE demonstrated a statistically significant improvement ($p < 0.05$) over the TYPE$_{\text{DM}}$ model for the COM, EVT and LOC classes. For the HUM and ORG classes, although there is a slight improvement (0.003 for ORG and 0.017 for HUM), these improvements did not demonstrate statistical significance. Furthermore, the results also demonstrated an overall average accuracy increase of 16.0 and 6.0 points over the LINGLINE and DM models, respectively.

Class	LINGLINE*	$type_{\text{DM}}$*	WE
COM	0.80	0.88	0.93
EVT	0.81	0.83	0.93
HUM	0.86	0.91	0.93
LOC	0.83	0.81	0.93
ORG	0.77	0.84	0.84
Average $F1$-Score	0.81	0.84	0.91

Table 4.16: $F1$-Score for each class using each model over LARGE3BN. The results for the models marked with * were previously obtained, reported and elaborated upon in Section 4.3

Table 4.17 presents the results obtained we obtained using WE word representations with the BNC90M corpus as well as the results for the LINGLINE model, first presented in Section 4.3.

Class	LINGLINE*	WE
COM	0.72	0.90
EVT	0.71	0.93
HUM	0.89	0.90
LOC	0.73	0.87
ORG	0.76	0.83
Average $F1$-Score	0.76	0.89

Table 4.17: $F1$-Score for each class using each model over the BNC90M

The results obtained for WE using the BNC90M corpus data again demonstrated a statistically significant improvement ($p < 0.05$) over the LINGLINE model and with smaller corpus data. This improvement is consistent for each class, except for the HUM class (in which the improvement was not statistically significant at ($p < 0.05$) for its highest scoring model). Furthermore, the results obtained with the WE model demonstrated an average overall increase of 13.0 points in the $F1$-Score over the LINGLINE model.

We attribute the increase of $F1$-Score to the fact that the LINGLINE model is more dependent on size and availability of corpus data for training purposes. Thus, the results indicated that WE representations have an advantage over both of these models when considering a reduced corpus size (90 million tokens) for training. However, we do acknowledge that although we reduced the corpus size, a 90 million token corpus should by no means be considered a “small” amount of data.

4.4.4 Discussion

In this work, we evaluated the word representations built with WE in a nominal lexical semantic classification task to determine whether the learned vectors provide more informative word representations, resulting in improved classification decisions. Given the extremely successful results obtained, we explored the two main obstacles that we identified to have an impact on the results of word representations when analyzing the results, especially given the novelty of using this approach for nominal lexical semantic classification:

- The effects of frequency of target lemmas on classification when using WE representations.

- The impact of polysemy on classification results obtained when using WE representations.

Frequency effect on WE models

Our intuition was that word frequency would *not* have an effect on the accurate classification of target nouns when using WE word representations. This is due to the fact that feature vectors are weighted based on similarity of a target nominal to other nominals found in similar contexts, rather than on how many times the noun itself was seen (but, probably, on how many shared contexts were found). Moreover, the WE are using all of the contexts available to a given word, similarly to the DM model, hence the minimal differences between the results obtained with these two models.

We observed in our results that the frequency of the target nouns considered per class ranged from an absolute frequency of 192 in corpus data (*pugilist* of the HUM class) to an absolute frequency of $2,350,093$ in corpus data (*information* of the COM class). We then looked at the frequency of occurrence of those nouns per class that were misclassified to determine if the misclassifications occurred with nouns below a certain frequency. However, the results obtained did not demonstrate any particular, or recognizable, pattern in the amount of occurrences of the misclassified nouns. For instance, there were sixteen misclassified nouns from the COM class. Of these nouns, only one noun, (*notepaper*), had an absolute frequency of less than $1,000$, while nine nouns were observed to have an absolute frequency between $2,500$ and $8,500$ in corpus data, five of which had an absolute frequency of more than $20,000$.

This trend was also consistent for the other classes studied. The HUM class obtained twelve misclassifications, of which one noun (*highness*) has an absolute frequency of less than 500 times, while nine nouns had an absolute frequency ranging between $10,000$ and $100,000$, with two of these instances (*customer* and *human*) occurring each more than $1,000,000$ times in the corpus data. Although there were only eight misclassifications of the LOC class, the nouns misclassified were not necessarily low-frequency nouns with absolute frequencies ranging from $2,668$ (*domicile*) to $121,767$ (*port*).

Along this line, we cannot conclude that frequency of the noun in corpus data is one of the obstacles of classification when using WE. Our results indicated that the absolute frequency of a noun does not affect classification. This is in contrast to the LING model, for which the total frequency of target lemmas is crucial because the availability of distributional information is dependent on target nouns occurring with linguistically-motivated class-indicative contexts, and it is also in contrast to the LINE model, which is dependent on the availability of a large amount of corpus data to ensure a sufficient amount of surface information. The results obtained confirmed that the representations built by WE are not actually dependent on high-frequency nouns, as tends to be the case with count-context models. Thus, we can say that the construction of efficient WE representations are dependent on the occurrences of a target in contexts where other similar target nouns are occur, rather than the amount of occurrences of that target, such as in count-context models.

Polysemy and WE models

As clearly identified in Section 4.3, polysemy is an obstacle to distributional model due to the conflation of all senses of a target word in one vectorial representations. This issue mainly stems from the fact that all of the co-occurrence information of a target noun is being stored in one feature vector, although some nouns have the potential to occur as more than one sense within corpus data. This becomes a problem when providing feature vectors to a classifier especially because all of the senses that a noun can be selected for are dispersed within the vector and, therefore, the noun may not have enough distributional information as a member of any of its classes for classification. Moreover, nouns may not occur sufficiently as a member of the class to be classified, thus the distributional profile for that noun may not provide any information toward the indication of that specific class.

Our intuition here is that WE will lessen the effect that this sense dispersion can have in count-context models by more evenly distributing the available sense information within the distributional representation. An analysis of the results obtained with WE indicated that some cases of misclassifications can be due to polysemy within our results. One particular case that we observed was with the ORG noun *delegation*, which was misclassified as a ORG noun and then later classified as an EVT noun. Here, the noun *delegation* may have occurred as a nominalization in our corpus data in contexts that are more similar to other words indicative of *"the action of entrusting a responsibility to others"*, rather than in contexts of words indicative of a *"persons who are representative of others"* and thus acquired a feature vector more similar to EVT nouns. To illustrate this difference further, consider the noun *delegate* in Examples 9 and 10, where it is selected for as members of different classes in each context.

(9) The delegation (HUM) arrived to the summit prepared to propose their new

platform.

(10) The main responsibility of the manager is the delegation (EVT) of tasks among the group members.

Another case of misclassification due to polysemy that we observed was with the HUM noun *parent*, which was classified as a COM noun. In this case, we also observed a temporal shifting of senses within corpus data, which can be dependent on the type of corpus considered. For instance, in some contexts, the noun *parent* can be used as a term in a computer science domain to define *"a node that is one step higher in the hierarchy"*, rather than being used in the sense of a *"familiar caretaker of children"*. In this case, again, we have to consider not only the occurrence of the misclassified noun in contexts indicative of its target sense, but also the entire network of information that the WE model links to it through mapping, which could bias the resulting word representation to one sense over the other.

Thus, WE representations do not fully solve the problem of polysemy in nominal lexical semantic classification tasks; but they do seem to lessen the effects of its consequences, especially in comparison to results obtained using other types of models, as described in Section 4.3. Overall, although we identified punctual cases of misclassifications due to cases of lexical ambiguity, as discussed above, the results obtained still merit the conclusion that WE representations can much more efficiently handle polysemy than representations built with count-context models.

4.4.5 Final remarks

Considering the results obtained using WE for our nominal lexical semantic classification task, our results indicate that WE representations handle polysemy more effectively than count-context models, as the number of misclassifications due to polysemy is lower, in general, to what has previously been reported in the State of the Art as well as what has been observed in the results of previous Sections. However, we must note that the data set used for the evaluation of the work presented in this Section was not constructed to specifically evaluate the effects of polysemy on classification, as described in Section 3.1. The use of WE representations to the handle obstacles of polysemy in classification tasks will be revisited in Section 5.3.

Given the results obtained and our goal to evaluate the use of WE representations in nominal lexical semantic classification tasks, our results confidently confirmed that the information provided by WE in their distributional representations does result in more accurate and robust classification decisions, further verified by the statistically significant improvement in precision, recall and $F1$-Score, in comparison to other models that we considered in the scope of this thesis. This type of an analysis would require further investigation, including an inspection of the contexts that contribute to the construction of WE. Yet, WE representations do not allow for the intrinsic inspection of the behavioral characteristics that led to a particular classification decision. Thus, although the use of WE representations improves the accuracy of classification decisions for all of the classes studied, from a linguistic perspective, we are unable to draw conclusions regarding the lexical or linguistic boundaries that distributionally characterize each classes, as we are able to do with other models.

Furthermore, this raises the question of whether to pursue the identification of learnable lexical-syntactic contexts that can induce word representations or simply improve upon or refine these distributional representations that cannot be inspected. The debate of this question goes beyond the scope of the work conducted for this thesis, but it does open a door for extremely relevant and interesting future research lines.

Chapter 5

HANDLING POLYSEMY IN DISTRIBUTIONAL WORD REPRESENTATIONS

The work presented in the previous Chapter focused on different distributional models to build word representations of nouns from corpus data. Based on the results obtained, we concluded that lexical semantic classes can be automatically acquired using corpus data, using distributional characteristics, although the success of classification is directly tied the use of information by the model. However, one issue recurrently observed to be a cause of "errors" in classification decisions was lexical ambiguity, even after overcoming the obstacle of data sparsity with WE representations.

Furthermore, and as discussed in Chapter 2, this topic is also typically ignored in most related empirical work [Boleda et al., 2012b]. Lexical ambiguity is an important phenomenon to cover for several reasons, for instance, it that can introduce *noise* into a vector due to the "occurrence" of a target nouns with misleading corpus features, in which nouns occur in contexts that do not correspond to their assumed lexical class. This is also due to the fact that most data sets do not take into account the possibility of nouns belonging to more than one lexical semantic class. Furthermore, as the distributional representation of each word conflates all of the senses into one vector, misclassifications can occur if a target word occurs more frequently in contexts characteristics to another class. Because lexical ambiguity is an extremely broad topic, we focus on one aspect that we consider to have the biggest impact on distributional word representations: regular polysemy.

Regular polysemy, or the systematic alternation between two senses of a word, hinders distributional representations because it allows some nouns to be selected for as a member of multiple classes. This phenomenon affects a number of members of lexical semantic classes, and it occurs only in particular classes, where it has been recurrently seen. This results in the word representations of regular polysemous nouns to contain distributional information that is indicative of more than one lexical semantic class in the same vector. Moreover, regular polysemy presents a challenging problem for supervised classification tasks, such as in our case, because most authors do not distinguish among related senses of the same word in their data sets, considering individual item as part of a class or not ([Hindle, 1990]; [Bullinaria, 2008]). This is particularly problematic when words allow for multiple selection, i.e. when different senses of the same lexical item can be simultaneously selected for in one sentence, as illustrated in Example (11).

(11) noun: *church*

a. The church discussed its role in society at the gathering. (ORGANIZATION)

b. The choir rehearses on Saturdays at the church. (LOCATION)

c. There is a collection organized (ORGANIZATION) by the church on Mulberry Street (LOCATION) this Sunday.

Example (11) demonstrates how a noun like *church* can denote an ORG noun in (11a) and a LOC noun in (11b). Moreover, it also shows how *church* can denote both an ORG and a LOC noun in one context, in (11c). The complexity of complex-type selectional behavior in context, as illustrated in Example 11, makes it difficult to apply to complex types the standard notion of word sense, as used in automatic text processing tasks. As discussed in Chapter 2, traditional word sense disambiguation (WSD) systems are not an appropriate solution for this task due to the fact that a decision for a single sense must always be made, despite the fact that in a context such as (11c) both senses of the noun are activated by the context.

The different contexts in Example (11) indicates how distributional representations can be affected by the ability of a noun to be selected for as a member of different classes. This is because the noun will not occur equally as a member of each class in corpus data, which results in distributional vectors containing asymmetrical information for each class of the noun. As classification algorithms use the distributional information provided in a word representation to assign class membership, performance can be hindered by asymmetrical representation of different senses, resulting in a negative effect on classification decisions. Thus, it is imperative to correctly model and consider this phenomenon when building word representations to build lexical semantic classifier. Furthermore, accounting for this phenomenon in classification systems represents an important step towards implementing systems that can assign meaning to words dynamically depending on the context in which they occur [Cooper, 2005].

Additionally, we also consider regular polysemy an important topic to cover in this thesis because the incorporation of this type of rich complex-type nominal information into the word representations can serve to reduce the search space in disambiguation tasks, and thus the number of decisions needed. Moreover, it can also provide grounds to opt for the non-disambiguation of instances when relevant, for example in co-predication contexts like (11c), which more accurately models the contextual uses of these types of nouns. Furthermore, the knowledge of the entire sense potential of a given word is sometimes required for specific tasks (see for instance [Rumshisky et al., 2007] and [Lenci, 2014]), thus resulting in more complete and precise representations of these lexical items.

In the subsequent Sections, we propose to address regular polysemous nouns as members of a given ambiguity class (within a wider lexical semantic class) and making apparent the relation between members of different classes by identifying shared properties beyond class limits. Thus, we place ourselves within the Generative Lexicon (GL) framework [Pustejovsky, 1995], as it provides the tools to account for regular polysemous nominal lexical units that display rich variations of meaning in language use. Furthermore, GL allows for the identification of refined and relevant semantic features, as well as to capture information that does not necessarily emerge from a purely corpus-based collocation analysis.

As defined earlier, the nouns that can instantiate this phenomenon have been defined in GL [Pustejovsky, 1995] to be *complex-type* nouns. Complex-type nouns are formed by the intersection of two (or more) senses that they can be selected for in context, thus they are typically recognized in context by a bullet that joins the classes together ($x \cdot y$). In Section 5.1 we perform a detailed error analysis regarding the classification of complex-type nouns. In Section 5.2 we study the difference between super, or (x/y), classes that encompass all of the nouns related to each lexical-semantic class (or classes) of interest. In our case, the (x/y) class includes simple-type nouns from classes (x) and (y), as well as complex-type ($x \cdot y$) nouns.

To address the differences between the two types of nouns, we proposed a dedicated two-step approach to capture the distributional behavior of regular polysemous words in corpus data in a way that accurately reflects the semantic complexities of different types of (related) lexical-semantic classes. By classifying nouns into a broader super (x/y) class in a first step, we are able to also capture more nouns due to the larger distributional profile of the class. In a second step, we were then able to differentiate between complex types ($x \cdot y$) and their monosemous counterparts (x and y) that form the (x/y) super class. Finally, in Section 5.3 we expand our data set to also use the Cascade approach to classify nouns into CoreLex classes. Furthermore, in this Section we also used the WE model to build word representation, due to their success reported in Section 4.4.

5.1 Using the Qualia Structure to identify lexical-semantic classes

Authors such as [Pustejovsky, 1995, Ježek and Lenci, 2007, Lenci, 2014] have shown how distributional analysis and theoretical modeling interact to account for rich variation in linguistic meaning, especially because blind-theory distributional approaches have been shown to fail to account for the wide range of linguistic behavior displayed by words in language data (see [Pustejovsky, 1995]). In this section, we proposed and evaluate the use of automatically obtained FORMAL role descriptors as features to cluster nouns. As introduced in Chapter 2, the FORMAL role is one of the four roles that form the Qualia Structure (QS), the structure that defines the semantic properties of a noun, according to the GL [Pustejovsky, 1995].

We specifically used the FORMAL role of the QS because it directly corresponds to the facets of "what an object is", thus it can be considered as a feature that identifies lexical class membership. Figure 5.1 illustrates the semantic composition of the FORMAL role of complex-types nouns, according to the GL, in an attribute-value matrix (AVM), consisting of the feature structure of the noun [Copestake and Briscoe, 1995]. According to the GL, the FORMAL role distinguishes a lexical object within a larger domain, using the same concept as our lexical semantic classes, which are essentially the larger domains, or super-types, of a given noun.

Building on this definition, we assume that nouns belonging to a certain class display particular features shared by other nouns of that class, which should distinguish them from members of other classes. For example, the EVT class may extract the noun *activity* as a FORMAL role descriptor, representing a common feature shared between these lexical items of that class. In our case, we expect the use of the FORMAL role descriptors as features of the lexical-semantic classes of a noun to capture information regarding all of the possible lexical-semantic classes of a noun; moreover serving to identify those nouns that can be members

$$\begin{bmatrix} \boldsymbol{\alpha} \\ \mathit{ARGSTR} = \begin{bmatrix} \mathit{ARG1} = \mathbf{x} : \tau_1 \\ \mathit{ARG2} = \mathbf{y} : \tau 2 \end{bmatrix} \\ \mathit{QUALIA} = \begin{bmatrix} \tau_1 \tau_2 \_\mathbf{lcp} \\ \mathit{FORMAL} = R\,(x,y) \end{bmatrix} \end{bmatrix}$$

Figure 5.1: AVM of complex-type nouns

of more than one lexical semantic class. We part from the hypothesis that the FORMAL role of a noun, by providing information about "what an object is". We consider that this information is representative of a feature that is bound to be closely related to a particular lexical semantic class. Thus, we hypothesize that the FORMAL role of the QS is sufficient to discriminate between lexical semantic classes of English nouns.

5.1.1 Extracting Qualia role information

Unlike linguistically-motivated cues, the information extracted from the FORMAL role provides a guideline to understand to what extent we can justify the membership of certain nouns in just one lexical-semantic class. Along this line, the FORMAL role features provide us with information regarding what the elements are semantically "made of", in this way moving away from grammatical knowledge by trying to instead extrapolate semantic knowledge from corpus data.

As there were no available lexica annotated with FORMAL role information, we developed a method to obtain it automatically and carried out clustering experiments. Automatically extracting qualia with lexico-syntactic patterns has received attention for its success: [Hearst, 1992] identified lexico-syntactic patterns to acquire noun hyponyms, which correspond to the FORMAL role, whereas [Cimiano and Wenderoth, 2007] identified lexico-syntactic patterns to obtain information regarding the specific semantic relations that correspond to each qualia role. As we needed information regarding the FORMAL role, not full lexical entries, in order for clusters to emerge, following [Celli and Nissim, 2009], we bypassed the representation of the entire QS, assuming semantic relations can be induced by matching lexico-syntactic patterns that convey a relation of interest.

Along this line and, moreover, considering that there are nouns that are representative of more than one sense and, therefore may contain distributional information indicative to more than one lexical semantic class, we followed the cue-based methodology to employ our nominal classifiers, described in Section 4.1 to have a direct representation regarding when certain nouns occur in context as members of one class or of another. We conducted classification experiments for three different classes that commonly participate in regular polysemous alternations [Pustejovsky, 2005]: HUM, LOC and EVT.

5.1.2 Experiments

In the experiment performed, we employed two steps:

- In a first step, we extracted FORMAL role descriptors from corpus data;
- In a second step, we used the representations built from this data for clustering.

To obtain FORMAL role descriptors for our unsupervised clustering task, we used UKWAC60M, as described in Section 3.2. For our data set, we employed 60 target nouns, selected from the monosemous data sets described in Section 3.1.

Building the data set

As explained above, given the unavailability of lexica annotated with FORMAL role information, and considering our basic goal of evaluating whether this information is enough to cluster together nouns of the same class, we first extracted distributional information for targets nouns from a corpus using hand-crafted lexico-syntactic patterns indicative of the FORMAL role, adapted from [Hearst, 1992] and the list proposed by [Cimiano and Wenderoth, 2007]. Table 5.1 provides a complete list of the patterns. Each patterns was formalized through Regular Expressions with PoS tags given after each token.

The extracted information was stored in feature vectors representing co-occurrences with target nouns in relevant contexts, as defined by the patterns in Table 5.1. Each element corresponds to occurrences of a particular target noun (x) with a possible FORMAL role descriptor (y), following [Katrenko and Adriaans, 2008]. Using the patterns in Table 5.1, we obtained 185 FORMAL role

Lexico-syntactic patterns indicative of FORMAL roles
x_($or \vert and$)_other_y
x_such_as_y
x_($is \vert are$)_(a/an/the)_(kind(s)/types(s))_of_y
x_($is \vert are$)_also_known_as_y

Table 5.1: Patterns used to detect FORMAL role information in corpus data were built

descriptors for 55 of the 60 target nouns in 353 occurrences.

Given the properties of the clustering algorithm used, a random value would be provided to nouns not sharing feature information with any other noun in our data set. To avoid random cluster assignations and to provide more significant information to the system, we filtered out the features not shared between at least two target nouns. We did not control for what class the shared features belonged to. Though we employed a large set of data, there were not enough shared FORMAL role descriptors for an important part of our data set. For this reason, we devised a strategy to increase the information available to the clustering algorithm, which we will describe in detail in the following Section.

Bootstrapping for more features

To obtain more FORMAL role descriptors, we employed a bootstrapping technique [Hearst, 1998] that relies on monotonic patterns for natural language inference [Hoeksema, 1986, van Benthem, 1991, Sánchez Valencia, 1991] and as illustrated in Example (12). This strategy is consistent with the GL lexical inheritance structure [Pustejovsky, 1995, Pustejovsky, 2001] that assumes lexical items obtain their semantic representation by accessing a hierarchy of types and inheriting information according to their QS. Thus, qualia elements can be viewed as hierarchically organized categories in which some sub-categories are subsumed by more general super categories as one moves up in the hierarchy.

(12) noun: *mammal*
  a. A mammal is a [type of] animal.
  b. A zebra is a [type of] mammal.
  c. Therefore, a zebra is a [type of] animal.

To illustrate how this applies specifically in our case, the HUM noun *treasurer* obtained *officer* as a FORMAL role descriptor, whereas *officer* extracted *person* and *employee* as its own FORMAL role descriptors in Example 13.

(13) noun: *officer*

a. An *officer* is a [type of] employee.

b. A treasurer is a [type of] *officer*.

c. Therefore, a treasurer is (also) a [type of] employee.

Assuming this lexical organization, we consider FORMAL role descriptors extracted for *officer* to also be features of *treasurer*. Thus, we gathered additional information regarding the nouns to cluster, using the originally obtained FORMAL role descriptors as target nouns to extract more elements in an attempt to overcome biases due to sparse data, as well as to reinforce information already obtained.

We conducted one iteration of the aforementioned bootstrapping technique, going up one level of generalization to obtain the final distribution of information below. The newly obtained feature information was unified with the previously extracted features, filtering out any additional noise attained. Table 5.2 presents the final distribution of this information.

Class	Elements	Occurrences
HUM	61 elements	841 occurrences
LOC	43 elements	225 occurrences
EVT	36 elements	216 occurrences

Table 5.2: Distribution of FORMAL role descriptors extracted (after filtering and bootstrapping per class of target noun)

Basing our clustering experiment on automatically extracted FORMAL role descriptors, the accuracy of information obtained was a concern. To assess the accuracy of the information obtained, the FORMAL role descriptors extracted were revised manually. Extractions were considered erroneous if they provided information not in accordance with the class that the target nouns pertained to. Table 5.3 presents the results of this analysis. Erroneous extractions caused by faults of the extraction mechanism (i.e. problems handling phenomena such as PP attachment), PoS tagging errors, lexical ambiguity or erroneous statements in text [Katrenko and Adriaans, 2008], as well as errors due to regular polysemy. Note that although errors were identified, they were not filtered out for the clustering task, i.e. all information (erroneous or not) was included. We discuss the impact of these errors, specifically the errors attributed to regular polysemy, later in this section.

Class	% of accurate FORMAL role descriptors extracted
HUM	87.60%
LOC	63.54%
EVT	75.96%

Table 5.3: Percentage (%) of accurate FORMAL role descriptors obtained by class

Clustering word representations

The second step of our experiment consisted in clustering nouns using the automatically extracted FORMAL role descriptors. To empirically demonstrate the extent to which FORMAL role descriptors clustered together nouns from the same class. We used the SIB algorithm [Slonim et al., 2002] in the WEKA [Witten and Frank, 2005] implementation to cluster target nouns into lexical semantic classes, based only on the FORMAL role information obtained. We selected the SIB algorithm due to the manner that it manages larger data sets through reduced complexity.

The SIB algorithm measures the similarity between two vectors using the Jensen-Shannon divergence, which measures the similarity between two probability distributions, rather than Euclidean distance, which can unfairly bias the data if the number of attributes representing the factors is not equal [Davidson, 2002]. Furthermore, we selected the SIB algorithm because our feature spaces were dependent on the number of FORMAL role descriptors each target noun occurred with in the corpus.

5.1.3 Results

The goal of this experiment was to use FORMAL role descriptors to cluster together target nouns from the same lexical-semantic class. To evaluate this task, we compared the nouns of each cluster to the class information indicated in the corresponding monosemous data sets described in Section 3.1. Tables 5.4 and 5.5 present clustering results. The distribution of nouns across each cluster is given by the percentage of nouns pertaining to each lexical class included in it. The total number of target nouns in each cluster is also provided.

Class	Cluster 0	Cluster 1	Cluster 2
HUM	92.85	00.00	57.14
LOC	07.69	39.13	14.29
EVT	00.00	60.87	28.57
Total Number of Seed Nouns per Cluster	14	23	7

Table 5.4: Distribution of nouns in a 3-way clustering solution (%)

We experimented with both a 3-way and a 4-way clustering solution. The 3-way clustering solution resulted in the clustering of HUM nouns (Cluster 0). LOC and EVT nouns grouped together in Cluster 1, the remaining cluster being composed of nouns from all classes with very few features available (less than three), i.e. insufficient information for classification. Considering this, we employed a 4-way solution to see whether LOCATION and EVENT nouns could be discriminated. This solution distinguished between the three classes (Cluster 0, 1 and 3 in Table 5.5) with a fourth cluster containing those nouns that did not have sufficient information for an accurate class membership decision, and moreover, had a negative effect on the 3-way solution.

Class	Cluster 0	Cluster 1	Cluster 2	Cluster 3
HUM	0.00	0.00	57.14	0.93
LOC	0.00	90.00	14.29	7.69
EVT	100.00	100.00	28.57	0.00
Total Number of Seed Nouns per Cluster	13	10	7	14

Table 5.5: Distribution of nouns in a 4-way clustering solution (%)

The results show that even after filtering and bootstrapping the features extracted, sparse data still affected the results. However, nouns whose most salient common trait was the lack of sufficient information were consistently grouped together. In this was, the results demonstrate that the clustering solutions are able to discriminate between lexical semantic classes, as well as to detect those nouns for which there is not sufficient information by clustering them together.

5.1.4 Discussion

The clustering algorithm discriminated between the three classes considered, using only the FORMAL role descriptors extracted from corpora data as features. Leaving aside the nouns for which there was not enough information available (12.7% of our data set), EVT, HUM and LOC nouns were discriminated in the 4-way clustering solution (Clusters 0, 1 and 3, respectively, as presented in Table 5.5). In this Section we analyze the misclassified nouns to understand the reasons behind their misclassification and to evaluate to which extent they correspond to recurring phenomena in language, which can possibly be accounted for and overcome with additional strategies.

Although their impact is not significant, noisy extractions, as explained earlier, do play a role in misclassification. In the 4-way clustering results, for instance, an EVT noun is included in the cluster dominated by LOC nouns due to errors in extraction, specifically the incorrect identification as a FORMAL role descriptor of the noun in a PP modifying the head noun of the NP that should be extracted. This type of noise is mostly generated by the use of low-level NLP tools. However, the existence of this type of noise, caused by the tools used, in the data did not significantly affect the clustering algorithm, as demonstrated by the accuracy of the clustering results presented in the previous section.

Concurrently, although general patterns can be identified in language use, one of the main characteristics of language data is its heterogeneity, which means that elements of a given lexical-semantic class do not necessarily share all their features or show perfectly matching “expected” linguistic behavior. Moreover, considering that lexical items are complex objects with different semantic dimensions, they may also share properties with elements of more than one lexical class. This type of phenomenon is behind some of the misclassifications in our data, such as the clustering of the LOC noun *factory* with HUM nouns. This misclassification seems to be related to the fact that a part of HUM class members tended to obtain FORMAL role descriptors typical of HUM nouns, as well as of ORG nouns, making apparent that nouns do not always occur in the sense considered in our pre-classified list of nouns.

The case of the noun *factory*, which was clustered with HUM nouns, clearly demonstrates how the polysemy described above can partially apply to this noun. Among the descriptors obtained for *factory* we found, alongside descriptors typical of LOC nouns, nouns such as *sector*, *organization* and *profession*, which were descriptors that were also extracted for HUM nouns showing the HUMANGROUP·ORG logical polysemy, indicating that nouns like *factory* are also

$$\begin{bmatrix} \textbf{factory} & \\ \text{ARGSTR} = & \begin{bmatrix} \text{ARG1} = \textbf{x: location} \\ \text{ARG2} = \textbf{y: organization} \\ \text{ARG3} = \textbf{z: human} \end{bmatrix} \\ \text{QUALIA} = & \begin{bmatrix} \text{FORMAL} = \mathbf{x{\cdot}y{\cdot}z} \end{bmatrix} \end{bmatrix}$$

Figure 5.2: AVM of the noun *factory*

complex objects, as illustrated in Example (14).

(14) noun: *factory*

a. The factory on the corner of Main Street is big and brown. (LOC)

b. The factory summoned a protest against the new government sanctions. (ORG)

c. There was a protest organized (ORG) by the factory that burned down (LOC) last week.

In our data, *factory* shared features both with definite plural NPs headed by HUM nouns like *teacher* and *employee* and LOC nouns such as *kitchen* and *resort*. The linguistic behavior of *factory* can, therefore, be assumed to reflect the regular polysemy of ORG·LOC·HUMANGROUP complex-types identified by [Rumshisky et al., 2007], further illustrated in the AVM of the noun *factory* is presented in Figure 5.2.

Furthermore, some HUM nouns obtained FORMAL role descriptors typical of ORG nouns that indicate a type of polysemy that occurred in our data only with plural HUM nouns. This alludes to the works of [Copestake and Briscoe, 1995] and [Caudal, 1998], according to which some HUM nouns show a specific type of polysemy when heading definite plural NPs: the polysemy between the individual HUM sense and the collection of HUMANS sense, which in turn is polysemous between the HUMANGROUP and ORG senses. In Example (15) we see how the definite plural NP *the doctors* can select for the two senses typically denoted by collective nouns, while having also the possibility to denote individual entities, which is not possible with collectives, as demonstrated in Example (16)), that cannot occur in contexts that force a distinct individual entity reading.

(15) nouns: *doctors*

a. The doctors lay in the sun. (several individual HUM entities)
b. The doctors protested in front of the hospital. (HUMANGROUP)
c. The administration negotiated with the doctors. (ORG)

(16) nouns: *staff*, *employees*, *administration*
a. # The staff lay in the sun. (several individual HUM entities)
b. The employees lay in the sun. (several individual HUM entities)
c. The staff protested in front of the hospital. (HUMANGROUP)
d. The administration negotiated with the staff. (ORG)

As both collectives and definite plural NPs denote collections, [Caudal, 1998] states that it is desirable to account for the polysemy of such items morpho-syntactically. This analysis is further strengthened by the observation that, unlike pairs such as *employee* and *staff*, for nouns like *doctor* there is no lexicalization for *group of doctors* in English. The same being true for collective nouns like *audience* or *committee*, whose individual members are not lexicalized. Given such lexical gaps, morpho-syntax is the strategy available. However, though logically polysemous, plural definite NPs like *the doctors* do not allow for multiple selection as is typical of complex types: once the individual HUM sense has been selected for there is no access to the HUMANGROUP·ORG sense, as suggested by Example 17 (see [Buitelaar, 1998] and [Rumshisky et al., 2007]).

(17) *The administration negotiated with the doctors, who later lay in the sun. (several individual* HUM *entities)*

[Pustejovsky, 1995]:155 claims these patterns of linguistic behavior are due to the information in the QS. In the case of expressions like the *doctors*, the dot element denoting the individual HUM entity and the complex type HUMANGROUP·ORG correspond to different qualia roles, as represented by the AVM in Figure 5.3. Hence, the different senses of the expressions cannot be selected at the same time.

For our work, the most relevant aspect of the behavior displayed by nouns like *factory* and *doctors* is that it makes apparent how our strategy to extract FORMAL role descriptors reflects the ambiguity of nouns to be clustered, which is often difficult to handle in NLP, particularly in classification tasks. The clustering solutions we obtained, as described in the results, grouped together HUM nouns, both those that display the ambiguity discussed in this Section and those that do not, the same being true for LOC nouns. And yet, polysemous nouns display features that clearly point towards the existence of finer-grained distinctions, i.e. sub-classes within lexical semantic classes. We demonstrated that these

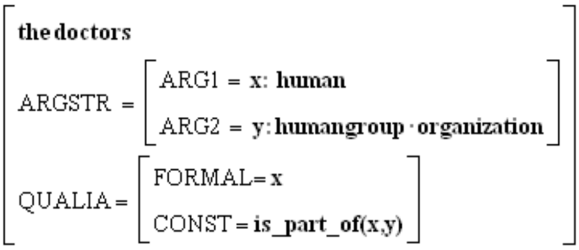

Figure 5.3: AVM of the noun *doctors*

fine-grained distinctions are mirrored in FORMAL role descriptors, we assume it should also be possible to automatically recognize groups of nouns within the same ambiguity class, i.e. complex-type nouns.

To validate this hypothesis, we performed an additional iteration of the clustering using the same features and algorithm over previously identified clusters. The iteration was run individually over Clusters 1 and 3 (LOC and HUM noun clusters, respectively) from our 4-way clustering solution, as both clusters contained regular polysemous nouns.

Finally, we obtained a 2-way clustering solution for each class, aiming to differentiate nouns strictly containing the LOC sense and those reflecting the polysemy described above for *factory*, on one hand, and nouns in the HUM·HUMANGROUP·ORG ambiguity class from those strictly denoting HUM individuals on the other. Cluster 1 split into 2 clusters distinguishing between polysemous LOC nouns and those that are not, whereas for Cluster 3 the clustering algorithm arrived at a near perfect distinction of complex-type nouns and non-ambiguous HUM nouns. The noun *factory* clustered with polysemous HUM nouns, once more confirming its semantic proximity with nouns of the HUM·HUMANGROUP·ORG type.

Therefore, a second iteration of the same clustering algorithm over the same feature vectors was able to identify finer-grained distinctions within lexical classes, automatically recognizing groups of nouns in the same ambiguity class. In doing this, we validate our analysis regarding the role of regular polysemy and complex-type nouns in the clustering solutions obtained, and further strengthen our original hypothesis.

5.1.5 Final remarks

In this Section, we proposed using automatically obtained FORMAL role descriptors as features to draw together nouns from the same lexical semantic class in an unsupervised clustering task. In line with the results, our initial hypothesis was supported: we can cluster nouns using automatically extracted FORMAL role descriptors as features. Moreover, we showed that they were sufficient to discriminate between elements of different lexical semantic classes.

Furthermore, we validated that some misclassifications observed were caused by regular polysemy, which directly effected the results obtained by the distributional word representations. The method outlined in this Section demonstrates that it is possible to handle the polysemous behavior of nouns in classification tasks by making finer-grained distinctions regarding lexical items that consistently belong to the same ambiguity class. Additionally, we provided a formal description of noise being caused in many cases by cases of regular polysemy.

Finally, one of the main causes of sparse data in the experiments presented in this Section can be attributed to the fact that we just used the information extracted for the FORMAL role of each noun. Consequently, in subsequent Sections, we revert back to using the manually-identified linguistically-motivated lexical-syntactic patterns, defined in Section 4.1 to extract more frequently occurring class-indicative information.

5.2 A cascade approach to complex type classification

Based on the results obtained in Section 5.1, in this Section we propose a strategy to classify complex-type nouns. The classification of regular polysemous nouns is a real challenge for classifiers because all of the senses of a word are conflated in one representations and machine learning classifiers cannot distinguish between the different senses of a noun in context. The exclusive discriminative information to regular polysemous nouns, while it is extremely indicative to members that can instantiate a given alternation, such as co-predication or underspecified contexts, for instance [Pustejovsky, 1995, Pustejovsky, 2005, Pustejovsky, 2013, Rumshisky et al., 2007], occur with extremely low frequency infrequency in corpus data. Because of this, their inclusion is typically not useful in distributional word representations. To overcome this problem, we use lexico-syntactic patterns to automatically classify nouns for which there is distributional evidence of their membership to more than one class. In this way, regular polysemous complex-type nouns should demonstrate characteristic and indicative lexico-syntactic distributional characteristics for each class that forms a given regular polysemous alternation.

In line with the argument presented above, we focus on two complex types representative of the general characteristics of dot objects, ORGANIZATION·LOCATION and EVENT·COMMUNICATION_OBJECT [Pustejovsky, 1995, Pustejovsky, 2005, Rumshisky et al., 2007, Ježek and Melloni, 2011, Copestake and Herbelot, 2012]:

(18) ORGANIZATION·LOCATION
$(\lambda x \cdot y \exists [\alpha(\textsc{org}(x) \cdot \textsc{loc}(y)) \wedge R(x, y)])$

a. *the church prays during mass* (ORGANIZATION)

b. *the church is a large building* (LOCATION)

(19) EVENT·COMMUNICATION_OBJECT
$(\lambda x \cdot y \exists [\alpha(\textsc{evt}(x) \cdot \textsc{com}(y)) \wedge R(x, y)])$

a. *the interview lasted for two hours* (EVENT)

b. *the interesting interview in the book* (COMM._OBJECT)

As in the previous Section, for the work presented in this Section we formally assume the [Pustejovsky, 1995] definition of complex-type nominals as a Cartesian product of types with a particularly restricted interpretation. This means that the product $(t_1 \times t_2)$, of types (t_1) and (t_2), each denoting sets, alone does not adequately determine the semantics of the complex-type nominal. The relation R, which structures the component types, must also be seen as part of the definition of the semantics of the lexical conceptual paradigm of the complex type. Thus, for the complex type $(t_1 \cdot t_2)$ to be well-formed, there must be a relation R that structures the elements (t_1) and (t_2), a concept that is formalized in GL ([Pustejovsky, 1995]: 149) as demonstrated in the AVM in Figure 5.1.

This formalization accounts for one of the properties that makes complex types unique and distinguishes them, for instance, from cases of homonymy: the possibility for their distinctive senses to be active at the same time ([Pustejovsky, 1995]: 223), as previously illustrated in Example (11c). The levels of representation and generative mechanisms in GL predict that a noun like *church* occurs not only in contexts typical of class (x): ORG (see again Example 11a) and of class (y): LOC (see again Example 11b), but also in contexts which activate the relation $R1(x, y)$, i.e. contexts where both ORG and LOC senses are simultaneously

$$\begin{bmatrix} \textbf{church} \\ ARGSTR = \begin{bmatrix} ARG1 = \textbf{x : organizati on} \\ ARG2 = \textbf{y : location} \end{bmatrix} \\ QUALIA = \begin{bmatrix} \textbf{org} \bullet \textbf{loc\_lcp} \\ FORMAL = R_1(x,y) \end{bmatrix} \end{bmatrix}$$

Figure 5.4: AVM of the noun *church*

activated (see 11c).

We attempt to capture these properties through the design of a two-step methodology that can account for the possibility of complex-type nouns to distributionally belong to multiple lexical-semantic classes.

5.2.1 Verifying distributional information to identify complex-type nouns

We considered that complex-type nouns are members of more than one lexical-semantic class, more precisely members of each class that forms a given regular polysemous alternation. Thus, complex-type nouns should occur sufficiently in the indicative distributional contexts of each corresponding lexical-semantic class. Using the distributional information extracted from each relevant lexical-semantic class that forms a regular polysemous alternation, we build word representations to automatically classify complex-type nominals.

We hypothesized that the classification of a noun as a member of each individual classes that forms a given regular polysemous alternation is an indicator of its membership in that alternation. Parting from the cue-based nominal lexical semantic classification work reported in Section 4.1, we applied this methodology also to complex-type nominals, which allows us to analyze the distributional behavior of nouns belonging to more than one class and to which extent binary classifiers can accurately deal with such items.

Due to their ability to be selected for as a member of more than one lexical-semantic class, we expect complex-type nouns to occur in indicative contexts of different classes. Because of this, their occurrences in contexts indicative of each class that form a regular polysemous alternation can be infrequent because they occur a finite number of times in corpus data and, therefore, their distributional information is divided among all classes for which they occur in indicative contexts. Thereby, we also evaluate to which extent this dispersion can be problematic to binary classifiers. Specifically, we verify whether the available distributional information indicative of each individual class is strong enough for an automatic cue-based classification of complex-type nouns.

Experiments

We conduct two main experiments in this Section. We first verify the ability of our binary classifiers to identify complex-type nouns as members of the class corresponding to their more prominent sense, according to the results obtained from our human annotation task. To do this, we classified (t_1) nouns as a member of the (t_2) class, and vice versa. Along this line, we confirm whether a noun has sufficient distributional information for classification in both classes that form a given regular polysemous alternation through their successful classification as an member of each individual class. Then, in a second experiment, we automatically classify complex-types from simple types by training a dedicated classifier by combining the distributional information characteristic of each individual sense component of the complex type into a single classifier.

We used the human-annotated regular polysemy data sets, as described in detail in Section 3.1, to extract distributional information for corpus data. The lexico-syntactic patterns indicative of each individual class, defined in see Section 4.1 were used to gather distributional evidence for each target noun in our data sets from the UKWAC60M corpus. The relative frequency of occurrence of each noun in each cue was stored in an n-dimensional vector, where n is the total number of indicative cues used for each class. For classification, we used a Logistic Model Trees (LMT) [Landwehr et al., 2005] Decision Tree classifier in the WEKA [Witten and Frank, 2005] implementation, as introduced and described in detail in Section 4.2.

A baseline based on the majority class would not allow us to assess the quality of the results depicted here. Thereby, to evaluate our results, we compare them against the performance of State-of-the-Art classifiers for simple types, reported in Section 4.1.

Complex-type nouns as members of individual simple-type classes

The results reported in Table 5.6 make apparent that complex-type nominals provide enough distributional indicative evidence toward their most frequently occurring sense so that their automatic classification as members of each class of the alternation is possible. The results obtained are actually in line with the performance of the same classifiers with simple-type nominals reported in Section 4.1, where a 66.21% and a 73.05% accuracy are obtained respectively for the LOC and the EVT nouns classifiers.

Class	complex types correctly classified as members of the class (%)	ratio of classified complex types per members of the class
ORG·LOC as ORG	58.69	0.22
ORG·**LOC** as LOC	89.47	0.25
EVT·COM as EVT	71.11	0.43
EVT·**COM** as COM	77.78	0.03

Table 5.6: Complex types correctly identified as members of the class corresponding to their frequent sense, in bold

With this in mind, we proceeded to verify whether this is also observed when considering the less frequent sense components of the complex-type noun by performing a cross-classification of the nouns using the binary classifiers mentioned above. More precisely, we used trained binary classifiers for each class to classify the human-annotated lists of nouns, i.e. each classifier trained for simple-type classification of nouns of semantic type (t_1) was provided with a list of nouns with (t_2).

In this way, the cross-classification experiment consisted of training a classifier with simple-type classification of (t_1) nouns and testing it with (t_2) nouns and vice versa. To further illustrate this, a noun like *church*, defined as a LOC (t_1), was checked for its occurrences in lexico-syntactic patterns indicative of ORG (t_2) nouns to determine whether it shows distributional evidence indicative of another class. Our claim is that (t_1) nouns that sufficiently occur in contexts indicative of t_2, resulting in an accurate classification as a member of (t_2), confirms that those nouns are members of more than one lexical semantic class, a fact that automatic classifiers should account for.

Table 5.7 presents the results of precision and recall of this cross-classification, indicating the less frequent sense component in bold.

Class	Precision	Recall	Ratio
ORG·**LOC** as ORG	57.14%	21.05%	0.06
ORG·LOC as LOC	77.78%	15.21%	0.06
EVT·COM as COM	6.67%	66.67%	0.03
EVT·**COM** as EVT	64.44%	32.22%	0.19

Table 5.7: Results of cross-classification

The results of the cross-classification verified that the information available generally indicates that certain nouns demonstrate a distributional behavior of members of more than one class, although the information available for each class may not be enough to correctly classify a part of the nouns studied, indicated by the low recall. From the results in Table 5.7, we made three main observations:

- The performance of cross-classification is in line with that of the classifiers used when dealing with simple-type nominals and when classifying complex-types nouns as members of the class corresponding to its most prominent sense component. This indicates that complex-types nouns do occur in contexts typical of the different classes corresponding to their sense components, i.e. they belong to more than one class and behave as such.

- The overall low recall indicates an imbalance of information regarding one of its class, which is consistent with the work of [Rumshisky et al., 2007] and the above discussion. Specifically, this hold true in regards to asymmetries in terms of frequency of the different meaning components of complex types. This is primarily reflected in the frequency of occurrences in contexts indicative of a given class, which represents the information provided to our classifiers of each class. The noun *church*, for instance, occurred in contexts typical of LOC nouns with a relative frequency of 0.015 and of 0.030 in contexts typical of ORG nouns. This is also the case of the noun *jurisdiction*, which occurred with a relative frequency of 0.039 in contexts typical of ORG nouns and just 0.014 in contexts typical of LOC nouns. This provides evidence that a noun occurs more frequently in the contexts indicative of one class, which is bound to affect classification results, particularly when the asymmetry is large. Thus, the asymmetric representations of senses impacts the recall, in particular, of our classification results, because there is insufficient distributional evidence towards class membership for an important part of nouns in our list.

- The absolute numbers are lower due to the aforementioned recall, the ratio of complex types per class shows similar tendencies to those observed in the results of the human annotation of the polysemous data set. In fact, the ratios of complex types for the ORG and LOC data sets are balanced, along the lines of the human annotation results (see Appendix A.2, whereas a bigger asymmetry is observed for the COM and EVT classes.

Finally, given our objective to verify whether complex-type nominals provide distributional evidence concurrent with more than one semantic class, our cross-classification experiment confirms that the distributional information available generally indicates that complex types demonstrate a distributional behavior typical of members of more than one class, though the information available is not always sufficient enough to correctly classify a part of the nouns studied.

However, in this experiment we only considered a part of the distributional data for each complex type at a time. Having demonstrated that complex-type nouns show distributional behavior typical of members of more than one class, we propose to include indicative contexts of each of the classes that form a regular polysemous alternation to the classifier, thus this way accounting for its full sense potential in one dedicated vector.

Distinguishing complex types from simple-type nouns

The experiments described above demonstrate that the distributional evidence of a complex-type noun can be indicative of one class, yet the information available is often not sufficient for automatic systems to perform accurately and robustly. Thus, we put forth a new experiment to classify complex types built upon these observations. The cross-classification experiments used the distributional information available for each word in contexts indicative of each class corresponding to one of its senses individually. In this experiment, we combine into one vector the contextual cues indicative of the individual classes that a the regular polysemous alternation.

Along this line, we collected distributional evidence of nouns by simultaneously using the cues for each class corresponding to the different sense components of the complex types considered in this work. As in the experiment presented in Section 4.2, we used the LMT classifier [Landwehr et al., 2005] in a 10-fold cross-validation setting in the WEKA [Witten and Frank, 2005]. Table 5.8 presents the results of the classification of ORG·LOC and EVT·ORG

	Accuracy	Precision	Recall	$F1$-Score
ORG·LOC	67.68%	0.62	0.67	0.62
EVT·COM	78.75%	0.72	0.78	0.72

Table 5.8: Results of complex-type classifiers

The results presented in Table 5.8 demonstrate that by combining indicative cues of different individual semantic classes and thus providing distributional evidence of the entire sense potential of a complex-type to the classifier, we are able to automatically classify complex types, by distinguishing them from simple-type nominals. As in the previous experiment, in order to be distinguished from simple-type nominals, complex types must demonstrate sufficient distributional evidence in contexts that are indicative of classes corresponding to their different sense components. By combining the distributional information that is indicative of both classes in one vector, we improve the results previously obtained and attain accuracy in line with State-of-the-Art simple-type classifiers (see Section 4.1 results regarding nominal lexical semantic classification in English).

We observed that the difference of more than 10% of accuracy between the classifiers for both complex types considered. Previously discussed work by [Ježek and Melloni, 2011] helped us identify possible causes for these contrasts, such as an ontological dependence between components of complex-type nouns like EVT·COM, whose occurrences have both sense components of the complex type generally simultaneously present. However, the same is not true for complex-types nouns such as ORG·LOC nouns, which results in a more disperse distributional behavior between indicative contexts of each sense component of the complex type, constituting a challenge for classifiers, naturally impacting performance.

Discussion

In order to discuss the results presented in this Section, we first comment on the data set that we used. We recall that the polysemy information available in the data set was obtained by a human annotation task, as described in Section 3.1. When analyzing the results of that task, we observed that there was an asymmetry of the sense distribution of the polysemous items. Previous work has reported asymmetries regarding the difference of the selection of senses in context that compose complex types (see, for example, [Rumshisky et al., 2007] and [Ježek and Melloni, 2011]), especially as one sense is generally more frequently used or constitutes a preferred interpretation. Confirming this observation, evidence

$$\begin{bmatrix} \textbf{illustration} & \\ ARGSTR = & \begin{bmatrix} ARG1 = \mathbf{x : event} \\ ARG2 = \mathbf{y : information} \end{bmatrix} \\ QUALIA = & \begin{bmatrix} \tau_1 \bullet \tau_2\_\mathbf{lcp} \\ FORMAL = R\,(x,y) \\ AGENTIVE = x(z,y) \end{bmatrix} \end{bmatrix}$$

Figure 5.5: AVM of the noun *illustration*

from psycholinguistic studies [Frisson, 2009] also claim that although more than one sense interpretation is available for a given word, one interpretation is consistently preferred over the other.

Several authors have established relations between this type of asymmetry and complex types, particularly with regard to the nature of the relations holding between their sense components. An important part of the work developed on this matter has focused on classes whose sense components are ontologically related, in particular on the PROCESS·RESULT complex-type. [Ježek and Melloni, 2011] characterized the properties of the polysemy involved in this case arguing it arises from the fact that a RESULT object type is temporally and causally dependent on a PROCESS type as an event is the pre-condition for the (coming into) existence of the object (RESULT). Thus, PROCESS readings can be considered more prominent as they are also reflected when the RESULT sense is active while the reverse does not hold true. The EVT·COM complex type, can be considered a sub-case of the former. Formalized in the AVM presented in Figure 5.5, the aforementioned unique properties of this complex type are represented in the AGENTIVE role.

Just as is the case for PROCESS·RESULT nominals, we expected the prominence of senses for this complex type to be asymmetric. The data obtained in our annotation task are consistent with this expectation (see Tables A.7 and A.6), as 90 of the 239 COM nouns in the data set are considered to also have an EVT sense, whereas only 9 of the 239 EVT nouns are annotated as also having an COM sense. Moreover, these human annotation results constitute a source of quantitative information that provide evidence that support the existence of asymmetries of prominence of the different sense components of complex types.

Regarding the LOC·ORG complex type, there is neither an ontological relation between its meaning components nor such a clear asymmetry in the prominence of its sense components. Yet, differences observed can be attributed to relations generally holding between objects in the world. For instance, an ORG, as a more abstract concept, is typically associated to a physical reality, namely the LOC which hosts this abstract object and makes it *perceivable*. Reversely, LOC, as a physical point in space, is often independent of any other reality. Thus, in the lexicon, we observe words primarily denoting an ORG that also refer to the LOC that hosts it, whereas the reverse is observed only in considerably stricter conditions, as illustrated by *congress* and *schoolyard* in Example (20).

(20) nouns: *congress* and *schoolyard*
   a. The congress (ORG) decided to vote the new rule into power after the recess.
   b. #The schoolyard decided to vote the new law into power after the recess.
   c. The new rule was voted to power in the schoolyard (LOC).

Asymmetry can, therefore, be said to be related to the nature of the systematic relation holding between them, which is different for each complex-type paradigm. Moreover, the ratio of nouns in each individual class annotated as having more than one potential sense, makes apparent the representativity of this phenomenon for each class. This provides crucial insight when evaluating our results, particularly in order to determine whether the asymmetries reported in this Section have an overall impact in the automatic identification of complex types.

Final remarks

The strategies proposed in this Section were able to automatically identify nouns that display characteristic properties of different simple types, namely LOC and ORG, and EVT and COM in spite of the strong biases that asymmetry has imposed on our data set. This is a very important point to note because it provides further evidences toward the treatment of a dedicated lexical-semantic class for nouns that instantiate regular polysemous alternations. This is because the nouns must be classified as members of each class that composes the regular polysemous relation, regardless of the bias a noun may have for one of the classes that instantiates that relation. By achieving this, we demonstrate the validity of our hypothesis that complex-type nouns can sufficiently display distributional characteristics of the different classes that form a regular polysemous alternation in one distributional vector.

Moreover, our cue-based lexical semantic classification methodology obtained an average overall performance of more than 70% when distinguishing complex types from simple-type nouns belonging to semantic classes that corresponds to any of the sense components of the former. Yet, these results are based on the classification of a complex-type noun from its simple-type counterparts. In the next Section, we expand upon this approach to demonstrate that this method can be adapted to also identify complex-type nominals from a given class by distinguishing them from any other noun in the language. Accomplishing this requires extending our approach to be able to not only separate complex-type nominals from simple-type nouns belonging to one of the classes corresponding to one of the sense components of the former, but to distinguish nouns belonging to a given complex-type class from any noun in the language, independently of the class to which they belong.

Finally, a task to consider for future work is the design of a strategy to also incorporate of contexts specific to complex types, i.e. contexts which *convoke* different sense components simultaneously (see, for instance, [Simon and Huang, 2010]; [Pustejovsky, 2013]; [Cruse, 2000]) into feature vectors, for a still more reliable classifier.

5.2.2 Implementing a 2-Step cascade classification approach

Based on the results and observations made in the previous Section, we expanded upon the approach proposed and designed a 2-step Cascade Approach to classify complex-type nouns from any other noun in language. As mentioned, the Cascade Approach consists of 2 steps:

- Step 1: to distinguish (x/y) group nouns from any other noun in the language for each polysemy alternation;
- Step 2: to take the nouns classified as belonging to the (x/y) group in Step 1 and distinguish them simple-type nouns from complex-type nouns.

Figure 5.6 outlines the workflow for each of the defined steps in our proposal for the Cascade Approach.

The Cascade Approach classifies consists of complex-type nouns in 2 steps. Having previously experimented with a single-step classification systems in Chapter 4 and Sections 5.1 and 5.2.1, the results obtained made apparent that the

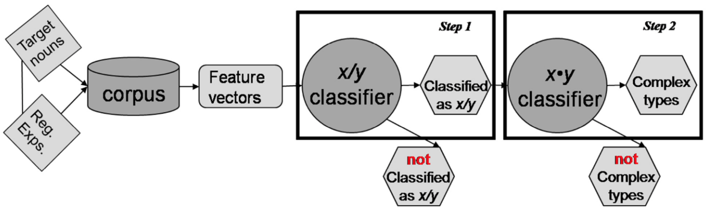

Figure 5.6: Workflow of the proposed approach for complex-type classification

nuanced distinctions a complex-type nominal classifier has to perform requires a different approach. Because complex-type nouns correspond to a very specific and complex linguistic phenomenon, with a strong impact in terms of semantic behavior in context, automatic systems have difficulty to accurately model them. More specifically, the characteristic properties of this type of nouns causes their distributional data to be more disperse, besides partially overlapping with that of the simple-type nouns that correspond to one of the sense components that form a regular polysemous alternation, which further raises problems to any automatic classification system.

These observations led us to search for an alternative approach to the problem of complex-type nominal classification, namely the design of a Cascade Approach that consists of two steps in order to ensure for the accurate classification of complex-types.

Step 1: Distinguishing nouns in the (x/y) group from any other noun:

The first step of the Cascade Approach consists of training a classifier to distinguish nouns of the (x/y) group from nouns of any other class in language. Along this line, we consider all $(x \cdot y)$ complex-type nouns (either LOC·ORG or EVT·COM, in the case of the classifiers discussed in this section), as well as the simple-type nouns that correspond to one of the classes that form the regular polysemous alternation. In other words, we classify each of the components of the complex-type classes (LOC and ORG and LOC·ORG; or EVT and COM and EVT·COM) into the (x/y) class. Thus, the goal of this step is to coarsely distinguish nouns belonging to the (x/y) group from nouns belonging to any other lexical semantic class.

Step 2: Identifying complex-type nouns with a ($x \cdot y$) classifier:

The goal of the second step of the cascade experiment is to distinguish ($x \cdot y$) complex-type nouns from simple-type (x) and (y) nouns. In this step, we use the output of Step 1 as the input to be classified. More precisely, we classified those nouns that were classified as members of the (x/y) class in Step 1. Testing our complex-type classifiers with this information allows us evaluate their ability to really capture complex-type nouns (i.e. identifying, on the one hand, LOC·ORG nouns and, on the other hand, EVT·COM nouns), as they have to deal with the noisier input consisting of (x/y) group nouns as identified by an automatic system whose average accuracy scores are in the mid-70% (see Table 5.8).

Experiments

Following the experimental design in Section 5.2.1, we gathered distributional data from the UKWAC60M corpus for each word in the human-annotated polysemous data set, described in Section 3.1, using the lexico-syntactic patterns, defined in Section 4.1, that correspond to each class considered. To build the word-representations for the complex-type classifier, we again combined the distributional information extracted from each class into one vector per word to provide to the classifier.

More specifically, to extract distributional information indicative of each (x/y) group, we combined the features indicative of class (x) with the features indicative of class (y), i.e. we combined class-indicative features of LOC and ORG, in the case of the LOC/ORG classifier, and indicative cues for the EVT and COM classes in the case of the EVT/COM classifier. The relative frequency of occurrence of each noun in each cue was stored in an n-dimensional vector, where n is the total number of cues used for each class. To classify, we used a Logistic Model Tree (LMT) [Landwehr et al., 2005] classifier in the WEKA [Witten and Frank, 2005] implementation, as introduced in Section 4.2.

For the purpose of training a classifier and testing it with unseen data, we divided our full data set into training and test sets (70% for training and 30% for test). The experimental results reported and discussed in the following sections are based on the results obtained when considering balanced data sets for training, whose constitution is presented in Table 5.9.

Thus, in order to implement Step 1 of the Cascade Approach, we trained two classifiers: for the (LOC/ORG) group and for the (EVT/COM) group, with a

Balanced Datasets						
	LOC·ORG			EVT·COM		
type	S	C	not LOC/ORG	S	C	not EVT/COM
training	56	56	112	68	68	136
test	128	23	211	315	31	356

Table 5.9: Distribution of nouns in training and test data sets for the complex-type classes considered in this experiment: C corresponds to complex-type nouns, S to simple-type nouns either of class (x) or (y), while the not (x/y) corresponds to nouns not belonging to the (x/y) group considered (see Footnote 1)

supervised LMT classifier, using 70% of our original data set in a balanced selection of data, as detailed in the previous Section (see Table 5.9). Each training classifier was trained in a 10-fold cross-validation setting. The (x/y) classifier model for each regular polysemous alternation was then tested on unseen data (i.e. the remaining portion of the original data set - cf. Table 5.9).

To implement Step 2 of the Cascade approach, we used the output of the classification of the test set in Step 1, more specifically, we used those nouns that were classified to be a member of the (x/y) class, whether it was a correct classification or not. We then classified those nouns using the trained ($x \cdot y$), or complex type, classifier, again, using a supervised LMT classifier.

Results

Table 5.10 presents the results regarding the performance of the classifiers used in the cascade experiment, both with training and test data, and for the two complex-type classes considered. As was to be expected, the performance of the complex-type classifiers in the training setting is consistent with the results reported in Section 5.2.1. Though slightly lower on the test setting, there is no statistically significant difference in the overall performance of the complex-type classifiers in the training and test settings, i.e. in a 10-fold cross validation setting and when used to classify the output of either the LOC/ORG or the EVT/COM group classifiers.

Our expectations were also confirmed by the results obtained and presented in Table 5.10. On the one hand, the overall precision of a 2-step classification system significantly improves when compared with that of a single-step approach to this problem. On the other hand, the automatic separation of items, which is inherent to the cascade approach, still have a negative impact on recall, although it was not statistically significant. However, this result crucially indicates an important

Step 1 of the cascade experiment: x/y **group classification**				
	LOC/ORG group classifier			
	accuracy	precision	recall	$F1$-Score
training set	74.55%	0.75	0.75	0.74
test set	75.69%	0.76	0.76	0.75
	EVT/COM group classifier			
	accuracy	precision	recall	$F1$-Score
training set	72.79%	0.73	0.73	0.73
test set	69.81%	0.71	0.69	0.69
Step 2 of the cascade experiment: complex-type classification				
	LOC·ORG complex-type classifier			
	accuracy	precision	recall	$F1$-Score
training set	60.71%	0.61	0.59	0.59
test set	57.14%	0.88	0.57	0.67
	EVT·COM complex-type classifier			
	accuracy	precision	recall	$F1$-Score
training set	59.56%	0.59	0.59	0.59
test set	56.69%	0.91	0.57	0.67
Overall Score of the cascade experiment: complex-type classification				
	LOC·ORG complex-type classifier			
	accuracy	precision	recall	$F1$-Score
training set	67.63%	.68	0.67	0.67
test set	66.42%	0.82	0.66	0.71
	EVT·COM complex-type classifier			
	accuracy	precision	recall	$F1$-Score
training set	66.18%	0.66	0.66	0.66
test set	63.25%	0.80	0.63	0.68

Table 5.10: Performance of classifiers in Step 1 and Step 2 of the Cascade Approach to classify complex-type nouns

noise reduction, which actually increases the reliability of the complex-type classification decisions made by the system.

Furthermore, the results obtained with the (LOC/ORG) group classifier, as well as the results obtained with the (EVT/COM) group classifier are consistent and promising: precision and recall are generally above 70% and there are no statistically significant differences between the performance of the classifier in a 10-fold cross validation training setting and when the classification models are confronted with an input of unseen data in the test setting. However, with regard to the results obtained in the second step of this experiment, further discussion is required.

Discussion

The considerably higher precision of the complex-type classifier in the test setting when compared with the results obtained in the training setting has, nonetheless, to be underlined and commented upon. Overall, this seems to indicate that the complex-type classifier successfully handle instances corresponding to noise proceeding from the first step of the cascade experiment, ubiquitous in any production-level scenario. Moreover, it also indicates that, although one of the concerns with using a cascade approach was the possibility of error accumulation, the results obtained, and the significant increase in precision in the classification of complex-type nouns in particular, point towards the opposite. More precisely, these results indicate that the information provided to the $(x \cdot y)$ classifiers is somehow cleaner. Yet, this is a direct reflection of the Cascade Approach itself. As we are using the output of Step 1 as the input for Step 2, the data set is no longer equally balanced between members of the class and nouns that are not members of the class, resulting in the higher numbers of precision reported.

We attribute these results to the increase of the distributional profile of nouns to be classification in the (x/y) step in Step 1, which can introduce higher levels of "noise" and result in the misclassification of a noun as a member of the (x/y). However, this smaller number of nouns used in Step 2 results in a higher precision, mainly due to the fact that there are less nouns to classify in the data set. Furthermore, it is interesting to note that this increase in precision does not have a relevant impact on recall: although the scores are slightly lower in the test setting, the difference between the recall scores in the training and test settings is not statistically significant.

Complex-type classification of automatically identified (x/y) group nouns from Step 1: impact on precision

In order to evaluate to which extent the first step of our Cascade Approach is actually capturing nouns that meet the criteria of the (x/y) class, we also re-ran our complex-type classification with the full test set, i.e. as if the first step of the cascade workflow was performing with an accuracy of 100%, and compared the results obtained. By doing this, we aimed to identify what types of nouns are being eliminated in the first step of our Cascade Approach to verify whether the candidates that we are losing would be correctly dealt with by our complex-type classifiers. In this context, we observed that an important part of the nouns being eliminated in Step 1 are nouns that occur in corpus data with a low frequency.

On the one hand, in the case of the (EVT/COM) classifier, 7 of the 11 nouns misclassified as not belonging to the (EVT/COM) group, and thus not included in the set of candidates provided as input to the (EVT·COM) complex-type classifier in Step 2, occurred with an absolute frequency of less than 200 times in the corpus. In fact, of those 7 nouns, 5 occurred with an absolute frequency of less than 20. Thus, in this case, the generalization of the distributional profile to the (x/y) class is not capturing some low-frequent nouns in Step 1. On the other hand, in the case of the (LOC/ORG) class, the absolute frequency of 6 of the 12 misclassified nouns not considered to belong to the (LOC/ORG) group was lower than 200 occurrences in the corpus, while the absolute frequency of 3 of those 6 nouns being lower than 20. Thus, as a large part of the misclassifications observed in the first step of the Cascade Approach is due to low frequency of occurrence in corpus data, it is bound to also affect the classification decisions in the second step. To further explore the cause of this, we submitted all the nouns misclassified in the first step of the cascade workflow to the complex-type classifiers in the second step to determine to what extent they are still able to successfully classify such candidates.

We obtained the following results: in the case of the (EVT·COM) class, 9 of the 11 nouns eliminated in the Step 1 of our cascade experiment would still be misclassified in Step 2 if they were to arrive to this step of the experiment, the 7 low-frequency nouns mentioned above being among these 9. In the case of the (LOC·ORG) class, the same would happen to 6 of the 12 nouns misclassified in Step 1, the overlap between the set of low-frequency nouns and that of misclassified complex-type nouns by the (LOC·ORG) classifier being perfect. These data make apparent that the increase in precision in Step 2 is directly explained by the fact that there are low-frequent complex-type nouns misclassified in Step 1, which results in a smaller amount of candidate nouns to the $(x \cdot y)$ classifier in Step 2 and has a direct impact on its performance in terms of an increase in precision, on the one hand, but at the cost of recall, on the other hand. However, we also consider that the inclusion of these nouns in Step 2 would have increased recall at the cost of precision, which is not desirable for the classification of nouns representative of a complex-semantic phenomenon, such as regular polysemy.

Naturally, not all of the nouns misclassified in Step 1 are necessarily "problematic". For instance, we observed 2 cases of EVT·COM nouns and 6 cases of LOC·ORG nouns that are incorrectly classified in Step 1, and therefore not considered in Step 2, although they would have been correctly classified, which further reduced the coverage of our classifiers and, moreover, impacted the potential increase of precision. But this is not the only aspect determining the scores of our classifiers in terms of recall, which is clearly the weak aspect of the classifiers developed. In order to further understand what impacts the recall scores obtained in complex-type classification we conducted an error analysis, which we discuss below in detail, focusing on those nouns that were misclassified as not belonging to a complex type by our classifiers.

Analyzing the recall of x·y classifiers from Step 2

In the case of the (EVT·COM) classifier, the final results obtained in Step 2 demonstrate 5 incorrectly classified complex-type nouns, which are considered to be non-members of the (EVT·COM) class by our classifier. Of these 5 cases, the noun *newsflash*, is the only one caused by insufficient distributional information (15 occurrences in total of this noun in corpus data). Due to its low frequency in our data, this noun only occurs 3 times with our class-indicative lexico-syntactic patterns, which did not provide sufficient information for the complex-type classifier to arrive at an accurate class membership decision for that particular noun. We have to underline that the complex-type classifiers in Step 2 must make more nuanced decisions, distinguishing between complex and simple-type nouns. This is because the complex-type nouns should display characteristic features of both class (x) and class (y) for an accurate classification decision, which makes the availability of sufficient class-indicative distributional information all the more important.

As to the 4 remaining cases of incorrectly classified complex-type nouns, their misclassification cannot be attributed to low frequency, as these nouns have an absolute frequency of 190, 3881, 1779 and 538 times in the corpus. However, when looking into their individual feature vectors, we observed that the information being provided to the classifiers demonstrated considerable asymmetry in terms of the frequency of use in language data of the different sense components of the complex type. When considering the distributional data as represented in the feature vectors of each of these complex-type nominals, we observed, for instance, that in the case of the complex-type noun *notice* 613 of its 697 co-occurrences with class-indicative lexico-syntactic patterns corresponded to features that are indicative of the COM class, while only 42 were indicative of the EVT class, and another 42 occurrences corresponded to negative cues. This same trend was also observed with the (EVT·COM) noun *quote* for which 123 of its 130 occurrences were in COM-indicative patterns, only 5 being in EVT-indicative patterns, and 2 with negative cues.

This way, we attribute misclassification in these cases to context behavior that is reflective of more than one lexical semantic class, which was reflected in the representativity in corpora data of the features indicative of the different sense components of these particular lemmas. This point is further verified by the fact that these lemmas are correctly classified in Step 1 as members of the (EVT/COM) group, as the classifier is trained to identify nouns from each of the individual simple-type classes that form a regular polysemous alternation. Thus, even though there is an asymmetry in the frequency of use in language data of the distributional information represented in the feature vector of a complex-type noun provided to our classifiers, which can cause its misclassification as a non-member of the $(x \cdot y)$ complex type, these nouns are not misclassified in the first step of our experiment as they have a significant number of features in common with nouns of one of the simple-type classes being considered by the classifiers in this step, and are therefore correctly classified as members of the (x/y) group.

In the case of (LOC·ORG), the final results obtained in Step 2 demonstrate 4 incorrectly classified complex-type nouns, which were considered not to belong to the LOC·ORG class by our classifier. Of these 4 cases, one is caused by insufficiency of distributional information (23 occurrences in total of this noun in corpus data) while the remaining 3 cases also displayed an asymmetry of occurrences in class-indicative lexico-syntactic patterns of the different sense components of the (LOC·ORG) complex type.

In the case of the LOC·ORG noun *borough*, 37 of its 54 occurrences in class-indicative lexico-syntactic patterns are indicative of the LOC class, while only 14 of its occurrences are class-indicative features for the ORG class. This same trend was also observed with the LOC·ORG noun *unit*, for which 339 of its 387 occurrences corresponded to features considered indicative of the LOC class, while only 38 were features considered indicative of the ORG class, and 10

amounted to negative cues. The same was also true for the LOC·ORG noun *agency*, which has 189 of its 286 occurrences in features indicative of the LOC class and only 33 in features of the ORG class, while 14 occurrences corresponded to negative cues.

These examples further serve to demonstrate the impact of asymmetry in the frequency of use of different sense components of a complex-type noun can have on results, as discussed in detail in Section 5.2.1. In fact, although theses nouns are considered to be complex-type nominals in our data set, their distributional data is still heavily biased towards one of the two sense components of the complex type.

Final remarks

The Cascade Approach presented in this Section confirms that we can obtain State-of-the-Art results when running a 2-Step complex-type classification on a data set consisting also of nominals that belong to any lexical semantic class, in contrast to the work presented in Section 5.2.1.

Overall, our approach can successfully identify very specific lexical items such as complex-type nominals with high accuracy, and distinguish them from those instances that are not complex types using a combination of the lexico-syntactic patterns indicative of each classes corresponding to the different sense components that form a regular polysemous alternation to build word representations. Moreover, the Cascade Approach increases the precision of the complex-type nominal classification, further providing evidence of recurrent contextual characteristics that are distinctive to regular polysemous nouns.

Due to the success we have has using the WE representations in Section 4.4, including overcoming problems of sparse data in word representations for classification, which resulted in higher quality classification decisions. In the final Section of this Chapter, we used again WE representations to expand upon our Cascade Approach as an attempt to overcome the reported effects of asymmetry in the representations obtained using the LING model. Furthermore, we also extend our approach to the larger CoreLex [Buitelaar, 1998], as described in Section 3.1, which will increase the number of regular polysemous alternations that we classify to 60, providing more evidence for our conclusions.

5.3 Using WE representations to classify complex-type nouns

As introduced in Section 5.2, and in parallel with the timeline provided in Chapter 4, the last experiments conducted for this thesis used WE distributional word representations for classification.

In this final experiment, we implemented the Cascade Approach proposed in Section 5.2 using WE word representations, as defined in Chapter 4.4. In this Section, we used WE representations because they demonstrated to more aptly handle the problem of sparsity in comparison to representations produced with count-context models, such as the LING model. The success of the Cascade Approach to classify complex-type nouns of a given regular polysemous alternation depends on the ability of the classifier to handle and classify low-frequent lexical items that consequently have sparser data in their vectorial representations.

Moreover, in this Section, we used the CoreLex data set, built by [Boleda et al., 2012a] and described in more detail in Section 3.1, to increase the number of complex-types studied and provide further empirical evidence for the approach. Furthermore, the inclusion of more regular polysemous alternations will confirm the transferability of this Approach to all complex-type nouns.

5.3.1 WE for complex-type classification

As detailed in Section 5.2.2, the Cascade Approach consists of 2-steps that account for all of the relevant classes of a complex-type nominal, by first distinguishing target nouns from any other noun found in the corpus not belonging to any class that contributes to the sense alternation, i.e. as members of the (x/y) class, as described in Section 5.2, and second, by distinguishing those polysemous nouns $(x \cdot y)$ from their monosemous counterparts (x) and (y), i.e. as members of the $(x \cdot y)$ class.

Our goal with using WE representations in the the Cascade Approach is to improve the quality of classification decisions for complex-types nominals. As the Cascade Approach has demonstrated to be effective even though frequency is an issue to handle instances of when frequency *is* an issue because it focuses on identifying those low frequent items first as members of the super class of the regular polysemous alternation (i.e. those nouns that are members of x, y and/or $x\cdot$). However, as we saw in the previous Section, this resulted in classification decisions with high precision and low recall. In this Section, with the use of WE representations, we aim to reduce the issues of recall while at least maintaining, or increasing, precision. Furthermore, we hypothesize that the use of WE representations can serve to provide further empirical evidence to support our conclusions regarding the reduction of sparse data in Chapter 4.

5.3.2 Experiments

Each experiment conducted in this Section was conducted using the disemous classes indicated in the CoreLex data set [Boleda et al., 2012a], as described in Section 3.1. This data sets provided gold standard information regarding nouns belonging to specific regular polysemous alternations.

Following the methodology already proposed in previous Section and the strategy of [Boleda et al., 2012a], we based our experiments on each of the 60 disemous alternations defined in the CoreLex data set (i.e. alternations that consist of only two different senses), such as (LOC · ORG) or (COM · EVT). As previously explained, [Buitelaar, 1998] designed the alternations in the CoreLex resource using a frequency criteria, which did not account for other types of lexical ambiguity, such as homonymy. In order to validate that the regular polysemous alternations being studied are exemplary, we also conducted a human annotation task to identify which of the disemous alternations are considered to be prototypical regular polysemous alternations. A description of the annotation task, as well as the details of the results are presented in Appendix A.3.

Each experiment was conducted with the 3 billion token LARGE3BN corpus, as described in Section 3.2. We trained our classifiers with WE representations built with dependency-based word embeddings [Levy and Goldberg, 2014a], using the same method as described in Section 4.4. A binary classification was conducted for each regular polysemous alternation using again a supervised LMT [Landwehr et al., 2005] classifier in the WEKA implementation [Witten and Frank, 2005], as previously introduced and described in Section 4.2. Each classifier was evaluated in a 10-fold cross validation setting.

In order to evaluate the approach using WE representations, we conducted two experiments:

- **Experiment 1** - implementation of the Cascade Approach:

- **Step 1**: We classify the nouns in Step 1 into a super class (or (x/y) class), which classifies any noun that is related to a class that participates in the alternation from any other noun, following the methodology outline in Section 5.2.

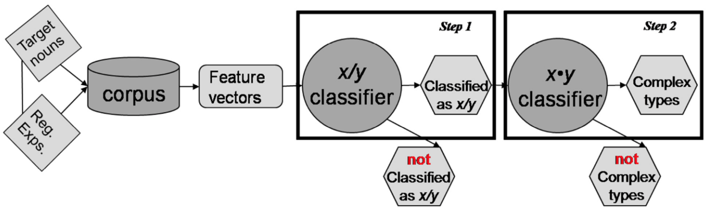

Figure 5.7: Workflow of Two-Step Cascade Approach (Experiment 1) using WE representations

- **Step 2**: Using the nouns classified to be a member of the (x/y) class in Step 1, we classify nouns as members of their own class $(x\dot{y})$. In this way, the goal is for the classifier to separate them from their potentially monosemous or homonymous counterparts. Figure 5.7 illustrates each Step in Experiment 1.

- **Experiment 2** - Direct classification of complex-type nouns:

  We directly classify the regular polysemous nouns into their own separate lexical semantic class, thus eliminating the 2-step process. In this way, the goal of this experiment is to directly classify nouns of a regular polysemous alternation in their own lexical-semantic class $(x \cdot y)$.

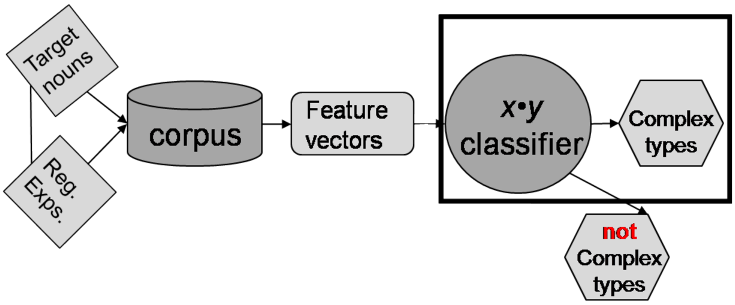

Figure 5.8: Workflow of Experiment 2 using WE representations

In this way, the classifier would directly separate complex-type nouns from all other nouns in a corpus including their monosemous or homonymous counterparts. Moreover, this experiment provides further evidence toward the existence of regular polysemous alternation as a lexical semantic class. Figure 5.8 illustrates the workflow for the direct classification Experiment 2.

5.3.3 Results

Table 5.11 presents the average overall results obtained from the 60 disemous classes of the CoreLex data set [Boleda et al., 2012a]. Table 5.12 presents the average results obtained from nouns of those regular polysemous alternations those classes that had 100% agreement between the human annotators in our annotation task. The results obtained using WE representations are quite promising and demonstrate a clear improvement upon the results previously presented in Section 5.2.2, leading us to form new conclusions regarding complex-type nominal classification.

Cascade Classification Experiment				
	accuracy	precision	recall	$F1$-Score
Step 1: x/y group classifiers: Experiment 1	74.66%	0.74	0.75	0.75
Step 2: $x \cdot y$ classifiers: Experiment 1	71.72%	0.70	0.71	0.70
Overall Score: Cascade Classifiers:	73.19%	0.72	0.73	0.73
Direct Classification Experiment				
	accuracy	precision	recall	$F1$-Score
$x \cdot y$ classifiers Experiment 2	82.55%	0.80	0.82	0.81

Table 5.11: Average results from the CoreLex Cascade Classification experiments over all 60 disemous polysemous classes

As indicated in the results presented in Table 5.11, Step 1 verified that we are able to successfully classify nouns into a super class (x/y) of regular polysemous with an average of almost 75% accuracy. This result is also consistent with the results reported for Step 1 in Table 5.12, although there is a slight overall improvement, which indicates that the Cascade Approach has a stronger performance when considering classes that are actually representative of this phenomenon.

Cascade Classification Experiment				
	accuracy	precision	recall	$F1$-Score
Step 1: x/y group classifiers: Experiment 1	76.41%	0.76	0.76	0.76
Step 2: $x \cdot y$ classifiers: Experiment 1	73.52%	0.73	0.73	0.73
Overall Score: Cascade Classifiers:	75.09%	0.75	0.75	0.75
Direct Classification Experiment				
	accuracy	precision	recall	$F1$-Score
$x \cdot y$ classifiers Experiment 2	86.09%	0.85	0.86	0.85

Table 5.12: Average results from the CoreLex Cascade Classification experiments over those disemous polysemous classes annotated by human annotators

The results of Step 1, also demonstrated an improvement over the results obtained with this method from count-context models, as presented in Section 5.2.2. Besides the average increase in accuracy over the test set, we also, more importantly obtained an overall average increase in recall of approximately 20 points, when considering the recall of the prototypical regular polysemous classes and $approximately 10$ points when considering the recall of all 60 disemous classes. This increase in recall further indicates that the use of WE representations drastically reduces the effects sparsity in vectors provided to the classifiers, especially in the case of more low-frequent items that we cannot classify with count-context models, which directly improves the performance of the classifier.

In the case of the $(x \cdot y)$ classifiers in Step 2 of Experiment 1, we achieved an average accuracy of 71.72 in the case of all 60 disemous classes and an average accuracy of 74.39 in the case of the annotated regular polysemy alternations. These results are also important because they indicated that the distributional representations of regular polysemous items is sufficient enough not only for an accurate classification decision, but also sufficient enough to directly discriminate these nouns from all other nouns that do not instantiate both classes. Hence, the experiment confirms that the distributional characteristics of complex-type nouns can be learned by a classification system.

More importantly, the results obtained in Experiment 2 *are* statistically significant from the final results obtained in Experiments 1, with a 10 point increase in the $F1$-Score and a 12 point increase in accuracy in the results of the prototypical regular polysemous alternations and a 9 point increase in accuracy and an 8 point increase in $F1$-Score in the results of all of the 60 disemous alternations. This is an important result because it directly highlights the main difference between WE representations and those built with count-context models. Furthermore, this result indicates that Step 1 of the Cascade Approach is rendered ineffective when using representations that are not as highly affected by data sparsity, such as is the case with WE representations. This is further indicated by the statistically significant increase ($p < 0.05$) of results obtained in Experiment 2 and those obtained in Step 2 of Experiment 1.

Moreover, the strong results obtained in Experiment 2 provide empirical evidence for the treatment of complex-type nouns of a given regular polysemous alternation as members of a separate, individual lexical-semantic class. Furthermore, these results provide evidence that complex-type nominals do, in fact, have common characteristics that are both identifiable and learnable.

5.3.4 Discussion

As indicated in Tables 5.11 and 5.12, WE representations are able to classify complex-type nouns efficiently as members of a separate lexical-semantic class. Along this line, we comment on the average accuracy score of 76.41% reported in Table 5.12 for the prototypical alternations in Step 1 of Experiment 1, and the fact that the score did not increase in Step 2 of Experiment 1. This result is attributed to the fact that the (x/y) classifiers are trying to classify a broader range of nouns that are essentially from 3 different lexical-semantic classes: (x), (y) and $(x \cdot y)$. Classifying nouns from the single-sense (x) and (y) classes into the super (x/y) class can be a challenge to the classifier because simple-sense nouns do not necessarily share distributional characteristics with each other, as required by nouns of the $(x \cdot y)$ class. Therefore, these distributional word representations behave differently in context, which can exclude certain nouns from classification into the super (x/y) class, thus accounting for the 10 point increase in accuracy when directly classifying complex-type nouns in Experiment 2.

We observed this directly in the data of the GRS·LOG (SOCIAL_GROUP·GEO_LOCATION)class. In Step 1 of Experiment 1, the GRS·LOG noun *village* was misclassified with a probability score of 0.96, thus indicating almost certainty of the class membership decision by the classifier. However, when we directly classified this noun as a member of the GRS·LOG class in Experiment 2, the classifier was able to correctly assign the noun class membership along with other members of this the regular polysemous alternation with a strong probability score of 0.80. Moreover, the misclassifications in Experiment 2 do not increase upon the misclassifications reported in Step 1 of Experiment 1, meaning that either the same number of nouns were misclassified in Experiment 2 or less nouns were misclassified. For instance, the GRS·LOG nouns *suburbia* and *neighborhood* were both misclassified in Step 1 of Experiment 1, with probability scores of 0.98 and 0.93, respectively. In Experiment 2, they remained misclassified, yet with probability scores of 0.78 and 0.58 respectively, indicating that the classifier was not as certain of their "non-membership" in the GRS·LOG class, which consequently implies that their representation, although not sufficient for accurate class membership decision, was still less similar to the representations of non-class members, which resulted in a lower probability score for their misclassification.

The results in both Tables 5.11 and 5.12 indicate that complex-type nouns share a large enough amount of distributional characteristics, resulting in robust class membership decisions, with more than 86% accuracy in Experiment 2. Furthermore, our results demonstrate that there are definitive characteristics that are unique to regular polysemous nouns because they are provided with more information when classified together as members of their own class, following the preliminary results that we obtained in Section 5.2.2, by combining the lexical semantic patterns from both classes of the alternation (LOC and ORG, for instance).

To further inspect the results obtained in Experiment 2, we consider again our conclusions in Section 4.4: WE representations eliminate the sparse data problem in distributional word representations. The results in Section 4.4 confirmed the success of WE representations to overcome the sparse data problem. We are able to further overcome even the need for a 2-Step procedure because the WE word representations are robust enough to allow for direct classification of regular polysemous nouns into a dedicated lexical-semantic class. This is also justified by the overall reduction in misclassifications when directly classifying regular polysemous nouns into their own lexical-semantic class. Moreover, the direct classification approach of the annotated alternations in Experiment 2 reported a recall of 0.86, which is an increase of 10 points in comparison to the recall reported in Step 1 of Experiment 1.

Furthermore, Experiment 2 reported an accuracy of 86.09%, which is almost a full 10 point increase from the 76.41% accuracy obtained with the classifiers of Step 1 of Experiment 1 for the prototypical alternations. With this comparison, we do not need to refer to the final results in Step 2 of Experiment 1 because they do not obtain a perfect classification, which automatically implies that the reported precision and recall decreases, as a direct effect of the results obtained in Step 1. For this reason, when using WE for complex-type nominal classification, the results obtained clearly indicate that the WE representations are robust enough to directly classify the nouns as members of their own class, without a preliminary step to capture low-frequent relevant lexical items.

Thus, the results obtained from Experiment 2 further support our hypothesis that when distributional information is accurately and usefully represented, regular polysemous items form their own lexical semantic class. This is verified by their ability to be classified together with quite a high precision, which is statistically significantly higher from the results obtained in the previous experiments conducted.

5.3.5 Final remarks

The experiments presented in this Section provided a strategy using WE representations to overcome the issues of sense asymmetry and data sparsity that occur in the vectors of complex-type nouns, due to the low frequency of these nouns in some contexts that are indicative of their alternations, as observed in Section 5.2. The Cascade Approach attempted to overcome these challenges by combining distributional information available to complex-type nouns in one feature vectors to smooth the asymmetry problems that can result in insufficient distributional information.

The results obtained using WE representations for classification demonstrated that sparsity is reduced in WE representations. This reduction in sparsity is indicated by the increase of recall in both Experiments 1 and 2 in comparison to the results obtained in Section 5.2 using word representations built with the LING model. Furthermore, the classification results using WE representations provide strong evidence toward the treatment of complex-type nouns of a given regular polysemous alternations as members of their own separate lexical semantic class. We draw this conclusion based on the ability of these nouns to be directly classified as such based on their own distinctive, identifiable, and, most importantly, learnable characteristics.

Furthermore, the results presented in this Section seem to demonstrate a generalization of the behavior of these types of nouns in corpus data, that smooths the asymmetry of occurrences in the vector. We further consider that the WE representations may also contain those very low-frequent yet highly indicative contexts characteristic to nouns that instantiate regular polysemous alternations, although we cannot directly inspect this fact.

Chapter 6

CONTRIBUTIONS AND CONCLUSIONS

Nominal lexical semantic class information is critical for a broad variety of NLP tasks, yet its manual production has been known to be costly and time-consuming. One approach to solve the construction and maintenance of large-coverage lexica for feeding NLP systems is the automatic acquisition of lexical information, which involves the induction of the semantic class related to a particular word from distributional data gathered within a corpus.

In this thesis, we concentrated our work precisely on automatically acquiring lexical-semantic information through empirical validations of different distributional representations of nouns that are used for the classification and the prediction of unknown nouns. We identified—and concentrated on overcoming—two of the main challenges of this task, pinpointed in the State of the Art and observed in our empirical validations.

6.1 Contributions

The main contribution of this dissertation is an empirical study of distributional representations used for nominal lexical semantic classification. We expanded upon six peer-reviewed articles [Bel et al., 2012, Romeo et al., 2012, Romeo et al., 2013b, Romeo et al., 2014a, Romeo et al., 2014b, Romeo et al., 2014c] related to the topics of nominal lexical semantic classification, distributional representations and regular polysemy.

The following are the contributions that resulted from the work presented in this thesis:

- A methodology to define the criteria to identify linguistically-motivated class-indicative features based on major linguistic categories including predicate selectional restrictions, grammatical functions, prepositions and suffixes.
- The definition and identification of class-indicative lexico-syntactic patterns for five lexical-semantic classes in English that achieve an average $F1$-Score of 0.76. The use of these patterns confirms that theoretically postulated explicit rules do certainly represent necessary boundaries for nouns belonging to a given lexical-semantic class.
- The definition and application of what we call unmarked contexts. Furthermore, we outlined a strategy to include these unmarked contexts in distributional vectors that encode the deviation of the occurrences of a noun in a specific context from the average occurrence of all nouns in that same context, achieving an average increase of accuracy of 5.19 points.
- A method for a Cascade Approach that achieves an overall average $F1$-Score of 0.69 with the LING model, and an average $F1$-Score of 0.75 with the WE model, whose main characteristic is a two-step approach that first classifies the members of a so-called super class (x/y) to broaden the distributional profile of nouns being classified to include all of the potential components of a given regular polysemous alternation, and then classifies in a second step complex-type $(x \cdot y)$ nouns from simple-type (x) and (y) nouns.
- Automatically constructed single-sense (i.e. monosemous) data sets for five lexical-semantic classes in English (EVT, HUM, ORG, LOC, COM) that have been used and empirically verified in several experiments and for several different models.
- Human-annotated polysemous data sets for two prototypical regular polysemous alternations (LOC·ORG and EVT·COM) that have been used and empirically verified in the Cascade Approach for complex-type nominal classification. These data sets contain information regarding the potential of nouns to be systematically interpreted in both senses that form a given regular polysemous alternation.
- The identification of the disemous regular polysemous alternations of the CoreLex repository that are prototypical of regular polysemy, according to

the GL, and not of other types of lexical ambiguity, such as homonymy, in a human annotation task. We confirm the validity of these alternations as true examples of regular polysemy by the increase of average accuracy of more than 3.5 points from the entire list of 60 disemous classes, as demonstrated in Section 5.3.

All resulting lexical-resources and data sets have been made freely available for download and use[1].

6.2 Main conclusions

The main research question proposed in this dissertation addresses: (i) whether corpus data provides sufficient distributional information to build efficient word representations that result in accurate and robust classification decisions; and (ii) whether automatic acquisition can handle polysemous nouns. The results in Chapters 4 and 5 allow us to draw the following conclusions.

Distributional information obtained in corpus data is sufficient to automatically acquire lexical semantic classes. However, a word representation, ultimately like those built from the WE model, that maintains the "semantics" underlying the occurrences of a noun in corpus data by mapping it to a dense feature vector, is necessary because it offers reduced dimensionality of the vector space, and considers only real numbers. Thus, it provides actual informative data to the classifiers, avoiding the zero values that negatively effect classifications decisions when using count-context models.

Furthermore, count-context models have been proven to maintain an upper limit related to—not caused by—the information available in corpus data; specifically in regards to their representation of distributional information. The results obtained with the LINGLINE and the DM models (an average $F1$-Score of 0.83) verify that the information available in corpus data is, in fact, sufficient to accurately separate lexical semantic classes. However, it still does not achieve the performance that mapped WE representations are able to achieve.

In order to handle polysemy, nouns that instantiate a given regular polysemous alternation should be treated as members of their own separate class for any lexical-semantic classification task. Unlike simple-type nouns, complex-type nouns can occur in contexts of each of the semantic classes that form the

[1] http://repositori.upf.edu/handle/10230/24562

alternation, as well as in contexts in which both senses are selected for, resulting in characteristic patterns of occurrence in corpus data that differ from simple-type monosemous nouns that cannot instantiate the alternation.

Distributional word representations and lexical-semantic classes

In Sections 4.1, 4.2 and 4.3, we successfully confirmed the learnability of lexical semantic classes using manually identified patterns of linguistic information, achieving an overall average $F1$-Score of 0.76 with an average of only 18 features per class.

On the one hand, the combination of this linguistic information with other commonly occurring—yet general—information, what we defined as unmarked contexts, or bag-of-words-type information, such as in our LINGLINE model, alleviated sparsity in the vectorial representations. This was confirmed by the reduction of errors by an average of 5 points and also through a clear reduction of false negatives. On the other hand, the inclusion of syntactic information provided by a dependency parse in the DM model provides structure to the lexical information in the features, which filters out noise and results in an overall $F1$-Score that is a statistically significant improvement over all the other count-context models, as observed in its higher average $F1$-Score of 0.84. Yet the DM models, which include syntactic dependency information, require the largest number of features—an average of more than $600,000$ per class—and are trained on more than 3 billion tokens of corpus data.

Thus, distributional models based on linguistically motivated information proved to be a viable solution to build word representations also on differently sized corpora, achieving consistent results on corpora ranging from 3 million tokens to 3 billion tokens, and with such a small amount of features. This further confirms that features based on linguistic knowledge accurately generalize and characterize indicative marks of a lexical-semantic class.

Finally, in Section 4.4, we empirically provided evidence (an overall average $F1$-Score of 0.91) that WE models almost completely overcome the sparse data problem. This dramatic increase in performance indicated that our analysis of the count-context models had a limited efficiency. Furthermore, linguistically motivated features do not always necessarily uncover the more implicit relations that can exist between similar words that occur in similar contexts that a mapped model, such as WE models, can procure. Likewise, we also consider that some of the more meaningful relations between words of a given class might require the information from all of language use to be able to uncover them. The representations built using WE tend to more closely simulate this idea, and the resulting performance of these classifiers with WE representations is so strong that they (almost) perfectly represent the occurrences of nouns of given lexical semantic classes, both monosemous and polysemous, in corpus data. These results confirm that WE models overcome the sparse data problem encountered with count-context models. Ultimately, using real numbers and condensing the number of mapped dimensions provide a better representation to the classifier.

With regard to the significant improvement of classification using WE, we also recall that the number of features (we used vectors of 200 dimensions) needed to obtain these results is much higher than the number of features required by LING, for instance, and it implied a large amount of corpus data (we trained the WE models on 3 billion tokens of corpus data), which must be considered when evaluating this model.

Finally, one of the main results of this thesis highlights the strong performance of WE models for nominal lexical semantic classification. Nonetheless, these models did not allow us to draw certain conclusions based on the fact that they are essentially not inspectable, namely in regards to the actual contextual evidence they represent. On the one hand, it is because they are not actually representing individual components in each dimension, as count-context models do. On the other hand, as mentioned earlier, the representations built using WE may stimulate more closely both the implicit and explicit relations that occur between words. Either way, we underline the fact that this limitation resulted in an inability to draw specific conclusions about the information contributing to the formation of a lexical class based on the contextual behavior of nouns when using this model for classification. From a linguistic point of view, this represents a major constraint toward the further understanding of the contextual boundaries required to define a lexical semantic class.

Thus, one future avenue of this work is to extensively and exhaustively explore the relations between word representations built using WE and actual occurrences of a word in corpus data. Although some intrinsic correlations have been identified [Levy and Goldberg, 2014a], an in-depth and dedicated study is not only warranted but needed to further understand the implications of the generalization process that occurs during the mapping of data into word representations for monosemous and regular polysemous nouns alike.

Representing regular polysemous complex-type nouns

We verified that the distributional information indicative of each class is sufficient to classify complex-type nouns, as proposed by GL [Pustejovsky, 1995], by classifying complex-type nouns individually as a member of each class that forms a given regular polysemous alternation. An average overall accuracy of 64.84% confirms that complex-type nouns show distributional behavior of each individual class that forms the regular polysemous alternation of that complex type.

We identified that the main limitation for the classification of complex-type nominals using count-context models is a resulting asymmetry in word representations caused by the unequal occurrence of a target noun in contexts indicative of each individual class that forms a regular polysemous alternation in context, indicated by the results discussed in Section 5.2.

With the results obtained in Chapter 5, we conclude that regular polysemous complex-type nouns should be treated as members of their own separate lexical-semantic class. This is because these nouns share sufficient distributional evidence that is identifiable and, more importantly, learnable, which permits their classification into a separate lexical-semantic class, with an impressive 0.85 average overall $F1$-Score using WE representations.

Thus, we conclude that complex-type nouns should be treated as members of their own lexical-semantic class because they do not equally occur in corpus data as members of each class that forms a regular polysemous alternation, resulting in biased distributional representations. A dedicated word representation that considers this asymmetry as a feature is imperative to the efficient representation of complex-type nouns. Likewise, the classification of an unknown noun into a nominal lexical semantic class must consider the possibility of that noun being a member of either a monosemous or regular polysemous class.

Distributional representations built with WE models appear to smooth this asymmetry due to its mapping of occurrences from corpus data to vector space. We consider that one of the differences between complex-type nouns and simple-type nouns is the occurrence in contexts representative of characteristics specific to regular polysemous complex-type nouns (i.e., co-predication contexts or underspecified contexts, as outlined in the GL [Pustejovsky, 1995]). Although it cannot be inspected, the mapping conducted in WE representations may include this information, which is not typically considered in count-context models due to its low frequency, and can permit the differentiation between complex and simple-type nouns, resulting in the classification of complex-type nouns into their own lexical-semantic class.

Appendix A

The following Appendices contain supplementary information regarding the models, experiments and data sets presented and discussed in detail throughout this dissertation.

A.1 Identification of linguistically-motivated class-indicative lexico-semantic cues

Below we present lists of the linguistically-motivated cues for each class studied, which served as the base to build the LING model, as presented in Section 4.1. All of the part-of-speech tags follow the Penn Tree Bank tag-set. The lists of the Regular Expressions used for extracting the relevant information from corpus data are freely available for download and use[1].

Each of the cues presented in the Tables below (EVT: Table A.1, HUM: Table A.2, COM: Table A.3, ORG: Table A.4 and LOC: Table A.5) were described in detail in Section 4.1. The exact methodology and steps followed to identify the cues listed in the following tables is described in detail in Chapter 4.

[1] http://repositori.upf.edu/handle/10230/24562

Cues for EVENT nouns	
Cue Type	Regular Expressions Examples
Target noun is preceded by certain adverbs/prepositions.	$(during\vert before\vert after)$ + ##target_noun##
Target noun occurs with certain expressions of time in specific NPs.	$(end\vert beginning\vert day\vert month\vert year\vert second\vert minute\vert hour\vert moment\vert century\vert period\vert time\vert age\vert decade\vert frequency\vert occurrence\vert repetition\vert regularity\vert happening\vert epoque\vert morning\vert night\vert afternoon\vert week\vert occasion\vert date)$ + of + ##target_noun##
Target noun is preceded by certain expressions of time	##target_noun## + [a-z]+\V*+$(day\vert month\vert year\vert second\vert minute\vert hour\vert moment\vert century\vert period\vert time\vert age\vert epoque\vert decade\vert week\vert date\vert while)$
Target noun occurs as subject of certain aspectual verbs or other "occurrence" verbs	##target_noun## +$(throw\vert transpire\vert organize\vert organise\vert happen\vert occur\vert initiate\vert begin\vert commence\vert inaugurate\vert launch\vert induct\vert open\vert originate\vert close\vert conclude\vert end\vert terminate\vert start\vert stop)$\V*
Target noun is an object of certain verbs indicative of EVENT nouns	$(initiate\vert begin\vert commence\vert inaugurate\vert launch\vert induct\vert open\vert originate\vert close\vert conclude\vert end\vert terminate\vert start\vert stop\vert throw\vert lock\vert result\vert involve\vert run\vert experience\vert refuse\vert plan\vert inject\vert complete\vert win\vert hold\vert follow\vert mark\vert launch\vert sustain\vert order\vert miss\vert convene\vert need\vert speak)$\V* + ##target_noun##
External argument of target noun is realized as genitive	[a-z]+('s) + ##target_noun##
Target noun is subject of common verbs that select for EVENT nouns	##target_noun## + $(gain\vert seek\vert spike\vert require\vert cover\vert run\vert erupt\vert race\vert slow\vert continue\vert average\vert hold\vert snap\vert skyrocket\vert stand\vert remain\vert mean\vert disintegrate)$\V*
Target noun is modified by certain adjectives	$(neutral\vert economic\vert third\vert political\vert interior\vert easy\vert full-service\vert minimal\vert standard\vert black\vert compact\vert first\vert annual\vert natural\vert human\vert presidential\vert final\vert other\vert rapid\vert future)$\J* + ##target_noun##
Target noun is preceded by certain agentive prepositions: "$by\vert for$"	##target_noun## + $(by\vert for)$\IN
Target noun precedes certain relative pronouns "$(when\vert where)$"	##target_noun## + $(when\vert where)$\WRB
Target noun modified by certain PP	[a-z]+\IN + ([a-z]+\DT)? [a-z]+\N* + ##target_noun##
Delimiting point of target noun signified by PP	$(until\vert since\vert till)$ + ##target_noun##
Target noun precedes certain post-adjectival modifiers	##target_noun## + $(early\vert late)$
Target noun contains a suffixes indicative of EVENT nominalizations	##[a-z]+$(ment\vert ion)$\target_noun##

Table A.1: Linguistic cues and corresponding lexical-syntactic patterns formalized as Regular Expressions used to extract distributional data indicative of nouns the EVENT class

<table>
<tr><th colspan="2">Cues for HUMAN nouns</th></tr>
<tr><th>Cue Type</th><th>Regular Expression Examples</th></tr>
<tr><td>Target noun is
subject of 35 frequent verbs
(that co-occur with
HUMAN nouns)</td><td>##target_noun## + (excel|support|tell|believe|name|spend|think|describe|ask|write|decide|show|receive|face|die|seem|put|tend|set|buy|consider|bring|represent|meet|appear|feel|agree|relate|start|pay|perform|sit|arrive|argue)\V*</td></tr>
<tr><td>Target noun is
subject of 35 frequent verbs
(that co-occur with
ORG nouns because
they tend to be similar
to HUM nouns due
to lexical gaps)</td><td>##target_noun## + (work|feature|include|live|want|call|base|follow|need|begin|win|continue|offer|hold|consist|report|remain|look|become|operate|lead|move|build|announce|comprise|rely|grow|leave|lose|try|dedicate|found|own|play)\V*</td></tr>
<tr><td>Target noun is
a complement of specific
agentive PP</td><td>(by|for) + ##target_noun##</td></tr>
<tr><td>Target noun is
an indirect object of “give” verbs
[Levin, 1993]</td><td>(give|lend|loan|pass|refund)\V* + ##target_noun##</td></tr>
<tr><td>Target noun is an
indirect object of indicative verbs
from [Levin, 1993], according to
their tendency to occur with
HUMAN nouns</td><td>(manage|trade|bet|evolve|register|protect|work|earn|found|carry|enable|empower|enhance|enable|crown|help|elect|lead|serve|require|strike|preside|appoint|designate|ail|link|become|visit|cost|stay|crew|allow|arm|oust|purchase|kill)\V* + ##target_noun##</td></tr>
<tr><td>Target noun is
modified by a genitive</td><td>[a-z]+(’s)\PoS + ##target_noun##</td></tr>
<tr><td>Target noun is
modified by a possessive pronoun</td><td>[a-z]+\PRP$ + ##target_noun##</td></tr>
<tr><td>Target noun is
headed by a specific relative pronoun</td><td>##target_noun## + (who|whom|whose)\WRB</td></tr>
<tr><td>Target noun is
included in “group of” constructions</td><td>(group|aggregation|association|clique|congregation|crowd|party|assemblage|band|club|coterie|gang|posse|assembly|class|company|crew|gathering|society|troup|troope) + of
+ ##target_noun##</td></tr>
<tr><td>Target noun is
modified by nationality, religion, governmental
affiliation, etc.</td><td>[a-z]+(ish|an|ch|ese|ss|ek|ino|ic|ant|ive|al|nt)\NP
+ ##target_noun##</td></tr>
<tr><td>Target noun
precedes
certain adverbs</td><td>##target_noun## + [a-z]+ly\RB</td></tr>
<tr><td>Target noun
occurs in certain “be located/found at” NPs</td><td>##target_noun## + at + [a-z]+\NP</td></tr>
<tr><td>Target noun is
preceded by
specific JJ that indicate <i>age|growth</i></td><td>(teenage|year − old|senior|junior|pre − pubescent)\J*
+ ##target_noun##</td></tr>
<tr><td>Target noun is preceded by
specific JJ that indicate
personality, political, spiritual preferences, etc.</td><td>(political|executive|proper|other|good|general|valuable|bright|civil|local|religious|spiritual|modern|military|old|medical|traditional|former|first|deputy|dear|global|close|little|public|good|online|scientific|new|innocent|senior|archaic|artistic|graphic|young)\J*
+ ##target_noun##</td></tr>
<tr><td>Target noun contains
specific suffixes indicative of
HUMAN nouns</td><td>##[a-z]+(er|or|man|men|mate|ist|arian|naut|yst|ster|ess)\target_noun##</td></tr>
<tr><td>Target noun contains
specific prefixes indicative of
HUMAN nouns</td><td>##(radio|anti|paleo|vice|epi|ex|neo|col|grand)[a-z]+\target_noun##</td></tr>
</table>

Table A.2: Linguistic cues and corresponding lexical-syntactic patterns formalized as Regular Expressions used to extract distributional data indicative of nouns the HUMAN class

Cues for COMMUNICATION_OBJECT nouns	
Cue types	Regular Expression Examples
Target noun occurs in specific frequent PP phrases	about\IN (+ [a-z]+\DT?) + ##target_noun##
	to\IN (+ [a-z]+\DT?) + ##target_noun##
	for\IN (+ [a-z]+\DT?) + ##target_noun##
	within\IN (+ [a-z]+\DT?) + ##target_noun##
	by\IN (+ [a-z]+\DT?) + ##target_noun##
	without\IN (+ [a-z]+\DT?) + ##target_noun##
Target noun occurs in specific frequent compound PP phrases	$(between \vert against \vert before \vert after)$\IN (+ [a-z]+\DT?) + ##target_noun##
Target noun contains suffixes indicative of COMM nouns	##[a-z]+$(gram \vert graph \vert tion \vert ario \vert phia \vert sion \vert dence \vert graphy \vert logue \vert logy \vert list \vert book \vert tale \vert note \vert chart \vert word \vert letter \vert paper)$\target_noun##
Target noun is an object of frequent verbs that select for COMM nouns	$(close \vert relate \vert mail \vert submit \vert report \vert write \vert send \vert believe \vert ask \vert issue \vert entertain \vert detail \vert manage \vert register \vert shock \vert answer \vert serve \vert interest \vert release \vert publish \vert collect \vert accompany \vert add \vert start)$\V* + ##target_noun##
Target noun is a subject of frequent verbs that select for COMM nouns	##target_noun## + $(complete \vert excel \vert need \vert submit \vert include \vert continue \vert contain \vert break \vert send \vert submit \vert violate \vert show \vert turn \vert call \vert refresh \vert regard \vert personalize \vert release \vert delay \vert collect \vert issue \vert notify \vert direct \vert please \vert mention \vert design)$\V*
Target noun is modified by "-ly" adjectives that are derivates from time/direction words or verbs (i.e. "weekly").	[a-z]+ly\J* + ##target_noun##
Target noun is modified by certain adjectives	$(identifiable \vert personal \vert unrelated \vert other \vert long \vert certain \vert commercial \vert new \vert content \vert online \vert legal)$\J* + ##target_noun##

Table A.3: Linguistic cues and corresponding lexical-syntactic patterns formalized as Regular Expressions used to extract distributional data indicative of nouns the COMMUNICATION_OBJECT class

Cues for ORGANIZATION nouns	
Cue type	Example
Target noun is subject of frequent agentive verbs	##target_noun## + $(work \vert feature \vert include \vert live \vert want \vert call \vert$ $base \vert follow \vert need \vert begin \vert win \vert continue \vert offer \vert hold \vert consist \vert report \vert$ $remain \vert look \vert become \vert operate \vert lead \vert move \vert build \vert announce \vert$ $comprise \vert rely \vert grow \vert leave \vert lose \vert try \vert dedicate \vert found \vert own \vert play)\backslash$V*
Target noun precedes past “founder”-type verbs	##target_noun## + $(create \vert found \vert preside \vert establish \vert$ $endow \vert organize \vert constitute \vert inaugurate \vert institute \vert originate \vert$ $decree)\backslash(VBD \vert VBN)$
Target noun is captured by agentive complements headed with the preposition “for”	for\IN ((+[a-z]+\DT)?) + ##target_noun##
Target noun is an indirect object headed by “to”	[a-z]+\V* + to\TO (+ [a-z]+$\backslash(DT \vert J*)$?) + ##target_noun##
Target noun occurs in independent PP complements headed by “in”	in\IN(+ [a-z]+$\backslash(DT \vert J*)$?) +##target_noun##
Target noun occurs in independent PP complements headed by “within”	within\IN(+ [a-z]+$\backslash(DT \vert J*)$?) +##target_noun##
Target noun occurs in independent PP complements headed by “from”	from\IN(+ [a-z]+$\backslash(DT \vert J*)$?) +##target_noun##
Target noun occurs as a direct object without a dependent PP	([a-z]+\V*)+ (+ [a-z]+$\backslash(DT \vert J*)$?) + ##target_noun##
Target noun is a subject of 35 frequent verbs that select for HUM nouns, which have similar characteristics to ORG nouns	##target_noun## + $(excel \vert support \vert tell \vert believe \vert name \vert$ $spend \vert think \vert describe \vert ask \vert write \vert decide \vert show \vert receive \vert face \vert$ $die \vert seem \vert put \vert tend \vert set \vert buy \vert consider \vert bring \vert represent \vert meet \vert$ $appear \vert feel \vert agree \vert relate \vert start \vert pay \vert perform \vert sit \vert arrive \vert argue)\backslash$V*
Target noun is preceded by certain relative pronouns	##target_noun## + $(who \vert whose \vert whom)\backslash$WRB
Target noun precedes dependent NP headed by “in”	##target_noun## + in\IN (+ [a-z]+$\backslash(DT \vert J*)$?) + [a-z]+\NP
Target noun contains suffixes for (ORG|HUMAN) (nouns tend to be similar to HUM NOUNS due to lexical gaps)	##[a-z]+$(hood \vert racy \vert man \vert men \vert mate \vert naut \vert ity \vert ship \vert ate)\backslash$target_noun##

Table A.4: Linguistic cues and corresponding lexical-syntactic patterns formalized as Regular Expressions used to extract distributional data indicative of nouns the ORGANIZATION class

Cues for LOCATION nouns	
Cue type	Regular Expression Examples
Target noun is headed by locative/simple/ compound PP complements	on\IN (+ [a-z]+$\backslash DT\vert J*$?) + ##target_noun##
	between\IN (+ [a-z]+\($DT\vert J*$)?) + ##target_noun##
	above\IN (+ [a-z]+\($DT\vert J*$)?) + ##target_noun##
	with\IN (+ [a-z]+\($DT\vert J*$)?) + ##target_noun##
	without\IN (+ [a-z]+\($DT\vert J*$)?) + ##target_noun##
	at\IN (+ [a-z]+\($DT\vert J*$)?) + ##target_noun##
	outside\IN (+ [a-z]+\($DT\vert J*$)?) + ##target_noun##
	inside\IN (+ [a-z]+\($DT\vert J*$)?) + ##target_noun##
	along\IN (+ [a-z]+\($DT\vert J*$)?) + ##target_noun##
	through\IN (+ [a-z]+\($DT\vert J*$)?) + ##target_noun##
	toward\IN (+ [a-z]+\($DT\vert J*$)?) + ##target_noun##
	to\IN (+ [a-z]+\($DT\vert J*$)?) + ##target_noun##
Target noun is subject of frequent verbs that select for LOC nouns	##target_noun## + ($thrive\vert regard\vert require\vert appreciate\vert forbid\vert include\vert remain\vert spring\vert conceal\vert request\vert buy\vert stop\vert nurture\vert confess\vert landscape\vert forfend\vert poison\vert stretch\vert lead\vert divide\vert break\vert cover\vert envision\vert own\vert lost\vert end\vert receive\vert describe\vert place\vert help$)\V*
Target noun occurs as Direct object of verb that selects for LOC nouns	($film\vert extend\vert tour\vert exist\vert conduct\vert stand\vert pass\vert educate\vert turn\vert respond\vert subdivide\vert originate\vert write\vert thrive\vert proceed\vert engage\vert divide\vert act\vert roll\vert start\vert call\vert feel\vert set\vert administrate\vert arrive\vert travel\vert walk\vert follow\vert fly\vert move\vert meet\vert run\vert bring\vert drive\vert pass$)\V* + ##target_noun##
Target noun is preceded by certain relative pronouns	##target_noun## + where\WRB
Target noun contains suffixes that are indicative of LOC nouns	##[a-z]+($port\vert teria\vert dom\vert ory\vert topy\vert ium\vert polis\vert way\vert place\vert ground\vert space\vert point\vert land\vert field$)\target_noun##
Target noun is modified by adjectives of dimension	($far\vert close\vert distance\vert high\vert low\vert nearby\vert remote\vert wide\vert narrow\vert north\vert south\vert east\vert west\vert near$)\J* + ##target_noun##
Target noun is modified by frequent "distance" or "location" adjectives	($overseas\vert spiritual\vert intellectual\vert upper\vert healthful\vert posterior\vert early\vert anterior\vert private\vert coral\vert medial\vert third\vert daily\vert southern\vert common\vert lateral\vert safe\vert political\vert later\vert last\vert early\vert natural\vert different\vert next\vert southern\vert human\vert common\vert northern\vert public\vert international\vert final\vert western\vert main\vert modern\vert first\vert exclusive\vert inferior\vert various\vert social\vert right\vert exterior\vert lower$)\J* + ##target_noun##

Table A.5: Linguistic cues and corresponding lexical-syntactic patterns formalized as Regular Expressions used to extract distributional data indicative of nouns the LOCATION class

A.2 Human annotation of polysemous nouns

In this thesis, we built data sets of complex-type nouns using human annotators, as described in Section 3.1.2. These data sets were used for training and testing in Chapter 5. In order to provide guidelines to the human annotation task, we used the worksheets presented in Tables A.7 and A.6.

The human annotators were asked to complete a worksheet providing an annotation for each individual noun in the form of a *yes* or *no* response to one of the following four questions (Examples 19-22), according to the data set that was being considered.

(21) **Q1.** *Consider the following definition*: **ORGANIZATION** is an entity that has a collective goal.

*Mark "yes" or "no" if you think the current noun can be interpreted as an "*ORGANIZATION*" noun (besides potentially having any other sense)*

(22) **Q2.** *Consider the following definition:* **LOCATION** is a place, a specific position or a point in physical space.

*Mark "yes" or "no" if you think the current noun can be interpreted as an "*EVENT*" noun (besides potentially having any other sense)*

(23) **Q3.** *Consider the following definition:* **EVENT** is an occurrence, something that happens or is regarded as happening.

*Mark "yes" or "no" if you think the current noun can be interpreted as an "*EVENT*" noun (besides potentially having any other sense)*

(24) **Q4.** *Consider the following definition:* **COMMUNICATION_OBJECT** is any sort of knowledge communicated or received.

Mark "yes" or "no" if you think the current noun can be interpreted as an "communication object" noun (besides potentially having any other sense)

Thus, there were four worksheets provided to each annotator to annotate, which consisted of a total of 743 words to annotate. Once those annotations were obtained, we then used a voting scheme to select the majority annotation between the human annotators to assign a sense to each of the individual nouns in the data set.

Tables A.6 and A.7 present the information regarding each individual noun, the class assigned to it and the number of annotators agreeing with that particular class assignment. Each Table contains all of the words considered for each regular polysemous alternation studied. The implications of this task and the annotation results obtained were discussed in detail in Sections 3.1 and 5.2.

WORD	CLASS	TOTAL	WORD	CLASS	TOTAL	WORD	CLASS	TOTAL	WORD	CLASS	TOTAL	WORD	CLASS	TOTAL
riverside	LOC	3	fatherland	LOC· ORG	3	electorate	ORG	2	pinnacle	LOC	2	crew	ORG	3
midway	LOC	2	boundary	LOC	3	readership	ORG	3	coalfield	LOC	3	brotherhood	ORG	3
environs	LOC	3	sphere	LOC· ORG	2	gerontocracy	ORG	3	jurisdiction	LOC· ORG	2	hierocracy	ORG	3
wasteland	LOC· ORG	2	ground	LOC	3	college	LOC· ORG	3	encampment	LOC	2	army	LOC· ORG	2
scenery	LOC	3	hometown	LOC· ORG	3	academe	LOC· ORG	2	square	LOC	2	monarchy	LOC· ORG	2
northwest	LOC· ORG	2	bottom	LOC	2	troupe	ORG	3	fountainhead	LOC	3	entourage	ORG	3
bilocation	LOC	3	touchline	LOC	3	squad	ORG	2	borderline	LOC	2	conglomeration	LOC· ORG	2
sprawl	LOC	2	precinct	LOC· ORG	2	confederation	LOC· ORG	3	seafront	LOC	2	meritocracy	ORG	3
campground	LOC	2	borough	LOC· ORG	3	institute	LOC· ORG	3	epicentre	LOC	2	platoon	ORG	3
turf	LOC	2	oasis	LOC	3	colloquium	ORG	2	birthplace	LOC	3	knighthood	ORG	3
equator	LOC	3	megalopolis	LOC· ORG	3	alliance	ORG	3	prefecture	LOC· ORG	3	church	LOC· ORG	3
viewpoint	LOC	3	hemisphere	LOC· ORG	2	troop	ORG	2	heartland	LOC· ORG	2	sisterhood	ORG	3
archdiocese	LOC· ORG	3	cornfield	LOC	3	faculty	LOC· ORG	3	bedside	LOC	3	syndicate	LOC· ORG	2
resort	LOC· ORG	2	countryside	LOC· ORG	2	lineage	ORG	3	destination	LOC	3	clique	ORG	3
airspace	LOC	2	endpoint	LOC	3	chorus	ORG	2	dockside	LOC	3	academia	ORG	2
frontier	LOC	2	playground	LOC	2	tribunal	LOC· ORG	3	forefront	LOC	3	republic	LOC· ORG	3
harbourage	LOC	2	reservation	LOC· ORG	2	affiliate	LOC· ORG	2	haven	LOC· ORG	2	aggregate	ORG	3
airway	LOC	3	retreat	LOC· ORG	2	oligarchy	ORG	2	stage	LOC	3	inspectorate	LOC· ORG	2
ghetto	LOC· ORG	3	waterline	LOC	3	bureaucracy	ORG	3	cemetery	LOC	3	masonry	ORG	3
checkpoint	LOC	2	plaza	LOC· ORG	2	staff	ORG	3	hamlet	LOC· ORG	3	dictatorship	LOC· ORG	2
backwater	LOC	3	habitat	LOC	3	throng	ORG	2	enclave	LOC· ORG	3	unit	LOC· ORG	3
rooftop	LOC	3	outside	LOC	3	squadron	ORG	3	pasture	LOC	3	population	ORG	3
midpoint	LOC	2	fireside	LOC	3	convoy	ORG	3	borderland	LOC	2	parliament	LOC· ORG	3
coastline	LOC	2	diocese	LOC· ORG	3	triumvirate	ORG	3	terminal	LOC	3	subgroup	ORG	3
acre	LOC	3	harbour	LOC	2	assembly	LOC· ORG	2	slum	LOC· ORG	3	nobility	ORG	3
township	LOC· ORG	2	desktop	LOC	3	armada	LOC· ORG	2	graveyard	LOC	3	fraternity	ORG	2
latitude	LOC	3	port	LOC· ORG	2	minority	ORG	3	locality	LOC	2	directorate	ORG	2
path	LOC	3	underside	LOC	3	family	ORG	3	atmosphere	LOC	3	desert	LOC	3
biosphere	LOC	2	seaport	LOC· ORG	2	choir	LOC· ORG	2	ionosphere	LOC	3	epicenter	LOC	2
skyline	LOC	3	scenario	LOC	3	academy	LOC· ORG	3	hierarchy	ORG	3	darkness	LOC	3
battlefield	LOC	2	border	LOC	2	dynasty	ORG	3	autocracy	ORG	2	funfair	LOC	2
heaven	LOC	3	scene	LOC	3	womanhood	ORG	3	congregation	ORG	2	stopover	LOC	3
landmark	LOC	3	environment	LOC	3	gentry	ORG	3	patriarchy	ORG	2	dealership	LOC· ORG	2
continent	LOC· ORG	2	homeland	LOC· ORG	3	personnel	ORG	3	workforce	ORG	3	theocracy	LOC· ORG	2
heliosphere	LOC	3	tidewater	LOC	3	clientele	ORG	3	team	ORG	3	secretariat	LOC· ORG	3
oilfield	LOC	3	fairground	LOC	2	legion	ORG	2	poor	ORG	2	cooperative	LOC· ORG	3
grassland	LOC	2	battleground	LOC	2	caste	ORG	3	sainthood	ORG	3	herd	ORG	3
hilltop	LOC	3	sideline	LOC	2	caravan	LOC· ORG	3	jurisprudence	ORG	2	papacy	ORG	2
dukedom	LOC· ORG	3	territory	LOC· ORG	2	association	LOC· ORG	3	sorority	LOC· ORG	2	corps	ORG	3
overhead	LOC	3	circumference	LOC	3	homeless	ORG	2	organization	LOC· ORG	2	admiralty	LOC· ORG	2
minefield	LOC	3	paradise	LOC	3	horde	ORG	3	cadre	ORG	3	mob	ORG	3
schoolyard	LOC	2	aerospace	LOC	2	society	LOC· ORG	2	proletariat	ORG	2	aristocracy	ORG	2
savannah	LOC	3	pole	LOC	3	guild	ORG	2	elite	ORG	3	womankind	ORG	3
beachhead	LOC· ORG	2	battlefront	LOC· ORG	3	copartnership	ORG	3	matriarchy	ORG	3	democracy	ORG	2
crawlspace	LOC	3	solitude	LOC	3	administration	LOC· ORG	3	senate	LOC· ORG	3	technocracy	ORG	3
property	LOC	3	farmland	LOC	2	fellowship	ORG	3	company	LOC· ORG	3	university	LOC· ORG	3
laboratory	LOC· ORG	3	hearth	LOC· ORG	2	elderly	ORG	3	club	LOC· ORG	3	jury	ORG	3
habitation	LOC	3	seascape	LOC	3	agency	LOC· ORG	3	kinfolk	ORG	3	leadership	ORG	3
reef	LOC	3	domicile	LOC	2	crowd	ORG	3	organisation	LOC· ORG	3	priesthood	ORG	3
churchyard	LOC	2	hotspot	LOC	3	pontificate	ORG	2	nation	LOC· ORG	3	congress	LOC· ORG	3
lookout	LOC	2	authority	ORG	3	rabbinate	ORG	2	regime	ORG	3	midfield	LOC	3
campus	LOC· ORG	2	committee	LOC· ORG	2	consortium	LOC· ORG	2	league	ORG	2			
tip	LOC	3	forum	LOC· ORG	3	pastorate	LOC· ORG	2	plutocracy	ORG	3			

Table A.6: Human annotation results and class assignment for LOC/ORG data set

NOUN	CLASS	TOTAL	NOUN	CLASS	TOTAL	NOUN	CLASS	TOTAL	NOUN	CLASS	TOTAL	NOUN	CLASS	TOTAL
observance	EVENT·COM	2	materialisation	EVENT	3	experience	EVENT	2	infomercial	EVENT·COM	2	exam	EVENT·COM	3
flashing	EVENT	3	swerve	EVENT	3	severance	EVENT	3	lemma	COM	3	briefing	EVENT·COM	3
marathon	EVENT	3	headway	EVENT	3	disturbance	EVENT	3	database	COM	3	fiction	COM	3
suspension	EVENT	3	dislocation	EVENT	3	splash	EVENT	3	abbreviation	EVENT·COM	3	calendar	COM	3
bonanza	EVENT	3	heave	EVENT	3	densification	EVENT	3	command	EVENT·COM	3	inquiry	EVENT·COM	3
setback	EVENT	3	shame	EVENT	3	extinction	EVENT	3	genre	COM	3	authorisation	EVENT·COM	3
warp	EVENT	3	trip	EVENT	3	mortification	EVENT	3	headline	COM	3	copyright	COM	3
strike	EVENT	3	growth	EVENT	3	progression	EVENT	3	memo	COM	3	movie	EVENT·COM	2
smudge	EVENT	2	fundraiser	EVENT	3	demolition	EVENT	3	content	COM	2	billet	COM	3
turning	EVENT	3	wallow	EVENT	3	doomsday	EVENT	3	prognosis	COM	2	word	COM	3
impregnation	EVENT	3	consolidation	EVENT	2	affair	EVENT	3	nonfiction	COM	3	reservation	EVENT·COM	3
diving	EVENT	3	blackout	EVENT	3	ascension	EVENT	3	software	COM	3	printout	COM	2
lightning	EVENT	3	downfall	EVENT	3	undercurrent	EVENT	3	propaganda	EVENT·COM	2	caption	COM	3
climbing	EVENT	3	uplift	EVENT	2	competition	EVENT	3	pseudonym	COM	3	magazine	COM	3
contest	EVENT	2	remission	EVENT	3	impulse	EVENT	3	memorandum	COM	3	playscript	COM	2
rectification	EVENT·COM	3	disruption	EVENT	3	meeting	EVENT	2	theorem	COM	2	schedule	COM	2
permutation	EVENT	3	humiliation	EVENT	3	leakage	EVENT	2	symbol	COM	3	grade	EVENT·COM	2
playoff	EVENT	3	disappearance	EVENT	3	meltdown	EVENT	3	mail	COM	2	submission	EVENT·COM	3
substitution	EVENT	2	resurrection	EVENT	3	breakup	EVENT	3	statement	EVENT·COM	3	foreword	COM	2
plunge	EVENT	3	shock	EVENT	3	trial	EVENT	3	notepaper	COM	3	quote	EVENT·COM	3
earthquake	EVENT	3	sideshow	EVENT	2	squeeze	EVENT	3	metonym	COM	3	permission	EVENT·COM	3
recurrence	EVENT	3	wreck	EVENT	3	sunset	EVENT	3	interpretation	EVENT·COM	3	stenograph	COM	2
upheaval	EVENT	3	expiration	EVENT	2	decision	EVENT·COM	3	lecture	EVENT·COM	3	gag	EVENT·COM	3
separation	EVENT	3	dressing	EVENT	3	eruption	EVENT	3	vocabulary	COM	3	antithesis	COM	3
pileup	EVENT	2	regurgitation	EVENT	3	semifinal	EVENT	3	portrayal	EVENT·COM	3	testimony	EVENT·COM	3
coronation	EVENT	3	race	EVENT	3	wedding	EVENT	3	novel	COM	2	encyclopedia	COM	3
romp	EVENT	3	cracking	EVENT	3	precipitation	EVENT	3	character	COM	3	permit	COM	2
exhaustion	EVENT	3	whirlpool	EVENT	3	landslide	EVENT	3	term	COM	2	logbook	COM	2
deflection	EVENT	3	campaign	EVENT·COM	3	extermination	EVENT	3	nomination	EVENT·COM	3	firmware	COM	3
outgrowth	EVENT	2	installation	EVENT	3	stampede	EVENT	3	checklist	COM	2	context	COM	2
rumble	EVENT	3	deformation	EVENT	3	departure	EVENT	3	questionnaire	EVENT·COM	2	monologue	EVENT·COM	2
hardship	EVENT	3	thunder	EVENT	3	smolder	EVENT	3	sonnet	COM	3	editorial	COM	2
debacle	EVENT	3	standoff	EVENT	3	interruption	EVENT	2	newsletter	COM	3	idiom	COM	3
tsunami	EVENT	3	sacrifice	EVENT	3	schism	EVENT	3	shortlist	COM	2	tale	COM	2
aftershock	EVENT	3	relaxation	EVENT	3	burst	EVENT	3	application	EVENT·COM	3	paragraph	COM	2
relief	EVENT	3	beginning	EVENT	3	escape	EVENT	3	misquotation	EVENT·COM	3	preface	COM	3
steeplechase	EVENT	3	bounce	EVENT	3	pollination	EVENT	3	channel	COM	2	guideline	COM	3
shrinkage	EVENT	3	commencement	EVENT	3	ramification	EVENT	2	speech	EVENT·COM	3	oratory	COM	2
fading	EVENT	3	rebound	EVENT	3	displacement	EVENT	3	reportage	EVENT·COM	3	timetable	COM	3
spike	EVENT	3	procession	EVENT	3	breach	EVENT	2	notebook	COM	3	travelogue	EVENT·COM	2
overflow	EVENT	3	rebirth	EVENT	3	approaching	EVENT	2	prolog	EVENT·COM	2	holograph	COM	3
sleepover	EVENT	3	maelstrom	EVENT	3	ruination	EVENT	3	autograph	COM	2	guidepost	COM	2
outbreak	EVENT	3	fiasco	EVENT	3	outcome	EVENT	2	address	EVENT·COM	2	directory	COM	2
wildfire	EVENT	3	incident	EVENT	3	championship	EVENT	3	referral	EVENT·COM	2	cookbook	COM	3
misadventure	EVENT	3	visitation	EVENT	3	miracle	EVENT	3	journal	COM	2	credential	COM	2
shipwreck	EVENT	3	knock	EVENT	3	burial	EVENT	3	edict	EVENT·COM	2	website	COM	3
sunrise	EVENT	3	repercussion	EVENT	3	elevation	EVENT	2	broadcast	EVENT·COM	3	flashcard	COM	3
break	EVENT	3	joust	EVENT	3	snore	EVENT	3	testimonial	EVENT·COM	2	studbook	COM	3
passing	EVENT	3	eclipse	EVENT	3	depredation	EVENT	2	catalog	COM	2	film	EVENT·COM	2
collapse	EVENT	3	strengthening	EVENT	3	emission	EVENT·COM	2	blacklist	COM	2	document	COM	2
abatement	EVENT	2	disaster	EVENT	3	climb	EVENT	3	measure	EVENT·COM	3	mayday	EVENT·COM	2
cessation	EVENT	2	discharge	EVENT	3	progress	EVENT	3	gazette	COM	3	logograph	COM	3
intrusion	EVENT	3	loss	EVENT	3	preservation	EVENT	3	leaflet	COM	3	commercial	EVENT·COM	2
rise	EVENT	3	cotillion	EVENT	3	plague	EVENT	3	telecast	EVENT·COM	2	diploma	COM	3
celebration	EVENT	3	irradiation	EVENT	3	playlist	COM	2	overview	EVENT·COM	3	webpage	COM	3
malfunction	EVENT	3	entrance	EVENT	3	definition	EVENT·COM	2	recipe	COM	3	volume	COM	3
appearance	EVENT	3	comeuppance	EVENT	3	ideogram	COM	3	portfolio	COM	3	criterion	COM	3
discrepancy	EVENT	2	rupture	EVENT	3	slogan	COM	3	gazetteer	COM	2	songbook	COM	3
coincidence	EVENT	2	adjustment	EVENT·COM	2	literature	COM	2	prayerbook	COM	3	announcement	EVENT·COM	3
compression	EVENT	2	degeneration	EVENT	3	decree	EVENT·COM	2	dictation	EVENT·COM	3	inscription	EVENT·COM	3
replay	EVENT	2	tribulation	EVENT	3	instruction	EVENT·COM	3	recommendation	EVENT·COM	3	flowchart	COM	2
exit	EVENT	2	destruction	EVENT	3	chat	EVENT·COM	3	biography	COM	3	illustration	EVENT·COM	3
catastrophe	EVENT	3	recovery	EVENT	3	acronym	COM	3	acknowledgement	EVENT·COM	3	epilogue	EVENT·COM	2
finish	EVENT	3	avalanche	EVENT	3	copybook	COM	2	folktale	COM	2	workbook	COM	3
cascade	EVENT	3	epidemic	EVENT	3	quiz	EVENT·COM	3	trademark	COM	2	parody	EVENT·COM	2
epiphany	EVENT·COM	2	decrease	EVENT	3	newspaper	COM	3	codex	COM	3	homograph	COM	3
immersion	EVENT	3	constriction	EVENT	3	lithography	COM	3	manifesto	COM	3	horoscope	COM	3
emergency	EVENT	3	mudslide	EVENT	3	coverage	EVENT·COM	3	jargon	COM	2	charade	EVENT·COM	3
brawl	EVENT	3	reversion	EVENT	3	documentary	EVENT·COM	2	advisory	COM	3	fairytale	COM	2
triumph	EVENT	3	rip	EVENT	2	typescript	COM	2	cryptogram	COM	3	message	COM	3
defeat	EVENT	3	ceremony	EVENT	3	lexicon	COM	3	radiogram	COM	2	gpa	COM	2
beep	EVENT	3	torment	EVENT	3	formula	COM	3	semicolon	COM	3	insignia	COM	3
destabilization	EVENT	3	creation	EVENT	2	picture	COM	2	atlas	COM	3	variable	COM	3
settling	EVENT	3	standstill	EVENT	3	anecdote	EVENT·COM	2	paraphrase	COM	2	communique	EVENT·COM	2
inception	EVENT	3	invasion	EVENT	3	interview	EVENT·COM	3	correspondence	EVENT·COM	3	passport	COM	3
ordeal	EVENT	2	divergence	EVENT	2	summary	EVENT·COM	2	epigraph	COM	3	letterpress	COM	2
victory	EVENT	3	swell	EVENT	3	password	COM	3	quotation	EVENT·COM	3	assignment	EVENT·COM	3
replacement	EVENT	3	trample	EVENT	3	appendix	COM	3	anagram	COM	3	hieroglyphic	COM	3
surge	EVENT	3	party	EVENT	3	warranty	EVENT·COM	2	dialogue	EVENT·COM	3	ultimatum	EVENT·COM	2
crucifixion	EVENT	3	occurrence	EVENT	3	centerfold	COM	2	covenant	EVENT·COM	3	query	EVENT·COM	3
walloping	EVENT	3	funeral	EVENT	3	album	COM	3	hieroglyph	COM	3	filename	COM	3
accident	EVENT	3	perturbation	EVENT	3	epigram	COM	3	article	COM	3	playbill	COM	3
conception	EVENT	2	vision	EVENT·COM	2	finale	EVENT·COM	2	register	EVENT·COM	3	abridgement	EVENT·COM	3
shower	EVENT	3	fatality	EVENT	3	coupon	COM	3	postcard	COM	2	newscast	EVENT·COM	2
reversal	EVENT	3	respite	EVENT	3	question	EVENT·COM	3	footnote	COM	2	notice	EVENT·COM	3
phenomenon	EVENT	3	retrogression	EVENT	3	textbook	COM	3	handbook	COM	3	video	COM	2
downhill	EVENT	3	modification	EVENT·COM	2	consent	EVENT·COM	2	thesis	COM	2	vignette	COM	2
radiation	EVENT	3	changeover	EVENT	3	screenplay	COM	3	motto	COM	3	analysis	EVENT·COM	3
smash	EVENT	3	onrush	EVENT	3	thesaurus	COM	3	spreadsheet	COM	3	syllabus	COM	3
crackling	EVENT	3	inflation	EVENT	3	consonant	COM	3	newsflash	EVENT·COM	2	sentence	COM	3
puncture	EVENT	3	travel	EVENT	3	invitation	EVENT·COM	3	prescription	EVENT·COM	3	conference	EVENT·COM	3
devastation	EVENT	3	blowout	EVENT	3	timeline	COM	3	book	COM	3	audio	COM	2
initiation	EVENT	3	convulsion	EVENT	3	introduction	EVENT·COM	3	guide	COM	2	lesson	EVENT·COM	3
outline	EVENT·COM	2	commentary	EVENT·COM	3	soliloquy	EVENT·COM	3	information	COM	2	letter	COM	3
graph	COM	3	eulogy	EVENT·COM	2	phonebook	COM	3	counterpoint	COM	2	program	EVENT·COM	2
prologue	EVENT·COM	2	itinerary	EVENT·COM	2	glossary	COM	3	worksheet	COM	2	greeting	EVENT·COM	3

Table A.7: Human annotation results and class assignments for EVT/*textsccom* class

A.3 Human annotation of regular polysemous CoreLex classes

The CoreLex lexical resource [Buitelaar, 1998] was built as a repository of regular polysemous alternations. As described in Chapter 2, this resource was constructed using on frequency counts of nouns that recurrently occurred as a member of more than one class in WordNet. Therefore, other types of lexical ambiguity combinations were also included[2]. The classes from the CoreLex data set annotated here are considered by the annotators to be prototypical examples of regular polysemous alternations and were used solely for the discussion in Section 5.3.

In order to further verify that our method was able to correctly classify nouns that are representative of regular polysemous alternations, we decided to identify which of the disemous class combinations available in CoreLex exemplified the theoretical characteristics of a regular polysemous alternation, according to the GL [Pustejovsky, 1995]. Each annotator was provided with a list of the disemous alternations specified in the CoreLex repository. We provided 5 examples of target words that instantiate the alternation from the [Boleda et al., 2012a] data set. The annotators were asked to indicate whether the examples provided for each alternation were representative of regular polysemy or another type of lexical ambiguity.

Table A.8 below presents the results that we obtained for each disemous alternation considered in CoreLex by each of our five human annotators. The Table is organized as follows: Column 1 contains the short-form of each alternation, while Column 2 presents the long-form name of the two classes that form the regular polysemous alternation. In Columns $3-7$, there are 5 example words for each alternation, extracted directly from the [Boleda et al., 2012a] data set, as explained above. These words provide lexical examples to the annotators without context so that they can objectively determine what kinds

[2] Examples of other types of lexical ambiguity include cases of homonomy, for instance, in which there is no relation between the two senses of a word. Consider, for example, the difference of meaning of the word *bank* in the two following contexts: **savings *bank*** and **river *bank***, in which there are two clearly different meanings, which therefore require distinct lexical entries. Although noted here for posterity, this phenomenon is beyond the scope of this thesis.

of words are available in WordNet in both of the classes that form the alternation[3].

Column 8 provides the majority decision of the five annotators, as to whether that given alternation is representative of a regular polysemous alternation. Column 9 provides the total number of annotators that assigned the majority vote. Column 10 presents the decision, which was *yes* only if all five annotators were in agreement. We decided to consider only those alternations that all 5 annotators marked to be regular polysemous. This is because we were interested in identifying only those most prototypical examples of the phenomenon.

After the results were analyzed, our human annotators fully agreed that 14 of the 60 original disemous regular polysemous alternations described in CoreLex [Buitelaar, 1998] (approximately 23% of the data set) were prototypical examples of regular polysemy.

[3] We did not provide any context to the annotators, so as not to bias the distinction between homonymy and regular polysemy. This is because all of the examples words provided *are* members of at least both of the classes forming the alternation. In this way, we were interested in having the annotators determine if there could be a relation between the two senses. (Because of this, we decided not to provide contextual examples for this task).

1	2	3	4	5	6	7	8	9	10
Abbrev.	ALTERNATION CLASSES	Example 1	Example 2	Example 3	Example 4	Example 5	(YES/NO)	TOTAL # AGR.	DECISION
ACT-ART	ACT-ARTIFACT	vignette	fresco	mend	improvisation	scan	YES	3	
ACT-ATR	ACT-ATTRIBUTE	betrayal	supplementation	disobedience	kindness	role	YES	3	
ACT-COM	ACT-COMMUNICATION	laughter	preaching	abuse	request	intro	YES	4	
ACT-EVT	ACT-EVENT	burial	easing	athletics	demolition	visitation	YES	5	RP
ACT-GRP	ACT-GROUP	following	mailing	traffic	legislation	parade	YES	4	
ACT-GRS	ACT-SOCIAL_GROUP	deputation	percussion	management	secession	delegation	YES	3	
ACT-HUM	ACT-HUMAN	arrival	minister	catcher	heroine	swagger	NO	0	
ACT-PHM	ACT-PHE0ME0N	flurry	flotation	breeze	transgression	sprinkle	NO	3	
ACT-POS	ACT-POSSESSION	allotment	spoil	holding	duty	atonement	YES	3	
ACT-PRO	ACT-PROCESS	growing	deflation	filtration	watering	pairing	YES	5	RR
ACT-PSY	ACT-PSYCHOLOGICAL_FEATURE	imposition	analogy	vaccination	imperialism	reorientation	NO	1	
ACT-STA	ACT-STATE	suffocation	diversification	participation	tumult	privation	YES	4	
ACT-TME	ACT-TIME	festival	regency	probation	continuance	leisure	YES	3	
AGT-HUM	AGENT-HUMAN	engineer	manipulator	shopper	jockey	promoter	YES	5	RP
ANM-ART	ANIMAL-ARTIFACT	stilt	kit	blower	turtle	rook	NO	0	
ANM-FOD	ANIMAL-FOOD	duckling	smelt	quail	carp	hare	YES	5	RP
ANM-HUM	ANIMAL-HUMAN	maverick	prey	predator	tiger	sheep	NO	0	
ART-ATR	ARTIFACT-ATTRIBUTE	piano	panache	glaze	fabric	still	NO	1	
ART-COM	ARTIFACT-COMMUNICATION	wire	fount	facade	well	directory	YES	3	
ART-EVT	ARTIFACT-EVENT	pic	serial	drip	shipwreck	grate	NO	2	
ART-FOD	ARTIFACT-FOOD	sausage	sub	lager	casserole	screwdriver	YES	3	
ART-FRM	ARTIFACT-FORM	prism	coil	flute	disc	rim	YES	5	RP
ART-GRP	ARTIFACT-GROUP	collage	motley	library	pantheon	repertory	YES	3	
ART-GRS	ARTIFACT-SOCIAL_GROUP	bastion	academy	divan	gang	gymnasium	YES	3	
ART-HUM	ARTIFACT-HUMAN	seeker	rocker	doll	organiser	tripper	NO	1	
ART-LOC	ARTIFACT-LOCATION	domicile	abode	mansion	roundabout	laundry	YES	5	RP
ART-LOG	ARTIFACT-GEO_LOCATION	tee	spa	apron	oasis	hearth	YES	4	
ART-NAT	ARTIFACT-NATURAL_BODY	radiator	ditch	curtain	plough	waterway	NO	1	
ART-PHO	ARTIFACT-PHYSICAL_OBJECT	tent	prop	widget	escarpment	trench	NO	2	
ART-POS	ARTIFACT-POSSESSION	manor	vat	bullion	store	hacienda	NO	2	
ART-PRT	ARTIFACT-PART	cistern	phone	claw	girdle	rostrum	NO	1	
ART-PSY	ARTIFACT-PSYCHOLOGICAL_FEATURE	credence	straitjacket	telecommunication	pitfall	magnet	NO	1	
ART-QUI	ARTIFACT-INDEFINITE_QUANTITY	raft	tub	bottle	keg	carton	YES	5	RP
ART-STA	ARTIFACT-STATE	hinge	bazaar	overdrive	limelight	maze	NO	3	
ART-SUB	ARTIFACT-SUBSTANCE	latex	linen	binder	asphalt	wicker	YES	5	
ATR-COM	ATTRIBUTE-COMMUNICATION	format	slur	hoot	publicity	leer	NO	2	
ATR-EVT	ATTRIBUTE-EVENT	glitter	discrepancy	glint	gleam	sparkle	YES	3	
ATR-PSY	ATTRIBUTE-PSYCHOLOGICAL_FEATURE	odour	chivalry	texture	pragmatism	relativity	YES	4	
ATR-REL	ATTRIBUTE-RELATION	odds	eccentricity	productivity	prevalence	inconsistency	NO	2	
ATR-STA	ATTRIBUTE-STATE	visibility	liability	degeneracy	uncertainty	optimism	YES	5	RP
COM-EVT	COMMUNICATION-EVENT	flick	genesis	prelude	chatter	broadcast	YES	5	RP
COM-HUM	COMMUNICATION-HUMAN	wanderer	flyer	counsel	cad	morse	NO	0	
COM-PSY	COMMUNICATION-PSYCHOLOGICAL_FEATURE	agenda	supposition	overtone	will	dictate	NO	1	
COM-STA	COMMUNICATION-STATE	reproach	disdain	acknowledgment	fugue	mystery	NO	3	
EVT-PSY	EVENT-PSYCHOLOGICAL_FEATURE	experience	fundamental	corollary	aetiology	instance	NO	2	
EVT-STA	EVENT-STATE	occurrence	triumph	malformation	affair	incident	YES	5	RP
FOD-HUM	FOOD-HUMAN	butter	honey	batter	frank	eater	NO	0	
fod-plt	FOOD-PLANT	currant	celery	potato	watercress	pineapple	YES	5	RP
GRP-GRS	GROUP-SOCIAL_GROUP	bunch	public	fleet	swarm	fraternity	YES	3	
GRP-PSY	GROUP-PSYCHOLOGICAL_FEATURE	zoology	underworld	jurisprudence	tableau	mythology	NO	1	
GRS-HUM	SOCIAL_GROUP-HUMAN	underwriter	dealer	bodyguard	acquirer	protestant	NO	2	
GRS-LOG	SOCIAL_GROUP-GEO_LOCATION	borough	metropolis	commonwealth	neighbourhood	parish	YES	5	RP
GRS-PSY	SOCIAL_GROUP-PSYCHOLOGICAL_FEATURE	humanism	democracy	christianity	genealogy	religion	NO	2	
HUM-NAT	HUMAN-NATURAL_BODY	creek	sculptor	firth	moor	marsh	NO	0	
HUM-PRT	HUMAN-PART	contractor	reverend	sigorina	bum	subordinate	YES	3	
HUM-PSY	HUMAN-PSYCHOLOGICAL_FEATURE	lord	paragon	trickster	successor	son	YES	3	
PHM-STA	PHEN0MENON-STATE	potential	aberration	fog	turbulence	polarization	YES	3	
PLT-SUB	PLANT-SUBSTANCE	maple	flax	sycamore	spruce	fir	YES	5	RP
PRO-STA	PROCESS-STATE	ulceration	bondage	fermentation	dehydration	glaciation	YES	5	RP
PSY-STA	PSYCHOLOGICAL_FEATURE-STATE	partiality	sensibility	pathology	predilection	feeling	YES	3	

Table A.8: Results and task description of the human annotation conducted to identify true examples of regular polysemous alternation from the disemous alternations described in the CoreLex data set [Buitelaar, 1998]

Bibliography

[Agirre et al., 2011] Agirre, E., Bengoetxea, K., Gojenola, K., and Nivre, J. (2011). Improving dependency parsing with semantic classes. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 699–703, Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

[Agirre et al., 2014] Agirre, E., López de Lacalle, O., and Soroa, A. (2014). Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, 40(1):57–84.

[Apresjan, 1974] Apresjan, J. (1974). Regular polysemy. *Linguistics*, 12(142).

[Atserias et al., 2005] Atserias, J., Padró, L., Rigau, G., and Salgado, J. G. (2005). An integrated approach to word sense disambiguation. In *Proceedings of the Recent Advances of Natural Language Processing RANLP-2005*.

[Azpeitia et al., 2014] Azpeitia, A., Cuadros, M., Gaines, S., and Rigau, G. (2014). NERC-fr: Supervised Named Entity Recognition for French. In Sojka, P., Horák, A., Kopeček, I., and Pala, K., editors, *Text, Speech and Dialogue*, volume 8655 of *Lecture Notes in Computer Science*, pages 158–165. Springer International Publishing.

[Baldwin, 2005] Baldwin, T. (2005). General-purpose lexical acquisition: Procedures, questions and results. In *Proceedings of the 6th Meeting of the Pacific Association for Computational Linguistics (PACLING-2005)*, pages 23–32.

[Baldwin and Bond, 2003] Baldwin, T. and Bond, F. (2003). Learning the countability of english nouns from corpus data. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 463–470, Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

[Bannard and Baldwin, 2003] Bannard, C. and Baldwin, T. (2003). Distributional models of preposition semantics. In *Proceedings of the ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, pages 169–180.

[Baroni et al., 2009] Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

[Baroni et al., 2014] Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 238–247.

[Baroni and Lenci, 2010] Baroni, M. and Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

[Baroni et al., 2010] Baroni, M., Murphy, B., Barbu, E., and Poesio, M. (2010). Strudel: A corpus-based semantic model based on properties and types. *Cognitive Science*, 34(2):222–254.

[Baroni and Zamparelli, 2010] Baroni, M. and Zamparelli, R. (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1183–1193, Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

[Bel, 2010] Bel, N. (2010). Handling of missing values in lexical acquisition. In (Conference, N. C., Chair), Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*.

[Bel et al., 2010] Bel, N., Coll, M., and Resnik, G. (2010). Automatic detection of non-deverbal event nouns for quick lexicon production. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 46–52, Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

[Bel et al., 2007] Bel, N., Espeja, S., and Marimon, M. (2007). Automatic acquisition of grammatical types for nouns. In *Human Language Technologies*

2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers, NAACL-Short '07, pages 5–8, Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

[Bel et al., 2012] Bel, N., Romeo, L., and Padró, M. (2012). Automatic lexical semantic classification of nouns. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, May 23-25, 2012*, pages 1448–1455.

[Bengio et al., 2003] Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

[Blacoe and Lapata, 2012] Blacoe, W. and Lapata, M. (2012). A comparison of vector-based representations for semantic composition. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 546–556, Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

[Bohnet et al., 2002] Bohnet, B., Klatt, S., and Wanner, L. (2002). An approach to automatic annotation of functional information to adjectives with an application to German. In *Proceedings of the 3rd LREC Conference, Workshop: Linguistic Knowledge Acquisition and Representation*, pages 24–33.

[Boleda et al., 2012a] Boleda, G., Padó, S., and Utt, J. (2012a). Regular polysemy: A distributional model. In **SEM 2012: The 1st Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*, pages 151–160, Montréal, Quebec, Canada. Association for Computational Linguistics.

[Boleda et al., 2012b] Boleda, G., Schulte im Walde, S., and Badia, T. (2012b). Modeling regular polysemy: A study on the semantic classification of Catalan adjectives. *Computational Linguistics*, 38(3):575–616.

[Breiman et al., 1984] Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, California, USA.

[Brent, 1993] Brent, M. R. (1993). From grammar to lexicon: Unsupervised learning of lexical syntax. *Computuational Linguistics*, 19(2):243–262.

[Buitelaar, 1998] Buitelaar, P. (1998). *Corelex: Systematic Polysemy and Underspecification*. PhD thesis, Brandeis University, Waltham, Massachusetts, USA.

[Buitelaar et al., 2005] Buitelaar, P., Cimiano, P., and Magnini, B. (2005). Ontology learning from text: Methods, evaluation and applications. *Computational Linguistics*, 32(4).

[Bullinaria, 2008] Bullinaria, J. A. (2008). Semantic categorization using simple word co-occurrence statistics. In *Proceedings of the "Bridging the gap between semantic theory and computational simulations" workshop at the European Summer School in Logic Language and Information (ESSLLI 2008), Hamburg, Germany, 4-15 August, 2008*, pages 1–8.

[Bullinaria and Levy, 2007] Bullinaria, J. A. and Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39(3):510–526.

[Bullinaria and Levy, 2012] Bullinaria, J. A. and Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming and SVD. *Behavior Research Methods*, 44(3):890–907.

[Burnard, 2007] Burnard, L. (2007). Reference guide for the british national corpus (XML edition), 2007.

[Bybee, 1985] Bybee, J. L. (1985). *Morphology: A study of the relation between meaning and form*, volume 9. John Benjamins Publishing.

[Bybee, 2007] Bybee, J. L. (2007). *Frequency of Use and the Organization of Language*. Oxford University Press.

[Bybee, 2010] Bybee, J. L. (2010). *Language, usage and cognition*. Cambridge University Press.

[Bybee and Hopper, 2001] Bybee, J. L. and Hopper, P. (2001). *Frequency and the Emergence of Language Structure*. John Benjamins, Amsterdam, The Netherlands.

[Carvalho and Ranchhod, 2003] Carvalho, P. and Ranchhod, E. (2003). Analysis and disambiguation of nouns and adjectives in Portuguese by FST. *Proceedings of the Workshop on Finite-State Methods for Natural Language Processing at EACL2003*, pages 105–112.

[Castellví Cabré et al., 2012] Castellví Cabré, M. T., Bach, C., and Vivaldi, J. (2012). 10 anys del corpus de l'IULA.

[Caudal, 1998] Caudal, P. (1998). Using complex lexical types to model the polysemy of collective nouns within the Generative Lexicon. In *Proceedings of the 9th International Workshop on Database and Expert Systems Applications, August 24-28, 1998*, pages 154–159.

[Celli and Nissim, 2009] Celli, F. and Nissim, M. (2009). Automatic identification of semantic relations in Italian complex nominals. In *Proceedings of the 8th International Conference on Computational Semantics*, pages 45–60. Association for Computational Linguistics.

[Chen et al., 2013] Chen, Y., Perozzi, B., Al-Rfou, R., and Skiena, S. (2013). The expressive power of Word Embeddings. *CoRR*, abs/1301.3226.

[Ciaramita and Altun, 2005] Ciaramita, M. and Altun, Y. (2005). Named-entity recognition in novel domains with external lexical knowledge. In *Proceedings of the NIPS Workshop on Advances in Structured Learning for Text and Speech Processing*.

[Cimiano and Wenderoth, 2007] Cimiano, P. and Wenderoth, J. (2007). Automatic acquisition of ranked qualia structures from the web. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*.

[Clark, 2014] Clark, S. (2014). *Vector space models of lexical meaning*. Blackwell, 2 edition.

[Collobert and Weston, 2008] Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 160–167, New York, New York, USA. ACM.

[Collobert et al., 2011] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

[Cooke and Gillam, 2008] Cooke, M. N. and Gillam, L. (2008). Distributional lexical semantics for stop lists. In *Proceedings of the 2008 BCS-IRSG Conference on Corpus Profiling*, IRSG'08, pages 5–5, Swinton, United Kingdom. British Computer Society.

[Cooper, 2005] Cooper, R. (2005). Do delicious lunches take a long time? In *Graduate School of Language Technology internal conference*.

[Copestake, 2013] Copestake, A. (2013). Can distributional approaches improve on Good Old-Fashioned Lexical Semantics? In *Proceedings of the 10th International Conference on Computational Semantics*, IWCS '13, pages 11–20, Potsdam, Germany. Association for Computational Linguistics.

[Copestake and Briscoe, 1995] Copestake, A. and Briscoe, T. (1995). Semi-productive polysemy and sense extension. *Journal of Semantics*, 12(1):15–68.

[Copestake and Herbelot, 2012] Copestake, A. and Herbelot, A. (2012). Lexicalised compositionality. *Unpublished draft*.

[Cruse, 2000] Cruse, D. A. (2000). Aspects of the micro-structure of word meanings. In Ravin, Y. and Leacock, C., editors, *Polysemy: Theoretical and Computational Approaches*, pages 30–51. Oxford University Press, Oxford, England.

[Cuadros and Rigau, 2006] Cuadros, M. and Rigau, G. (2006). Quality assessment of large scale knowledge resources. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 534–541, Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

[Davidson, 2002] Davidson, I. (2002). Understanding K-means non-hierarchical clustering. Technical report.

[de Marneffe et al., 2010] de Marneffe, M.-C., Manning, C. D., and Potts, C. (2010). "Was it good? It was provocative." Learning the meaning of scalar adjectives. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 167–176, Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

[de Marneffe et al., 2009] de Marneffe, M.-C., Padó, S., and Manning, C. D. (2009). Multi-word expressions in textual inference: Much ado about nothing. In *Proceedings of the 2009 Workshop on Applied Textual Inference*, TextInfer '09, pages 1–9, Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

[Dinu et al., 2013] Dinu, G., Pham, N. T., and Baroni, M. (2013). DISSECT - DIStributional SEmantics Composition Toolkit. In *51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, Proceedings of the Conference System Demonstrations, 4-9 August 2013, Sofia, Bulgaria*, pages 31–36.

[Dixon, 1982] Dixon, R. M. (1982). *Where have all the adjectives gone?*, volume 107. Walter de Gruyter, Berlin, Germany.

[Dorr and Jones, 1996] Dorr, B. J. and Jones, D. (1996). Role of word sense disambiguation in lexical acquisition: Predicting semantics from syntactic cues. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 1*, COLING '96, pages 322–327, Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

[Erk, 2012] Erk, K. (2012). Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653.

[Erk, 2013] Erk, K. (2013). Towards a semantics for distributional representations. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 95–106, Potsdam, Germany. Association for Computational Linguistics.

[Erk and Padó, 2008] Erk, K. and Padó, S. (2008). A structured vector space model for word meaning in context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 897–906, Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

[Evert, 2008] Evert, S. (2008). Corpora and collections. In Lüdeling, A. and Kytö, M., editors, *Corpus Linguistics. An International Handbook*. Mouton de Gruyter, Berlin, Germany, 58 edition.

[Ferrer, 2004] Ferrer, E. E. (2004). Towards a semantic classification of Spanish verbs based on subcategorisation information. In *Proceedings of the ACL 2004 Workshop on Student Research*, ACL Student '04, Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

[Fillmore et al., 2006] Fillmore, C. J., Narayanan, S., and Baker, C. F. (2006). What can linguistics contribute to event extraction? In *Proceedings of the 2006 AAAI Workshop on Event Extraction and Synthesis*, pages 18–23.

[Frisson, 2009] Frisson, S. (2009). Semantic underspecification in language processing. *Language and Linguistics Compass*, 3(1):111–127.

[Fu, 2009] Fu, G. (2009). Chinese named entity recognition using a morpheme-based chunking tagger. In *Proceedings of the 2009 International Conference on Asian Language Processing*, IALP '09, pages 289–292, Washington, DC, USA. IEEE Computer Society.

[Gillon, 1992] Gillon, B. S. (1992). Towards a common semantics for English count and mass nouns. *Linguistics and philosophy*, 15(6):597–639.

[Girju, 2009] Girju, R. (2009). The syntax and semantics of prepositions in the task of automatic interpretation of nominal phrases and compounds: A cross-linguistic study. *Computational Linguistics*, 35(2):185–228.

[Goldberg, 2006] Goldberg, A. E. (2006). *Constructions at Work: The Nature of Generalization in Language*. Oxford University Press.

[Grefenstette, 1994] Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Norwell, Massachusetts, USA.

[Griffiths et al., 2007] Griffiths, T. L., Steyvers, M., and Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological review*, 114(2):211.

[Grimshaw, 1990] Grimshaw, J. (1990). *Argument Structure*. MIT Press.

[Harris, 1954] Harris, Z. (1954). Distributional structure. *Word*, 10(23):146–162.

[Hatzivassiloglou and McKeown, 1997] Hatzivassiloglou, V. and McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the 8th Conference on European Chapter of the Association for Computational Linguistics*, EACL '97, pages 174–181, Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

[Hearst, 1998] Hearst, M. (1998). Automated discovery of word-net relations. In Fellbaum, C., editor, *WordNet: An Electronic Lexical Database and Some of Its Applications*, pages 131–153. The MIT Press, Cambridge, Massachusetts, USA.

[Hearst, 1992] Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, COLING '92, pages 539–545, Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

[Hindle, 1990] Hindle, D. (1990). Noun classification from predicate-argument structures. In *Proceedings of the 28th Annual Meeting on Association for Computational Linguistics*, ACL '90, pages 268–275, Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

[Hoeksema, 1986] Hoeksema, J. (1986). Monotonicity phenomena in natural language. *Linguistic Analysis*, 16:235–250.

[Huang et al., 2012] Huang, E. H., Socher, R., Manning, C. D., and Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 873–882, Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

[Jackendoff, 1973] Jackendoff, R. (1973). The base rules for prepositional phrases. In Anderson, S. and Kiparsky, P., editors, *Festschrift for Morris Halle*, pages 345–366. Holt, Rinehart and Winston, New York, New York, USA.

[Jackendoff, 1983] Jackendoff, R. (1983). *Semantics and Cognition*, volume 8. MIT Press, Cambridge, Massachusetts, USA.

[Jakobson, 1971] Jakobson, R. (1971). *Selected Writings: Word & Language*, volume 2. de Gruyter Mouton.

[Ježek and Lenci, 2007] Ježek, E. and Lenci, A. (2007). When GL meets the corpus: A data driven investigation of semantic types and coercion phenomena. In *Proceedings of the 4th International Workshop on Generative Approaches to the Lexicon (GL-2007), Paris, France, May 10-11, 2007.*

[Ježek and Melloni, 2011] Ježek, E. and Melloni, C. (2011). Nominals, polysemy and copredication. *Journal of Cognitive Science*, 12:1–31.

[Ježek and Vieu, 2014] Ježek, E. and Vieu, L. (2014). Distributional analysis of copredication: Towards distinguishing systematic polysemy from coercion. In *Proceedings of the 1st Italian Conference on Computational Linguistics CLIC-it*, pages 219–223, Pisa, Italy. Pisa University Press.

[Joanis et al., 2008] Joanis, E., Stevenson, S., and James, D. (2008). A general feature space for automatic verb classification. *Natural Language Engineering*, 14(3):337–367.

[Jones and Mewhort, 2007] Jones, M. N. and Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, 114:1–37.

[Katrenko and Adriaans, 2008] Katrenko, S. and Adriaans, P. (2008). Qualia structures and their impact on the concrete noun categorization task. *Proceedings of the "Bridging the gap between semantic theory and computational simulations" workshop at the European Summer School in Logic Language and Information - ESSLLI 2008, Hamburg, Germany, 4-15 August, 2008.*

[Kiela and Clark, 2014] Kiela, D. and Clark, S. (2014). A systematic study of semantic vector space model parameters. In *Proceedings of EACL 2014, Workshop on Continuous Vector Space Models and their Compositionality (CVSC).*

[Kilgarriff, 2003] Kilgarriff, A. (2003). Thesauruses for natural language processing. In *Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering*, pages 5–13.

[Korhonen, 2010] Korhonen, A. (2010). Automatic lexical classification: bridging research and practice. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 368(1924):3621–3632.

[Kotsiantis, 2007] Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. In *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*, pages 3–24, Amsterdam, The Netherlands. IOS Press.

[Krenn and Samuelsson, 1997] Krenn, B. and Samuelsson, C. (1997). The linguist's guide to statistics - don't panic.

[Landauer and Dumais, 1997] Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211–240.

[Landwehr et al., 2005] Landwehr, N., Hall, M. A., and Frank, E. (2005). Logistic model trees. *Machine Learning*, 59(1-2):161–205.

[Lapata, 1999] Lapata, M. (1999). Corpus-based induction of lexical representation and meaning. In *Proceedings of the 16th National Conference on Artificial Intelligence and the 11th Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence*, AAAI '99/IAAI '99, pages 948–, Menlo Park, California, USA. American Association for Artificial Intelligence.

[Lapata and Brew, 1999] Lapata, M. and Brew, C. (1999). Using subcategorization to resolve verb class ambiguity. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 266–274.

[Lee et al., 2001] Lee, G. G., Seo, J., Lee, S., Jung, H., Cho, B., Lee, C., Kwak, B., Cha, J., Kim, D., An, J., Kim, H., and Kim, K. (2001). Siteq: Engineering high performance QA system using lexico-semantic pattern matching and shallow NLP. In *Proceedings of The 10th Text REtrieval Conference, TREC 2001, Gaithersburg, Maryland, USA, November 13-16, 2001.*

[Lenci, 2011] Lenci, A. (2011). Composing and updating verb argument expectations: A distributional semantic model. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, CMCL '11, pages 58–66, Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

[Lenci, 2014] Lenci, A. (2014). Carving verb classes from corpora. *Word Classes: Nature, typology and representations*, 332:17.

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

[Lenci and Benotto, 2012] Lenci, A. and Benotto, G. (2012). Identifying hypernyms in distributional semantic spaces. In *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the 6th International Workshop on Semantic Evaluation*, SemEval '12, pages 75–79, Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

[Leonetti, 2004] Leonetti, M. (2004). Specificity and differential object marking in Spanish. In *Catalan Journal of Linguistics*, volume 3, pages 075–114.

[Levin, 1993] Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago, Illinois, USA.

[Levy and Goldberg, 2014a] Levy, O. and Goldberg, Y. (2014a). Dependency-based Word Embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, pages 302–308.

[Levy and Goldberg, 2014b] Levy, O. and Goldberg, Y. (2014b). Linguistic regularities in sparse and explicit word representations. In *Proceedings of the 18th Conference on Computational Natural Language Learning, CoNLL 2014, Baltimore, Maryland, USA, June 26-27, 2014*, pages 171–180.

[Li and Brew, 2008] Li, J. and Brew, C. (2008). Which are the best features for automatic verb classification. In *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA*, pages 434–442.

[Light, 1996] Light, M. (1996). Morphological cues for lexical semantics. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, ACL '96, pages 25–31, Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

[Lin, 1998a] Lin, D. (1998a). Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2*, ACL '98, pages 768–774, Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

[Lin, 1998b] Lin, D. (1998b). An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, ICML '98, pages 296–304, San Francisco, California, USA. Morgan Kaufmann Publishers Inc.

[Lin et al., 2003] Lin, D., Zhao, S., Qin, L., and Zhou, M. (2003). Identifying synonyms among distributionally similar words. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, IJCAI'03, pages 1492–1493, San Francisco, California, USA. Morgan Kaufmann Publishers Inc.

[Lund and Burgess, 1996] Lund, K. and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2):203–208.

[Martínez Alonso, 2013] Martínez Alonso, H. (2013). *Annotation of Regular Polysemy: An empirical assessment of the underspecified sense*. PhD thesis, University of Copenhagen.

[Martínez Alonso et al., 2013] Martínez Alonso, H., Pedersen, B. S., and Bel, N. (2013). *Annotation of regular polysemy and underspecification*, pages 725–730.

[Martínez Alonso and Romeo, 2014] Martínez Alonso, H. and Romeo, L. (2014). Crowdsourcing as a preprocessing for complex semantic annotation tasks. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014*, pages 229–234.

[McCarthy, 2000] McCarthy, D. (2000). Using semantic preferences to identify verbal participation in role switching alternations. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, NAACL 2000, pages 256–263, Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

[Merlo and Stevenson, 2001] Merlo, P. and Stevenson, S. (2001). Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*, 27(3):373–408.

[Mikolov et al., 2013] Mikolov, T., Yih, W., and Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013)*.

[Miller et al., 1990] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database*. *International Journal of Lexicography*, 3(4):235–244.

[Mitchell, 1997] Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, Inc., New York, New York, USA, 1 edition.

[Moon and Erk, 2013] Moon, T. and Erk, K. (2013). An inference-based model of word meaning in context as a paraphrase distribution. *ACM Transactions on Intelligent Systems and Technology*, 4(3):42:1–42:28.

[Nadeau and Sekine, 2007] Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

[Navigli, 2009] Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):10:1–10:69.

[Navigli, 2012] Navigli, R. (2012). A quick tour of word sense disambiguation, induction and related approaches. In *Proceedings of the 38th International Conference on Current Trends in Theory and Practice of Computer Science*, SOFSEM'12, pages 115–129. Springer-Verlag, Berlin, Heidelberg, Germany.

[Padó and Lapata, 2007] Padó, S. and Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

[Panchenko et al., 2012] Panchenko, A., Morozova, O., and Naets, H. (2012). A semantic similarity measure based on lexico-syntactic patterns. In Jancsary, J., editor, *Proceedings of the 11th Conference on Natural Language Processing, KONVENS 2012, Empirical Methods in Natural Language Processing, Vienna, Austria, September 19-21, 2012*, pages 174–178. ÖGAI.

[Pantel, 2005] Pantel, P. (2005). Inducing ontological co-occurrence vectors. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 125–132, Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

[Pecina et al., 2011] Pecina, P., Toral, A., Way, A., Prokopidis, P., Papavassiliou, V., and Giagkou, M. (2011). Towards using web-crawled data for domain adaptation in statistical machine translation. In Vadeghinste, M. F. H. D. V., editor, *Proceedings of the 15th Annual Conference of the European Association for Machine Translation*, pages 297–304, Leuven, Belgium.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[Poesio and Almuhareb, 2004] Poesio, M. and Almuhareb, A. (2004). Feature-based vs. property-based KR: An empirical perspective. In *Formal ontology in information systems: Proceedings of the 3rd conference (FOIS-2004)*, pages 177–184.

[Pustejovsky, 1995] Pustejovsky, J. (1995). *The Generative Lexicon*. MIT Press, Cambridge, Massachusetts, USA.

[Pustejovsky, 2001] Pustejovsky, J. (2001). Type construction and the logic of concepts. In Bouillon, P. and Busa, F., editors, *The Language of Word Meaning*, pages 91–123. Cambridge University Press, Cambridge, United Kingdom.

[Pustejovsky, 2005] Pustejovsky, J. (2005). A survey of dot objects. *Unpublished draft.*

[Pustejovsky, 2013] Pustejovsky, J. (2013). Type theory and lexical decomposition. In *Advances in Generative Lexicon Theory*, pages 9–38.

[Pustejovsky et al., 2004] Pustejovsky, J., Hanks, P., and Rumshisky, A. (2004). Automated induction of sense in context. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

[Pustejovsky et al., 2006] Pustejovsky, J., Havasi, C., Littman, J., Rumshisky, A., and Verhagen, M. (2006). Towards a generative lexical resource: The Brandeis Semantic Ontology. In *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006, 24-26 May 2006, Genoa, Italy*, volume 7.

[Pustejovsky and Ježek, 2008] Pustejovsky, J. and Ježek, E. (2008). Semantic coercion in language: Beyond distributional analysis. *Italian Journal of Linguistics*, 20(1):175–208.

[Pustejovsky and Rumshisky, 2008] Pustejovsky, J. and Rumshisky, A. (2008). Between chaos and structure: Interpreting lexical data through a theoretical lens. *International Journal of Lexicography*, 21(3):337–355.

[Quinlan, 1986] Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning*, 1(1):81–106.

[Rauh, 1993] Rauh, G. (1993). On the grammar of lexical and non-lexical prepositions in English. In Zelinsky-Wibbelt, C., editor, *The Semantics of Prepositions*, pages 99–150. De Gruyter, Berlin, Germany - Boston, Massachusetts, USA.

[Rigau et al., 1997] Rigau, G., Atserias, J., and Agirre, E. (1997). Combining unsupervised lexical knowledge methods for word sense disambiguation. In *Proceedings of the 8th Conference on European Chapter of the Association for Computational Linguistics*, EACL '97, pages 48–55, Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

[Ritter et al., 2011] Ritter, A., Clark, S., Mausam, and Etzioni, O. (2011). Named entity recognition in tweets: An experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1524–1534, Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

[Romeo et al., 2013a] Romeo, L., Alonso, H. M., and Bel, N. (2013a). Class-based word sense induction for dot-type nominals. In *Proceedings of the 6th International Conference on Generative Approaches to the Lexicon: Generative Lexicon and Distributional Semantic GL2013, September 24-24 2013 Pisa, Italy*, pages 76–83.

[Romeo et al., 2014a] Romeo, L., Lebani, G., Bel, N., and Lenci, A. (2014a). Choosing which to use? A study of distributional models for nominal lexical semantic classification. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014*, pages 4366–4373.

[Romeo et al., 2012] Romeo, L., Mendes, S., and Bel, N. (2012). Using qualia information to identify lexical semantic classes in an unsupervised clustering task. In *COLING 2012, Proceedings of the 24th International Conference on Computational Linguistics, Proceedings of the Conference: Posters, 8-15 December 2012, Mumbai, India*, pages 1029–1038.

[Romeo et al., 2013b] Romeo, L., Mendes, S., and Bel, N. (2013b). Towards the automatic classification of complex-type nominals. In *Proceedings of the 6th*

International Conference on Generative Approaches to the Lexicon: Generative Lexicon and Distributional Semantic GL2013, September 24-24 2013 Pisa, Italy, pages 21–28.

[Romeo et al., 2014b] Romeo, L., Mendes, S., and Bel, N. (2014b). A cascade approach for complex-type classification. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014*, pages 4451–4458.

[Romeo et al., 2014c] Romeo, L., Mendes, S., and Bel, N. (2014c). Using unmarked contexts in nominal lexical semantic classification. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 508–519.

[Rubenstein and Goodenough, 1965] Rubenstein, H. and Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

[Rumshisky et al., 2007] Rumshisky, A., Grinberg, V., and Pustejovsky, J. (2007). Detecting selectional behavior of complex types in text. In *Proceedings of the 4th International Workshop on Generative Approaches to the Lexicon (GL-2007), Paris, France, May 10-11, 2007.*

[Sahlgren, 2008] Sahlgren, M. (2008). The Distributional Hypothesis. *Italian Journal of Linguistics*, 20(1):33–54.

[Sánchez Valencia, 1991] Sánchez Valencia, V. M. (1991). *Studies on natural logic and categorial grammar*. PhD thesis, University of Amsterdam.

[Schulte im Walde, 2000] Schulte im Walde, S. (2000). Clustering verbs semantically according to their alternation behaviour. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 2*, COLING '00, pages 747–753, Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

[Schulte im Walde, 2006] Schulte im Walde, S. (2006). Experiments on the automatic induction of German semantic verb classes. *Computational Linguistics*, 32(2):159–194.

[Schütze, 1998] Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

[Schütze and Pedersen, 1995] Schütze, H. and Pedersen, J. (1995). Information retrieval based on word senses. In *Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175.

[Simon and Huang, 2010] Simon, P. and Huang, C. (2010). Cross-sortal predication and polysemy. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation, PACLIC 24, Tohoku University, Japan, 4-7 November 2010*, pages 853–861.

[Slonim et al., 2002] Slonim, N., Friedman, N., and Tishby, N. (2002). Unsupervised document classification using sequential information maximization. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pages 129–136, New York, New York, USA. ACM.

[Šnajder et al., 2013] Šnajder, J., Padó, S., and Agic, Ž. (2013). Building and evaluating a distributional memory for croatian. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 2: Short Papers*, pages 784–789.

[Snow et al., 2006] Snow, R., Jurafsky, D., and Ng, A. Y. (2006). Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 801–808, Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

[Socher et al., 2011] Socher, R., Lin, C. C., Ng, A. Y., and Manning, C. D. (2011). Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 129–136.

[Stevenson and Joanis, 2003] Stevenson, S. and Joanis, E. (2003). Semi-supervised verb class discovery using noisy features. In *Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 71–78, Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

[Stevenson and Merlo, 1999] Stevenson, S. and Merlo, P. (1999). Automatic verb classification using distributions of grammatical features. In *Proceedings of the 9th Conference on European Chapter of the Association for Computational Linguistics*, EACL '99, pages 45–52, Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

[Sun and Korhonen, 2009] Sun, L. and Korhonen, A. (2009). Improving verb clustering with automatically acquired selectional preferences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pages 638–647, Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

[Sun et al., 2008] Sun, L., Korhonen, A., and Krymolowski, Y. (2008). Verb class discovery from rich syntactic data. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, volume 4919 of *Lecture Notes in Computer Science*, pages 16–27. Springer Berlin Heidelberg.

[Taylor, 1993] Taylor, J. R. (1993). Prepositions: Patterns of polysemization and strategies of disambiguation. In Zelinsky-Wibbelt, C., editor, *The Semantics of Prepositions*, pages 151–176. De Gruyter, Berlin, Germany - Boston, Massachusetts, USA.

[Tkachenko and Simanovsky, 2012] Tkachenko, M. and Simanovsky, A. (2012). Named entity recognition: Exploring features. In Jancsary, J., editor, *Proceedings of the 11th Conference on Natural Language Processing, KONVENS 2012, Empirical Methods in Natural Language Processing, Vienna, Austria, September 19-21, 2012*, pages 118–127. ÖGAI.

[Toral et al., 2009] Toral, A., Monachini, M., Soroa, A., and Rigau, G. (2009). Studying the role of qualia relations for word sense disambiguation. In *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon GL2009, September 2009, Pisa, Italy*, pages 21–28.

[Tseng, 2001] Tseng, J. (2001). *The representation and selection of prepositions*. PhD thesis, University of Edinburgh, Edinburgh, United Kingdom.

[Turian et al., 2010] Turian, J., Ratinov, L., and Bengio, Y. (2010). Word representations: A simple and general method for semi-supervised learning. In *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden*, pages 384–394.

[Turney and Pantel, 2010] Turney, P. and Pantel, P. (2010). From frequency to meaning: Vector Space Models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.

[Utt and Padó, 2011] Utt, J. and Padó, S. (2011). Ontology-based distinction between polysemy and homonymy. In *Proceedings of the 9th International Conference on Computational Semantics*, IWCS '11, pages 265–274, Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

[van Benthem, 1991] van Benthem, J. (1991). *Language in Action Categories, Lambdas and Dynamic Logic*, volume 130 of *Studies in Logic and the Foundations of Mathematics*. Elsevier.

[Vázquez et al., 2000] Vázquez, G., Fernández, A., and Martí, M. A. (2000). Clasificación verbal alternancias de diátesis. *Quaderns de Sintagma*, 3.

[Witten and Frank, 2005] Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, California, USA, 2nd edition.

[Wu et al., 2008] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Philip, S. Y., et al. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37.

[Yamada et al., 2007] Yamada, I., Baldwin, T., Sumiyoshi, H., Shibata, M., and Yagi, N. (2007). Automatic acquisition of qualia structure from corpus data. *IEICE – Transactions on Information and Systems*, E90-D(10):1534–1541.

[Zipf, 1935] Zipf, G. (1935). *The Psychobiology of Language: An Introduction to Dynamic Philology*. M.I.T. Press, Cambridge, Massachusetts, USA.