eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Predicate Matrix: an Interoperable Lexical Knowledge Base for Predicates

Maddalen López de Lacalle Lekuona

PhD Thesis

informatika fakultatea
facultad de informática

Donostia, April 2023

Lengoaia eta Sistema Informatikoak Saila

eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Informatika Fakultatea

Predicate Matrix: an Interoperable Lexical Knowledge Base for Predicates

Thesis written by Maddalen López de Lacalle Lekuona under German Rigau i Claramunt and Egoitz Laparra Martín's guidance, presented to obtain the title of Doctor in Computer Science in the University of the Basque Country

Donostia, April 2023

Acknowledgements

Hasteko, eskerrak eman nahi dizkiot Elhuyarri, 2007an hizkuntza-teknologietan ikertzeko beka hura emateagatik. Josu, Xabier eta Igorrek egindako lehen elkarrizketa hura ondo gogoratzen dut. Esan daiteke hor hasi zela nire ibilbidea ikerketaren mundu honetan.

Eskerrak eman nahi dizkiet Elhuyarreko lankide guztiei, era askotara laguntzeagatik eta behar izan dudan guztietan zalantzarik egin gabe laguntzeko prest egoteagatik. Bereziki aipatu nahi dut Iñaki, laguntasunaren adibide ezin hobea. Baita Xabi ere, nire ibilbide osoan eta hasiera hasieratik emandako laguntzagatik eta pazientziaz erakutsitako guztiarengatik.

IXAko lantaldea ere eskertu nahi dut, bertan egon nintzen urtean eskuzabalik hartzeagatik. Etorkizunean ere lankidetza gehiagotan topo egitea espero dut. Itziar Aldabe, bereziki zuri, mila esker emandako laguntzagatik eta nitaz arduratzeagatik.

Quiero agradecer enormemente a los dos directores de esta tésis, German Rigau y Egoitz Laparra, su colaboración, ánimo y consejos durante este largo viaje. También por todos los *meets* de los jueves, y por las interesantes conversaciones mantenidas en los mismos sobre la tésis y más allá. Siempre se aprende algo con vosotros. Y a ti, Egoitz, gracias también por los madrugones de los jueves desde Tucson, Arizona.

Eskerrik asko bihotzez nire ama Marian eta aitta Andoniri txikitatik hasi eta gaur arte niretzat oinarri eta eredu izateagatik. Jaione eta Oier, zuek be bai.

Amama Ana, mila esker, beti laguntzeko prest. Unai, Martxel eta Martina: eskerrik asko danatik. Laguntza kontziente eta ez kontzientiatik. Zuek nere onduan egotia bihar-biharrezkua da neretzat.

Contents

INTRODUCTION

List of Figures

List of Tables

INTRODUCTION

CHAPTER 1

Introduction

In this introductory chapter, we first describe our research framework in Section 1.1 and we highlight the importance of linguistic resources to develop research in the field of Natural Language Processing (NLP). After that, Section 1.2 details the main goals of our research and the following Section 1.3 presents the main contributions of our research related with their corresponding chapters. Finally, the organization of the rest of the document is described in Section 1.4.

1.1 Research framework

Natural Language Processing (NLP) is one of the most relevant research areas in Artificial Intelligence (AI). Through NLP, computers are endowed with the ability to analyze, understand and generate human language, either written or spoken. Linguists and computer scientists, along with specialists from other fields or disciplines, have been working together for many years in the challenging task of developing systems capable of understanding human language, which allows humans to communicate with computers using normal, everyday language. Despite the inherent difficulty of many of the tasks performed, current NLP support allows many advanced applications which have been unthinkable only a few years ago. NLP is present in our daily lives, for example, through search engines, recommendation systems, virtual assistants, chatbots, text editors, text predictors, automatic translation systems, automatic subtitling, automatic summaries, inclusive technology, etc. Its rapid development in recent years predicts even more encouraging and also exciting results in the near future.

During the years, NLP has opened up a wide range of research lines that focus on investigating particular linguistic phenomena. Each of these tasks aims to provide a portion of information that contributes to a better interpretation of the text. In NLP, Lexical semantics studies the word meanings and how words structure their meaning. Words have different meanings based on the context where they appear, therefore, solving the word semantic ambiguity is one of the very first problems that NLP systems must face to understand a sentence. This task is called Word Sense Disambiguation (WSD) (Agirre and Edmonds, 2007) and consists of matching the words with their corresponding word sense in a specific dictionary, ontology or lexical knowledge base, such as WordNet (Fellbaum, 1998a). However, the semantics of the sentence does not only depends on the meaning of words that compose it. Semantic Role Labelling (SRL) (Gildea and Jurafsky, 2000, 2002) is a task that serves to find the semantic interpretation of a sentence by understanding "who did what to whom when and where". SRL detects the arguments associated with the predicates of a sentence and classify them into their specific semantic roles. In other words, the SRL task aims to model the predicate-argument structure of a sentence. Semantic resources describing predicate structures such as FrameNet (Baker et al., 1998a), PropBank (Palmer et al., 2005), NomBank (Meyers et al., 2004) and AnCora (Juan Aparicio and Martí, 2008) provide the knowledge on which the labeling of the predicate and its arguments will be based on.

Several systems have been developed for shallow semantic parsing an explicit and implicit semantic role labeling using these resources (Erk and Pado, 2006; Shi and Mihalcea, 2005; Giuglea and Moschitti, 2006; Laparra and Rigau, 2013). However, building large and rich enough predicate models for broad–coverage semantic processing takes a great deal of expensive manual effort involving large research groups during long periods of development. In fact, the coverage of currently available predicate-argument resources is still far from complete.

The recent emergence of new deep learning techniques have brought a disruptive change to the way NLP tasks are tackled. Large models pre-trained on huge collections of unannotated texts (Devlin et al., 2018; Yang et al., 2019; Lan et al., 2019; Liu et al., 2019) serve as starting-point for fine-tuning on a wide variety of downstream tasks achieving unprecedented results. Moreover, the generative capabilities of some of these models (Lewis et al., 2019; Raffel et al., 2020b; Radford et al., 2018) have boosted a paradigm shift on many NLP tasks, including SRL (He et al., 2017; Conia et al., 2021).

However, semantic resources such as WordNet, PropBank or FrameNet remain relevant and maintain the interest of the NLP community that continue working on projects investigating this type of knowledge bases (QasemiZadeh et al., 2019). Indeed, unless we resort to an unsupervised approach in any of the key lexical semantic tasks, from WSD to SRL, systems have to assign explicit labels from an existing inventory. And, traditional SRL is still considered as a fundamental step towards Natural Language Understanding (Navigli, 2018) and the number of benchmarks to evaluate SRL systems, including multilingual and cross-lingual settings, continue to grow (Tripodi et al., 2021).

1.2 Main goals

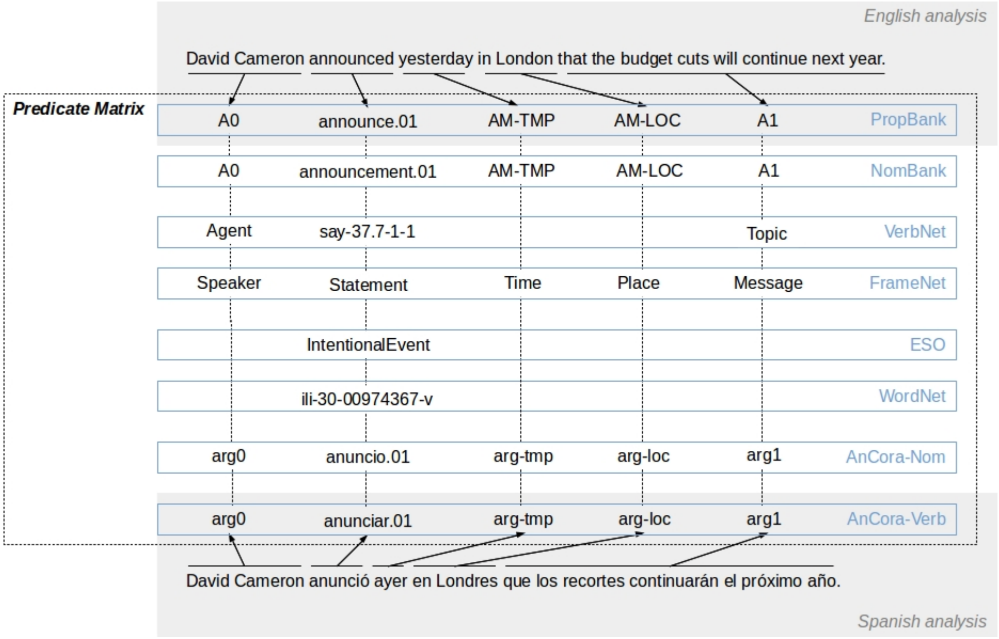
Large lexical resource projects such as FrameNet, VerbNet or PropBank are different attempts to formalize and encode predicative knowledge which have been later exploited by Natural Language Understanding systems, mostly through Semantic Role Labeling (SRL).

All these resources use different background predicate models guided by different design criteria and linguistic principles. At the same time, each of these resources offers some interesting characteristics not provided by its alternatives. Unfortunately, since they are developed independently and they are not integrated into a common platform, it becomes very difficult to exploit them jointly. However, a **common semantic framework** would allow the interoperability between all these tools and resources.

One of the few projects working on the integration of the predicate information is SemLink (Palmer, 2009) SemLink aimed to connect together different predicate resources such as FrameNet, VerbNet, PropBank and WordNet. Although it represents the greatest effort in this direction, since it was manually developed, SemLink has some limitations. Mapping all the resources together is a very costly process and, in consequence, its coverage is incomplete. Moreover, SemLink only takes into account English verbal predicates.

However, these limitations can be overcome by developing methods that automatize the resource integration in a more systematic manner. This is the motivation behind the building of the **Predicate Matrix**, a new lexical resource that incorporates multiple sources of predicate information to improve the interoperability between various semantic representations. Moreover, the Predicate Matrix extends its coverage beyond the English verbal predicates to their nominal counterparts and to different languages such as Spanish, Catalan and Basque.

A SRL-based system for event detection that identifies in texts what happened, who is involved, where and when, can exploit the Predicate Matrix to enrich its representation of the events obtaining a deeper semantic understanding. Besides, thanks to the cross-lingual interoperability provided by the Predicate Matrix, systems developed for different languages can contrast their outputs and conclude whether they correspond to the same events.



David Cameron announced yesterday in London that the budget cuts will continue next year.

A0	announce.01	AM-TMP	AM-LOC	A1	PropBank
A0	announcement.01	AM-TMP	AM-LOC	A1	NomBank
Agent	say-37.7-1-1			Topic	VerbNet
Speaker	Statement	Time	Place	Message	FrameNet
	IntentionalEvent				ESO
	ili-30-00974367-v				WordNet
arg0	anuncio.01	arg-tmp	arg-loc	arg1	AnCora-Nom
arg0	anunciar.01	arg-tmp	arg-loc	arg1	AnCora-Verb

David Cameron anunció ayer en Londres que los recortes continuarán el próximo año.

Figure 1.1: Example of cross-lingual and semantic interoperability provided by the Predicate Matrix.

For instance, Figure 1.1 provides the output of a SRL module based on PropBank for the English sentence *David Cameron announced yesterday in London that the budgget cuts will continue next year.* As PropBank is integrated into the Predicate Matrix, it is possible to obtain the corresponding predicate classes and roles for the rest of the predicate resources integrated in it. Thus, *David Cameron* identified as arg_0 role of the predicate **announce.01** by the PropBank-based SRL corresponds to the *Speaker* role of a **Statement** frame according to FrameNet or to the *Agent* role of the ***Say-37.7-1-1*** VerbNet verb class. The alignments between the different resources allows not only to enrich the annotation returned by a SRL system but also to merge the outputs given by models based on different schemes, e.g. PropBank and FrameNet.

The Predicate Matrix also allows to merge cross-lingual annotations. Figure 1.1 includes the output of an Ancora-based SRL system for the Spanish translation of the example above: *David Cameron anunció ayer en Londres que los recortes continuarán el próximo año.* Thanks to the mapping between Ancora and PropBank, it is possible to know that the Spanish verb *anunció* and its lemma **anunciar** are aligned to the PropBank predicate **anunciar.01**, the **Statement** frame from FrameNet, the ***Say-37.7-1-1*** VerbNet class and the rest of information included in the Predicate Matrix for this event description. Similarly, *David Cameron* identified by the Spanish SRL module as *arg0* role of the Ancora predicate **anunciar.01** which also corresponds to the *Speaker* role of a **Statement** frame. This way, it is possible to align the annotations returned by systems that work on different languages and to know if the sentences describe the same events.

A similar approach can be followed to include predicate nominalizations. For example, Figure 1.2 shows the NomBank-based (Meyers et al., 2004) annotation for the sentence *Steve Jobs gave his annual opening speech to the WWDC at Moscone Center, on Monday* that contains the nominal predicate **speech**. Thanks to the mappings in the Predicate Matrix, it is possible to obtain alignments to PropBank-based annotations to find out that two different sentences contain the same semantics.

In summary, the research we present on this dissertation aims to work on the automatic integration of multiple sources of predicate information in order to achieve a common semantic framework that would allow the interoperability between annotations based on different resources. Our research has the following main goals:

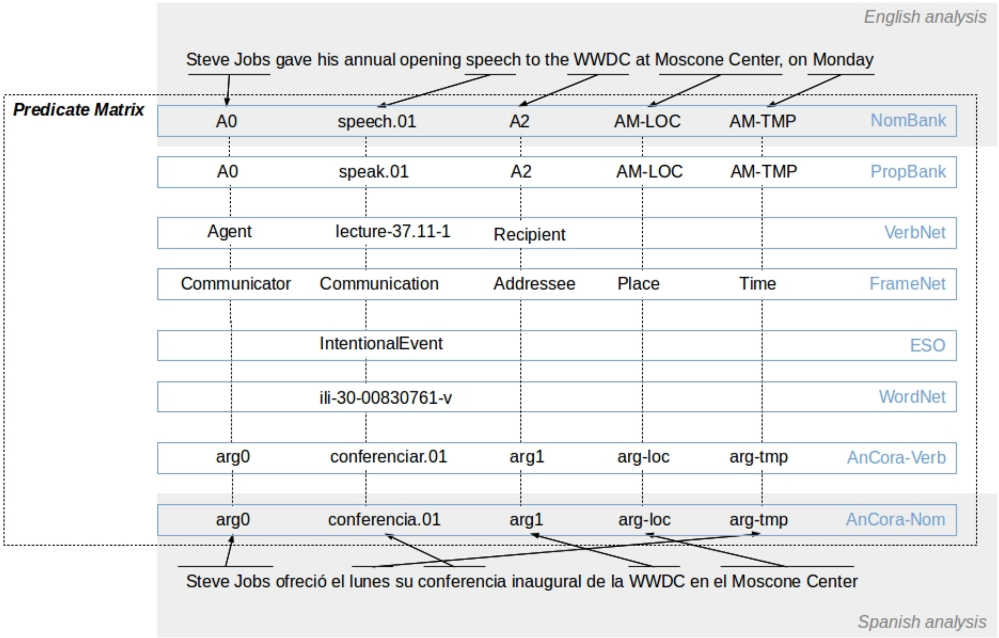


Figure 1.2: Example of semantic interoperability of a nominal predicate provided by the Predicate Matrix.

1. To define an automatic methodology to complete and extend the coverage of the mappings between the resources included in SemLink in a systematic way. The aim is to allow a more complete semantic interoperability between them. For that, we will work on the integration of predicate information at lexical and role levels.

2. To build the Predicate Matrix, a new lexical-semantic resource resulting from the integration of multiple multi-lingual sources of predicate information.

1.3 Main contributions

Integrating in a common platform different semantic resources that have been developed independently and are based on different design criteria and linguistic principles enables to exploit them together. Moreover, it allows the interoperability between resources that offer characteristics not provided by their alternatives. Nevertheless, both building manually large and rich enough predicate models and fully integrating them in a common semantic framework are very costly processes that requires a great deal of expensive manual effort. To avoid this manual effort, we propose an automatic methodology for mapping in a systematic way different semantic resources containing predicate information. The aim is to allow a more complete semantic interoperability between them. In order to achieve this aim, we have worked on the integration of predicate information at lexical and semantic role levels.

The approach for extending the lexical level mappings are centralized in WordNet in order to offer a wider coverage. For that, we apply graph-based Word Sense Disambiguation (WSD) algorithms in three different scenarios: (a) mappings between WordNet and VerbNet lexicons; (b) mappings between WordNet and FrameNet lexicons; and (c) mappings between WordNet and PropBank lexicons. The idea is to exploit WordNet to establish the appropriate correspondences among FrameNet, VerbNet and PropBank lexical entries at a WordNet sense level. In the case of FrameNet and VerbNet, graph-based WSD algorithms are applied to coherent groupings of words belonging to the same FrameNet frame or VerbNet class. For PropBank, the WSD approach is applied to a corpus annotated with information about basic semantic propositions.

Regarding the role level mappings, we concentrate our efforts to contribute on finding new mappings between FrameNet and VerbNet and between FrameNet and PropBank. On the one hand, we introduce an strategy that exploits already existing mappings and patterns of the examples that contain the resources for calculating role mappings frequencies in order to infer new role mappings between VerbNet and FrameNet roles. On the other hand, we study a corpus-based approach to extend the role mappings between FrameNet and PropBank.

In summary, some of the methods presented here are based on the use of existing information in these resources, while other methods are based on corpus annotations. In Figure 1.3 summarizes the methods applied at lexical level (left-side diagram) and role level (right-side diagram). The lines with a R refers to methods that obtain the mappings using the information contained in the resources and the C stands for methods that obtain the mapping using corpus information.

In addition, we have obtained further results derived from the work carried out during this investigation. The main contributions of this research

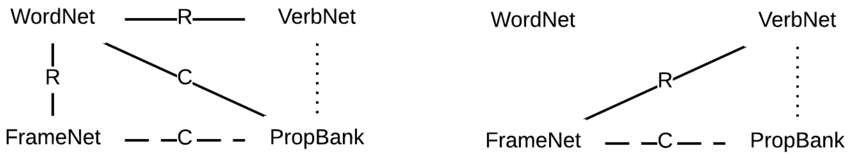


Figure 1.3: Type of automatic methods to obtain lexical and role mappings.

work can be listed as follows:

- The Predicate Matrix, a new lexical-semantic resource resulting from the integration of the new automatic mappings between resources and the mappings already offered by SemLink. See Chapter 3.
- A thorough overview of the state of the art about approaches for mapping lexical semantic resources and available prominent resources that model event and predicate-argument structures. See Chapter 2.
- A complete study of the coverage of the mappings included in SemLink. See Chapter 4.
- A new automatic methodology for creating automatic mappings between lexical entries and roles of multiple sources of predicate information including FrameNet, VerbNet, PropBank and WordNet. See Chapter 5.
- The extension of our resource to nominal predicates and other languages. As a result of including NomBank, AnCora-Nom, AnCora-Verb and the Basque Verb Index into the Predicate Matrix, we have obtained new mappings between these resources and VerbNet, FrameNet and WordNet. See Chapter 6.
- A review of the integration of the Predicate Matrix into different systems and projects, such as NewsReader (Vossen et al., 2016). See Chapter 8.

1.4 Organization of the document

This thesis presents the research we have carried out on automatic methods for mapping several existing predicate models on lexical and role level. Starting from a complete study of SemLink, one of the few predecessor projects working on the integration of predicate information, the following chapters describes our novel approaches and methods for mapping the semantic knowledge included in WordNet, VerbNet, PropBank and FrameNet, and, the projection of our resource to nominal predicates and languages other than English. The rest of this document is organized as follows:

- **Chapter 2: State of the art**

  This chapter presents a review of the state of the art of different research lines regarding Lexical Semantics. In particular, we have compiled the work concerning to the integration of lexical semantic resources. We first introduce some important concepts of the Lexical Semantics research area. After this, we present a deeper review of those predicate resources that are part of the Predicate Matrix. In addition, we give an overview of a variety of remarkable lexical semantic resources regarding verbal information. Finally, we include an overview of manual and automatic approaches for mapping different lexical semantic resources.

- **Chapter 3: A framework for Predicate Information Integration**

  This chapter summarizes the complete framework of the research that will be developed in the rest of chapters. It starts with an overview of the methodology to construct the new lexical resource resulting from the research presented in this work. After that, the new lexical semantic resource, the Predicate Matrix, is introduced. Finally, we present a schematic summary of the characteristics of the different versions of our resource.

- **Chapter 4: A study of SemLink coverage**

  In this chapter we present a study of SemLink coverage. We detail a study of the coverage of the mappings between each resource included in SemLink. Semlink uses Verbnet as a central resource, so for the analysis of the Semlink coverage we analyze the mappings between the following resource pairs: first, we analyze the alignments between WordNet and VerbNet, next, the coverage between PropBank and VerbNet is examined and finally, the coverage between FrameNet and VerbNet. We describe the coverage and gaps of these mappings with respect to the lexical entries and the role structures of each resource. The chapter finishes with some concluding remarks.

- **Chapter 5: Automatically extending the semantic interoperability between predicate resources**

  This chapter presents our approach to improve the interoperability between four semantic resources that incorporate predicate information. After a motivation of this work, the details of these techniques for creating automatic mappings between lexical entries and roles of WordNet, VerbNet, PropBank and FrameNet are further explained. We present empirical prove that our approach provides productive and reliable mappings in. Next, the new lexical semantic resource built applying this methodology is introduced. Finally, we present some concluding remarks about our approach.

- **Chapter 6: Nominalization and Multilingualism : extending to nominal predicates and to other languages**

  This chapter describes how to collect other resources to extend the predicate information included in the Predicate Matrix to languages other than English, in particular, SpanishAncoraVerb (Spanish), CatalanAncoraVerb (Catalan) and the Basque Verb Index (Basque) (Estarrona et al., 2015). Thus, we deal in a simple way with the problem of multilingualism. We also describe the strategy to extend the Predicate Matrix to include nominal predicates, in particular, by adding mappings to NomBank (English) and SpanishAncoraNom (Spanish). Finally, the resulting version of the Predicate Matrix is introduced, which provides a multilingual lexicon to allow interoperable semantic analysis in multiple languages.

- **Chapter 7: Using WordNet to extend the coverage of the PM**

  This chapter gives some insight into WordNet's exploitability to extend the coverage of the PM by exploiting its semantic relations hierarchy. First, we analyze the coverage of WordNet in the Predicate Matrix for its different types of verbs. Then, we study some straightforward methods to extend the mapping coverage of the Predicate Matrix through monosemous verbs, and the synonymy and hyperonimy relations from WordNet. We also show how semantic knowledge from the Multilingual Central Repository (MCR) (Atserias et al., 2004; Gonzalez-Agirre et al., 2012a) can be easily included in the PM and are presented some additional results concerning WordNet. Finally, The chapter finishes with some concluding remarks.

- **Chapter 8: Conclusion and Further work**

  Finally, in this chapter we draw the main conclusions of this research and present some possible future lines.

CHAPTER 2

State of the art

This chapter presents a review of the state of the art of different research lines regarding Lexical Semantics. In particular, we have compiled the work that has been done on the automatic integration of lexical resources. The chapter starts in Section 2.1 with an introduction to the Lexical Semantics research area. After this, we present a deeper review of those predicate resources that are part of the Predicate Matrix in Section 2.2. Next, in Section 2.3 other lexical-semantic resources regarding verbal information are introduced. Finally, in Section 2.4, we include an overview of manual and automatic approaches for mapping different lexical resources.

2.1 Lexical Semantics

Semantics is the study of how language encodes meaning so it can be used as a system of human communication. Although any linguistic expression carries several layers of meaning, the research work described in this document deals with the shallowest aspects of semantics like the meaning encoded in words and how they relate to sentence meaning through syntax. These are the subject matter of *lexical semantics* as well as of the present dissertation.

This section covers the basic theoretical concepts of lexical semantics and defines the fundamental terminology that will be used in the rest of the manuscript.

Word senses and lexical relations

Although the notion of the word *word* may be intuitive for any human speaker, finding a formal definition of such term is not a single task since it refers at the same time to both an utterance that plays a syntactic role and the concept it represents. Instead, we can introduce some terminology that will help to avoid ambiguities and provide a clearer landscape of the topic.

First, a *word* could be defined as the set of orthographic or phonological components that appear joined together in written or spoken language resulting in a particular form, i.e. a *word-form*. A set of *word-forms* that share the same base-form and the same underlying meaning are considered to be part of the same *lexeme*. By convention, a *word-form* of the set is chosen as the canonical form of the *lexeme*. This representative is called *lemma*. For example, the *word-forms* "solve", "solves" and "solved" belong to the same *lexeme* and from them, "solve" is used as the *lemma*. The specification of *lemmas* is critical for the construction of *lexicons*, large inventories of *lexemes* that represent the vocabulary of a particular language or knowledge branch. The *lemmas* are used as indices that ease organizing and querying those dictionaries.

As said above, all the *word-forms* associated with the same *lemma* represent the same meaning, but this meaning can vary greatly depending on the the context where they appear. Each individual meaning that can be referred by a *lemma* is called *word-sense*, or simply *sense*. For example, consider these two different uses of the *lemma* "arm" where it takes the meanings "a human limb" and "weapon" respectively: "*Arms* bend at the elbow", "They were licensed to carry *arms*".

Determining the number of *senses* that a *lemma* can have is very difficult. Different criteria are usually applied, for instance, two *word senses* are considered not to be the same if they have independent truth conditions, different syntactic behavior, independent sense relations, or exhibit antagonistic meanings.

The collection of *lemmas* that make up a *lexicon* can be organised on the basis of multiple types of connections or semantic relations that exist between their *word senses*. A possible strategy is to group *word senses* from a single domain or that commonly co-occur in real word situations into *semantic fields*, also called *lexical fields*. The intuition is that words belonging to the same semantic domain tend to exhibit strong semantic association between them. An alternative and complementary approach is to define more refined 1-to-1 semantic relations between pairs of *senses*. The set of these semantic relations studied in the literature is wide (Miller, 1995; Fellbaum, 1998b) and detailing all them is out of scope of this dissertation. By way of example, some of the most important ones are *synonymy*, *antonymy*, and *hypernymy*.

A *synonymy* relationship exists when two senses of two different lemmas are identical, or nearly identical (e.g a *"a little (or small)* group of scientists"). Whereas synonyms are *lemmas* with identical or similar meanings, antonyms are *lemmas* with an opposite meaning (e.g "a *big (not small)* group of scientists"). Hypernymy and hyponymy are taxonomic relations between word senses. A word sense is a hyponym of another word sense if the first is more specific, denoting a subclass of the other. For example, *swimming* is a hyponym of *sport* or *cat* is a hyponym of *animal*. Conversely, we say that *sport* is a hypernym of *swimming*, and *animal* is a hypernym of *cat*.

Event and predicate semantics

As mentioned above, besides addressing the meaning at the word-level, lexical semantics also studies how words relate with each other through syntax to give rise to sentences describing events or situations. The core component of such descriptions is the *predicate*, typically expressed by a verb or its nominalization, that carries the main semantic information of the event which is completed by the key participants that take part of it, the so-called *event participants* or *arguments*. All these elements along with the relations between them form the *argument structure*.

Lexical semantics studies how these structures generalizes over many surface forms of a sentence. For instance, an specific action of giving can be expressed in a variety of ways:

- Mary gives a ring to her mum.
- A ring is given to her mum by Mary.

- A ring is given by Mary to her mum.

All these paraphrases have the same semantic meaning, although they are expressed in diverse forms, encoding a giving event triggered by the predicate “give” that involves the arguments “Mary”, “her mum”, and “a ring”. In the three sentences, the predicate “give” has the same argument structure, with the same three *arguments*. Each argument of the structure plays a particular *semantic role* of the event. In the example above, “Mary” plays the the *giver* role, “her mum” is the *recipient* of the giving action and “a ring” is the *item* given. Besides the key arguments, the predicate “give” also can take other optional arguments expressing additional information of the event, such as the time or manner of the event, which are called *adjuncts*.

These structures allow to identify semantic commonalities between descriptions of different events. For example, in the following sentence, the predicate “give” has the same semantic roles as previously since “a ring” and “a letter” play the *item* role in the both cases:

- Mary gave her mum a letter.

Furthermore, semantic roles can be modeled into the more general *thematic roles* by identifying those that are shared across many event predicates and have a set of semantics properties in common. For example, the predicate “send” has a slightly different structure than the predicate “give”, like in the case bellow where “Mary” plays the role of the *sender*:

- Mary sent her mum a letter.

However, all the previous examples show that the *giver* and *sender* roles have a set of properties in common: their fillers (e.g. “Mary”) are usually animate (they are alive and sentient) and volitional (they choose to enter into the action). In contrast, the things that gets loaned or sent are usually not animate or volitional, furthermore, they remain unchanged by the event. Building on these intuitions, thematic roles generalize across predicates by leveraging the shared semantic properties of typical role fillers (Fillmore, 1967). For example, in examples above, “Mary” plays a similar role in all four sentences, which we call the *Agent*, reflecting several shared semantic properties: she is the one who is actively and intentionally performing the action, while “her mum” plays a more passive role and “the ring” or “the letter” are merely non-animate participants.

Thematic roles are one of the oldest linguistic models, but their modern formulation is due to Fillmore (1967) and Gruber (1965). There is no single universally agreed set of thematic roles, but some are very commonly used, such as: *Agent*, *Patient*, *Theme*, *Instrument* and *Goal*.

Thus thematic roles help us generalize over different surface realizations of predicate arguments. Although the *Agent* is usually realized as the subject of the sentence, in some cases, like when the verb is in passive voice, it can be the object, whilst the *Theme* or the *Instrument* are realized as the subject. Consider these two realizations of the arguments of the predicate “give”:

- Mary gives a ring.
- A ring is given by Mary.

In the former sentence, the subject (“Mary”) is the *Agent* of the action and the object (“a ring”) is the *Theme*. In the latter, the subject (“A ring”) plays the *Theme* role while the object (“Mary”) plays the *Agent*.

Just like in this example, many verbs allow their thematic roles to be realized in various syntactic positions. These multiple argument structure realizations are called *verb alternations* or *diathesis alternations*. Predicates can be clustered into *semantic classes* based on the kind of alternations they have in common. Levin (1993) listed, for a large set of English verbs, the semantic classes they belong to and the various alternations they participate in.

As mentioned, the generalization level provided by the semantic classes is attached to the syntactic realizations, but the criteria followed for grouping predicates can be further abstracted. Fillmore (1976a) proposed *frame semantics*, a linguistic theory where predicates targeting the same event or situation are arranged into schematic representations called *semantic frames*. These frames are defined by the semantic roles, known as *frame elements* in this paradigm, that take part of the situation described. Since the frames are defined as semantically coherent structures, the approach allows to incorporate ontological knowledge by defining frame-to-frame semantic relations.

Predicate-argument structures

Three well known *predicate-argument structures* models to represent event semantics are *VerbNet* (Kipper, 2005), *PropBank* (Palmer et al., 2005) and *FrameNet* (Baker et al., 1998a).

VerbNet is a lexicon of verbs which includes thirty “core” thematic roles played by arguments to these verbs. VerbNet organizes these roles in a hierarchy, so that a *Topic* is a type of *Theme* , which in turn is a type of *Undergoer*, which is a type of *Participant*, the top-level category. In addition, VerbNet organizes verb senses into a class hierarchy, in which verb senses that share similar core semantic and syntactic properties are grouped together. Each VerbNet class or subclass takes a set of thematic roles.

It has proved very difficult to come up with a standard set of roles and to produce a formal definition of these roles. For example, consider the *Agent* role; in most cases *Agents* are animate, volitional, sentient, sentient, causal, but it may happen that any individual noun phrase does not exhibit all of these properties. So, detailed thematic role inventories of the sort used in VerbNet are not universally accepted. These problems have led most research to alternative models of semantic roles. One such model is based on defining generalized semantic roles that abstract over the specific thematic roles. For example, *Proto-Agent* and *Proto-Patient* are generalized roles that express roughly agent-like and roughly patient-like meanings. These roles are defined, not by necessary and sufficient conditions, but rather by a set of heuristic features that accompany more agent-like or more patient-like meanings. Thus the more an argument displays agent-like properties (intentionality, volitionality, causality, etc.) the greater likelihood the argument can be labeled as *Proto-Agent.*

In addition to using proto-roles, many computational models avoid the problems with thematic roles by defining semantic roles that are specific to a particular verb, or specific to a particular set of verbs or nouns. Two other very commonly used lexical resources which make use of some of these alternative versions of semantic roles are PropBank, that uses both proto-roles and verb-specific semantic roles, and FrameNet, that uses frame-specific semantic roles.

In order to avoid the difficulty of establishing a common universal set of thematic roles, PropBank defines the semantic roles with respect to an individual verb sense. Each of these verb senses are linked to a list of numbered arguments where arg_0 is the proto-agent and arg_1 the proto-patient. The semantics of the rest of the roles will be related only to the particular verb sense. Thus, the arg_2 of one verb most probably has nothing to do with the arg_2 of a another verb.

While semantic roles in PropBank are specific to an individual verb sense, roles in FrameNet project are specific to a *frame*. A frame describes situations or events based on a script-like structure, which instantiates a set of frame-specific semantic roles played by the event participants and called *frame elements*.

FrameNet also relates frames and frame elements. Frames can inherit from each other, and generalizations among frame elements in different frames can be captured by inheritance as well. Rather than linking semantic roles such as *Sender* or *Giver* into thematic roles such as *Agent*, FrameNet groups verbs (or words) that trigger a kind of event into the same frame, and links semantically-related roles across frames. For example, the sentences *"Karl carried the books to the library on his head"* and *"The books were brought by Karl on his head to the library"* would be annotated identically by FrameNet, as *carry* and *bring* are both *lexical units* in the **Bringing** frame.

Semantic roles gave us a way to express some of the semantics of an argument in its relation to the predicate. Another way to express semantic constraints on arguments are the *selectional restrictions*. These are a kind of semantic type constraint that a verb imposes on the kinds, or categories, of concepts that are allowed to fill its argument roles. In other words, the predicates limit the semantic content of their arguments. Consider the sentence *"Tom drank a computer"*; the predicate drank selects an object argument that is a liquid or is liquid-like hence, the argument a car contradicts the selectional restrictions of the predicate drank.

In the following section we will introduce in more detail the three main predicate resources for English briefly mentioned above as well as other predicate semantics resources integrated in the Predicate Matrix.

2.2 Predicate Models

Several lexical-semantic resources describing predicate and role structures have been developed by different research groups. These resources are built on different theories and paradigms resulting on diverse and heterogeneous semantic representations focusing on predicate-argument structures. For instance, the role descriptions contained in these resources vary from syntactic arguments to thematic-roles. Certain predicate resources include only the verbal form of the predicates described, but some of the existing resources also take into account their nominalizations or other forms of the predicates.

In the following subsections we describe the structure and the semantic information offered by the different sources of predicate information integrated in the Predicate Matrix.

2.2.1 WordNet

WordNet[1] (Fellbaum, 1998a) is a large lexical knowledge base for English. It contains manually coded information about English nouns, verbs, adjectives and adverbs and it is organized around the notion of *synset* (synomyn set). A synset is a set of words with the same part-of-speech that refers to the same concept and can be interchanged in a certain context. For example, <*wage*, *pay*, *earnings*, *remuneration*, *salary*> form a synset because they can be used to refer to the same concept, as shown in Table 2.1. A synset is often further described by a gloss, e.g. *"something that remunerates"*, and, in most cases, it also contains one or more usage examples: *"wages were paid by check"*. Each synset is related with other synsets by means of explicit semantic relations, including hypernymy/hyponymy, meronymy/holonymy, antonymy, entailment, etc.

The synsets are organized into lexicographer files based on syntactic category and logical groupings. For instance, the *noun.possession* lexicographer file contains *nouns denoting possession and transfer of possession*, and *verb.possession* groups *verbs of buying, selling, owning.*

WordNet was designed as a lexico-semantic resource, and contains little syntactic information. However, verbal senses also encode sentence frames. There are 35 different sentence frames which illustrate the types of simple syntactic schemes in which the verbs in the synset can be used. Although the sentence frames are assigned to the synset, some of them are specific of a single member of the synset. Nevertheless, as they have not a direct

[1] `http://wordnet.princeton.edu/`

pay$^{1}_{n}$	wage, pay, earnings, remuneration, salary (Something that remunerates) *”Wages were paid by check”*; *”He wasted his pay on drink”*; *”They saved a quarter of all their earnings”* - lexicographer file: noun.possession
pay$^{1}_{v}$	pay (Give money, usually in exchange for goods or services) *”I paid four dollars for this sandwich”* ; *”Pay the waitress, please”* - lexicographer file: verb.possession - sentence frames: Somebody —-s Somebody —-s something Somebody —-s somebody Somebody —-s somebody something Somebody —-s something to somebody Somebody —-s somebody with something
pay$^{2}_{v}$	give, pay (Convey, as of a compliment, regards, attention, etc.; bestow) *”Don't pay him any mind”*; *”Give the orders”*; *”Give him my best regards”*; *”Pay attention”* - lexicographer file: verb.communication - sentence frames: Somebody —-s something Somebody —-s somebody something Somebody —-s something to somebody

Table 2.1: WordNet synsets for the lemma **pay**.

correspondence to more abstract semantic roles, they are not included into this study.

WordNet is one of the lexical knowledge-bases most widely-used by NLP researchers and developers. It has been applied to a large variety of knowledge-based NLP tasks such as semantic tagging, information retrieval or document classification. WordNet constitutes a large coverage sense inventory that have been used in several corpus annotations projects (Segond et al., 1997; Miller et al., 1993). In Information Retrieval(IR) (Gonzalo et al., 1998; Varelas et al., 2005; Meštrović and Calì, 2016) synonymy relations are used for query expansion to improve the recall of IR systems and cross-language synset correspondences are used for Cross-Language Information Retrieval (Agirre et al., 2008). Incorporating knowledge from WordNet hypernyms and senses leads to improvements on Document Classification (Scott and Matwin, 1998; Liu et al., 2007; Elhadad et al., 2017). Nevertheless, the main use of Wordnet has been in the area of Word Sense Disambiguation (WSD) (Burchardt et al., 2005; Agirre et al., 2009; Rutkowski et al., 2019; Vial et al., 2019), the identification of the most suitable meaning of a word in a given context.

WordNets have been built for several languages. Most of them are also linked to the original Princeton WordNet. The first effort for linking WordNets for different languages was EuroWordNet (Vossen, 1998a) project. Its purpose is to interconnect WordNets for several European languages through an interlinguistic index (called ILI). MultiWordNet (Pianta et al., 2002) and BalkaNet (Tufis et al., 2004) are also different efforts on creating, enriching, and maintaining WordNets for different languages. The Multilingual Central Repository (Gonzalez-Agirre et al., 2012a) extends these works connecting WordNets to other kinds of semantic resources as SUMO (Pease et al., 2002; Álvez et al., 2012) or the Top Ontology (Álvez et al., 2008). Given the increasing number of WordNets for languages other than English, the Global WordNet Association was founded to share and link WordNets for all languages in the world, and to promote the research related to these resources. These efforts continue today. For example building the Cantonese WordNet (Sio and da Costa, 2019) or new WordNets for some African languages Griesel et al. (2019).

Predicate **pay**, roleset 1:

arg_0 : the payer or buyer
arg_1 : the money or attention
arg_2 : the person being paid, destination of attention
arg_3 : the commodity, paid for what

[arg_0 Theatres] **pay** [arg_2 movie producers] [arg_3 for showing their films].

[arg_0 Investors] **pay** [arg_1 higher prices] [arg_3 for country funds].

Table 2.2: PropBank roleset and annotated sentences for sense 1 of the verb **pay**.

2.2.2 PropBank

PropBank[2] (the English Proposition Bank) aims to provide a wide corpus annotated with information about semantic propositions, including relations between the verbal predicates and their arguments. It is the result of more than 112,000 semantic annotations provided by Palmer et al. (2005) over the syntactic structures of the Penn TreeBank (Marcus et al., 1993) dependency parses of the Wall Street Journal corpus. PropBank also contains a description of the frame structure, called *roleset*, of each verbal sense from its lexicon. So, a polysemous verb would have a different frame or roleset for each sense that fits to its specific semantics. Rolesets are considered coarse-grained sense distinctions. This lexicon contains up to 3,256 different verbs. Unlike other similar resources, PropBank defines the arguments, or roles, of each verb individually and does not encode explicit relations between arguments of different predicates. In consequence, obtaining a generalization of the frame structures over the verbs becomes a hard task.

For example, consider the frame in Table 2.2 taken from PropBank for the sense 1 of the verbal predicate **pay** and two annotated sentences included in the corpus. In this case, the argument arg_0 represents *the payer or the buyer*, the argument arg_1 *the money or attention*, and the argument arg_3 *the commodity*). As shown Table 2.3, the arguments of the sense 1 of the predicate **charge** have the same meaning. Specifically, both predicates share

[2] `http://verbs.colorado.edu/~mpalmer/projects/ace.html`

Predicate **charge**, roleset 1:

arg_0 : the seller
arg_1 : the asking price
arg_2 : the buyer
arg_3 : the commodity

[arg_0 Movie producers] **charge** [arg_2 theatres] [arg_3 for showing their films].

[arg_2 Investors] are **charged** [arg_1 higher prices] [arg_3 for country funds].

Table 2.3: PropBank roleset and annotated sentences for sense 1 of the verb **charge**.

arguments referring to *the seller*, *the money or attention*, *the buyer*, and *the commodity* but, even though both predicates share the arguments buyer and seller arguments, they are not explicitly related since the descriptions of the roles are not systematic.

As already mentioned, the original PropBank corpus was created hand-annotating the syntactic trees of the Penn Treebank (Marcus et al., 1993) with predicate-argument structures. Since then, other corpora have been annotated with PropBank verb senses and semantic roles. One of the most well-known is Ontonotes[3] (Marcus et al., 2011) which comprises various genres of text (news, conversational telephone speech, weblogs, broadcast, talk shows) in three languages (English, Chinese, and Arabic). Furthermore, DARPA-BOLT(Broad Operational Language Translation)[4], NIH (National Institude of Health) and Google have used PropBank to annotate other kind of text such as SMS conversations, question answering corpora, the English Web Treebank, and even clinical notes (Paek et al., 2006). More recently, Moon et al. (2018) added semantic role labels and senses of verbs to an existing corpus of adult-child dialogues following PropBank guidelines.

PropBank resources have also been developed for languages other than English. For instance, Chinese, Finnish, French, German, Italian, Portuguese and Spanish. All of them are included in the IBM Universal Proposition

[3] https://catalog.ldc.upenn.edu/LDC2013T19

[4] https://www.ldc.upenn.edu/collaborations/current-projects/bolt

Banks[5] (Akbik and Li, 2016). They all use the frame and role labels from the English PropBank. Language-specific lexicons with PropBank annotations are available for languages as Hindi (Vaidya et al., 2013), Chinese (Xue, 2006), Arabic (Zaghouani et al., 2010), Finnish (Haverinen et al., 2014), Portuguese (Duran and Aluísio, 2012), Basque (Basque Verb Index (BVI)) Estarrona et al. (2015) or Turkish (Şahin and Adalı, 2018) among others. Màrquez et al. (2007) create a PropBank for Spanish and Catalan similar to the English PropBank. Nowadays, proposition banks that follow the English PropBank scheme are being developed for new languages taking into account language-specific properties, like the Russian Proposition Bank (Moeller et al., 2020). Daza and Frank (2020) translate English CoNLL-09 dataset using Machine translation into French, German and Spanish and project predicate and role annotations to the target languages using multilingual BERT (Devlin et al., 2018). As a result they have automatically constructed a parallel corpus in four languages with unified annotations.

PropBank is a commonly used gold standard for shallow semantic labeling. Since it was developed, PropBank soon became the most commonly used schema for the Semantic Role Labelling (SRL) task due to the success of the CoNLL challenges (Carreras and Màrquez, 2004; Carreras and Màrquez, 2005; Surdeanu et al., 2008; Hajič et al., 2009).

2.2.3 NomBank

NomBank[6] corpus (Meyers et al., 2004) provides argument structures for instances of common nominal predicates over the same Penn TreeBank corpus (Marcus et al., 1993) annotated by PropBank for verbal predicates. NomBank and PropBank follow the same annotation scheme and there is a coordinated effort between both projects to guarantee that role definitions keep consistent across parts of speech, whenever it is possible. Although the aim is to annotate every argument-taking noun in the Penn TreeBank corpus, most of the predicates annotated in NomBank are deverbal nominalizations derived from verbs already annotated in PropBank. These nominalizations share the argument structure with their verbal counterpart and, as a consequence, they inherit the rolesets and argument definitions from their corresponding verbal

[5] `https://github.com/System-T/UniversalPropositions`

[6] `https://nlp.cs.nyu.edu/meyers/NomBank.html`

Predicate **payment**, roleset 1:

arg_0 : the payer or buyer
arg_1 : the money or attention
arg_2 : person being paid, destination of attention
arg_3 : commodity, paid for what

Table 2.4: NomBank roleset for sense 1 of the nominal predicate **payment**.

The **payments** of [arg_0 theatres] to [arg_2 movie producers] [arg_3 for showing their films].

The **payments** [arg_3 for country funds] have [arg_1 higher prices] for [arg_0 investors].

Table 2.5: NomBank annotations for the predicate **payment**.

forms. For example, PropBank's frame file for the verb **pay** was used in the annotation of the noun **payment**. Consequently, the frame of the nominal predicate **payment.01** from NomBank shown in Table 2.4 matches the same frame of its verbal form **pay.01** from PropBank shown in Table 2.2.

The consistency across nominal and verbal predicate frames makes it possible to preserve the alignment in the verbal and nominal annotations of the same predicates. For example, in the example shown in Table 2.5, all the arguments of the nominal predicate **payment.01** from NomBank are in line with the arguments of the verbal predicate **pay.01** from PropBank.

NomBank 1.0 release includes a total of 114,576 propositions, and these annotated instances covers a list of 4,704 different nominal predicates which include several types of argument-taking nouns.

In the same way a significant number of studies on the task of verbal semantic role labeling (SRL) have focused on annotating resources such as PropBank, the development of the NomBank corpus inspired further investigations into semantic role labeling of nominal predicate-argument structure. Jiang and Ng (2006) and Liu and Ng (2007) applied machine learning methodologies and feature representations previously shown useful in PropBank-based verbal Semantic Role Labeling for NomBank-based automatic Semantic Role Labeling. Gerber and Chai (2008) also approached nom-

Class pay-68:		
	Members:	**serve, spend, squander, tithe, waste**
	Thematic-roles:	*Agent, Asset, Theme*
Subclass 68-1:		
	Members:	**pay, repay**
	Thematic-roles:	*Recipient*

Table 2.6: Description of the VerbNet class ***pay-68***.

inal SRL using NomBank. Recent neural SRL systems also use the NomBank dataset (Li et al., 2019; Xia et al., 2019).

2.2.4 VerbNet

VerbNet[7] (Kipper, 2005) is a hierarchical domain-independent and broad-coverage verb lexicon for English. VerbNet is organized into verb classes which are defined based on its syntactic and semantic behaviour. The resource is based on the hypothesis by Levin (1993) that verbs with similar semantics share similar syntactic properties. Kipper (2005) refined the Levin classes and added subclasses in order to achieve syntactic and semantic coherence among members of a class. Each of these verb classes in VerbNet is described by thematic-roles, selectional restrictions on the arguments, and frames consisting of an example sentence, a syntactic description in which thematic-roles are mapped to syntactic complements, and semantics that indicate how the participants are involved in the event. VerbNet groups up to 5,257 verb senses into 274 classes. For instance, the VerbNet class ***pay-68*** shown in Table 2.6 groups verbs related to *paying* acts.

Table 2.6 shows how the members of the subclass ***pay-68-1*** inherit the semantics from its parent class ***pay-68***. Subclass ***pay-68-1*** includes a more specific set of events grouped together because they share an additional thematic role, *Recipient*. Thus, instances of the members of the subclass ***pay-***

[7] `http://verbs.colorado.edu/~mpalmer/projects/verbnet.html`

[*Agent* Theatres] **pay** [*Recipient* movie producers] [*Theme* for showing their films].

[*Agent* Investors] **pay** [*Asset* higher prices] [*Theme* for country funds].

Table 2.7: Example annotations for the verbal predicate **pay**.

Selectional restrictions:

Agent:	animate or organization
Asset:	currency

Semantics :

has_possession(start(**Event**), *Agent*, *Asset*)
NOT_has_possession(start(**Event**), *Agent*, *Theme*)
has_possession(end(**Event**), *Agent*, *Theme*)
NOT_has_possession(end(**Event**), *Agent*, *Asset*)
transfer(during(**Event**), *Theme*)

transfer(during(**Event**), *Asset*)

Table 2.8: Information encoded in VerbNet for class ***pay-68***.

68-1 such as **pay** can be annotated as in the example sentences shown in Table 2.7.

Although VerbNet does not provide any corpus with annotations, this resource is very rich since it encodes semantic descriptions of the events that represent each class and also includes as selectional preferences a set of semantic types, hierarchically classified, for some of the roles of the classes.

Table 2.8 shows the information encoded in the VerbNet class ***pay-68***. Its semantics describe the exchanging process of the *paying* event of the examples in Table 2.7. In the first sentence, an *organization* type entity (*Theatres*) plays the role *Agent*. In this case, it is not clear if the role *Recipient* is played by an *animate* or an *organization* entity (*Movie producers*). It depends on whether it refers to a company or a person as a producer. At the start of the **paying** event the *Agent* possesses an *Asset* which will exchange during the event for the *Theme* (*for showing their films*). Consequently, when the event is finished, the *Agent* possesses the *Theme* but not the *Asset*. In the semantics described in VerbNet for the predicates in class ***pay-68*** and its subclass ***pay-68-1***, the role *Recipient* is not specified. But it can be deduced that the *Recipient* is the one that possesses the *Asset* at the end of the event after exchanging it for the *Theme* with the *Agent*.

VerbNet-based corpora have been developed adopting the general structure and content elements of the English VerbNet for various languages such as Arabic (Mousser, 2011), French (Falk et al., 2012), German (Mújdricza-Maydt et al., 2016), Dutch (Monachesi et al., 2007) , Urdu (Hautli et al., 2015), Mandarin (Liu, 2020), Italian (Busso and Lenci, 2016), Portuguese (Scarton et al., 2014) etc.

The VerbNet lexical resource has been used to support numerous NLP tasks due to its high coverage, useful verb groupings and systematic coding of thematic roles. Most notably, it has been used in semantic role labeling (Swier and Stevenson, 2004; Shi and Mihalcea, 2005; Di Fabio et al., 2019a) that in turn can be applied in many other application such as question-answering (Wen et al., 2008; Clark et al., 2018) or event extraction (Exner and Nugues, 2011). VerbNet has been used to aid many other NLP tasks. For example, Word Sense Disambiguation (Dang, 2004; Brown et al., 2014; Abend et al., 2008), automatic verb acquisition in spoken dialog systems (Swift, 2005) or building conceptual graphs (Hensman and Dunnion, 2004). However, Zapirain et al. (2008) propose to use automatic PropBank SRL for core role identification and then converting the PropBank roles into more meaningful VerbNet roles heuristically.

2.2.5 FrameNet

FrameNet[8](Baker et al., 1998a) is a very rich semantic resource that contains descriptions and corpus annotations of 12,940 English words following the Frame Semantics paradigm established by Fillmore (1976b). In Frame Semantics, a frame corresponds to a scenario that involves the interaction of a set of participants which play a particular role, being some of those roles essentials for the scenario. FrameNet groups 12,940 English words or *lexical-units* (LU) into 1,019 coherent semantic classes or frames. Each of these frames are further characterized by a list of participants or roles called

[8] `http://framenet.icsi.berkeley.edu/`

Commerce_pay:	
lexical-units:	**disburse.v, disbursement.n, pay.v, payer.n, payment.n, shell out.v**
frame-elements:	
Core:	*Buyer Goods Money Rate Seller*
Non-Core:	*Frequency Manner Means Place Purpose Time Unit*

Table 2.9: Lexical-units and frame-elements in the frame **Commerce_pay**.

[*Buyer* Theatres] **pay** [*Seller* movie producers] [*Goods* for showing their films].

Table 2.10: Example annotation for the verbal predicate **pay** in FrameNet.

frame-elements. Different word senses for a lexical-unit are represented in FrameNet by assigning different frames. These sets of words or lexical units will evoke the particular frames they belong to, which in turn are described with their respective frame-elements.

Table 2.9 shows the **Commerce_pay** frame. The *core frame-elements* are those essential frame-elements of the frame that can define the frames by themselves. In the **Commerce_pay** frame, the core frame-elements are the *Buyer*, *Goods*, *Money*, *Rate* and *Seller*, while the rest of the participants are considered less descriptive or too general. FrameNet contains a corpus of approximately 197,055 annotated frame instances as the example shown in Table 2.10.

However, not every *core frame-element* is always present in a sentence. For instance, in the example above it is not mentioned the *Money* given in the exchange between the *Buyer* (*Theatres*) and the *Seller* (*movie producers*) and neither the *Rate* or the price of the payment.

FrameNet defines a complex semantic network linking the frames and their frame-element with twelve different relationships of the type *subclass*, *causation* or *perspective*. The resulting network contains 10,076 direct relations between frames that allow to perform inferences involving the different events and their participants.

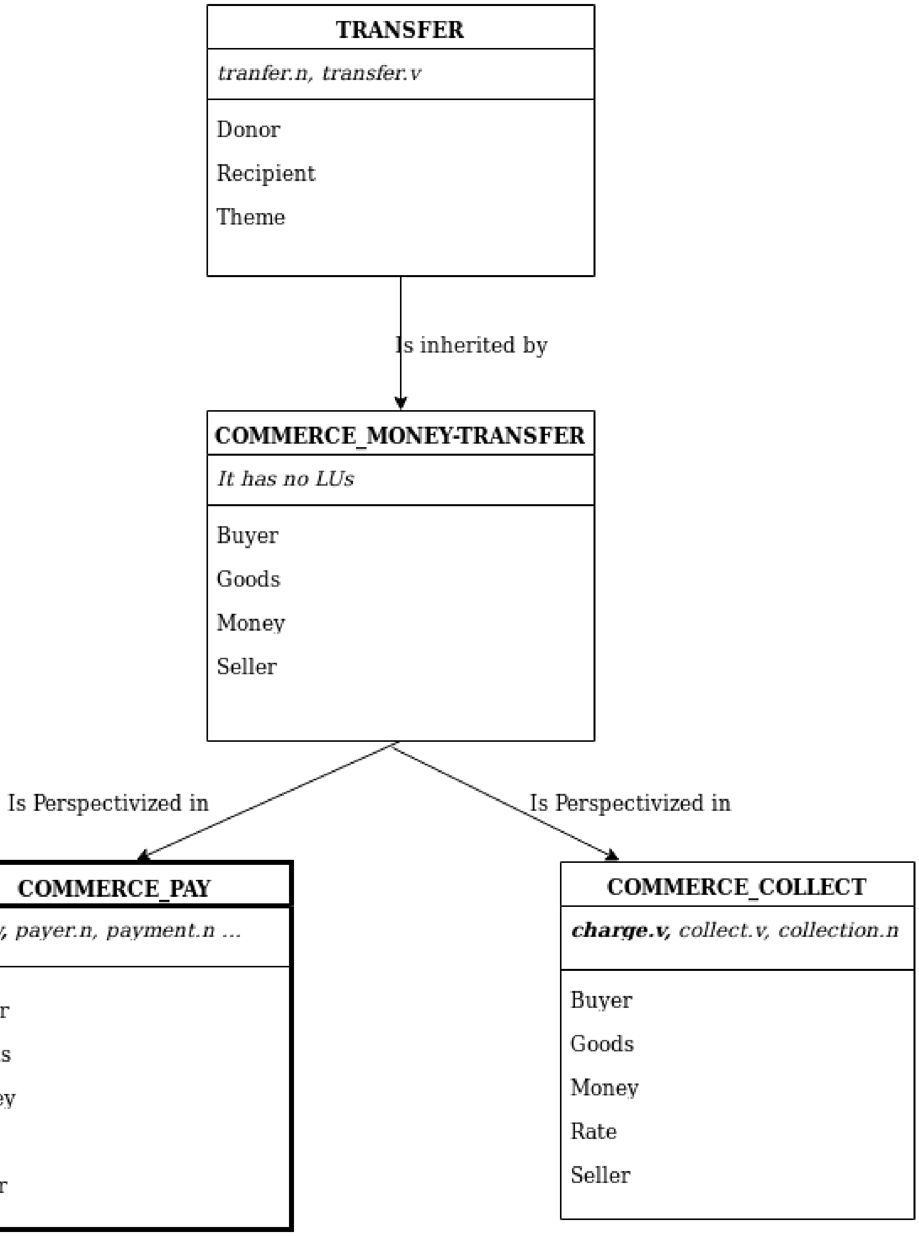


Figure 2.1: Examples of FrameNet frames connected by *Is Inherited by* and *Is Perspectivized in* relations.

Figure 2.1 shows a small portion of the whole ontology involving 4 different frames. The *Is Inherited by* relation between the frame **Transfer** and

[[*Buyer* Theatres] **pay** [*Seller* movie producers] [*Goods* for showing their films]] **Commerce_pay**.

[[*Seller* Movie producers] **charge** [*Buyer* theatres] [*Goods* for showing their films]] **Commerce_collect**.

Table 2.11: Example annotation for the frames **commerce_pay** and **commerce_collect** in FrameNet

Commerce_money-transfer means that the latter describes a more specific case of the first one. In consequence, all the properties of the frame **Transfer** are inherited by the frame **Commerce_money-transfer**, in particular, the corresponding frame-elements *Donor, Theme, Recipient.* However, the frame **Commerce_money-transfer** describes an scenario from a neutral point of view. That is why this frame does not contain any lexical-unit to evoke the frame. The event described by this frame varies its interpretation depending on the perspective of the participants. This is expressed by the relation *Is perspectivized in* that connects the frame **Commerce_money-transfer** to the frame **Commerce_pay**, if the perspective of the *Buyer* is assumed, and to the frame **Commerce_collect**, in the case of the perspective of the *Seller*. Thus, the lexical-units such as **pay.v**, **payer.v**, **payment.n**, **charge.v**, **collect.v** and **collection.n** are actually evoking the same event but from different points of view. Table 2.11 shows two sentences describing the same event, the difference is that the predicate **pay** in the first sentence sets the perspective on the *Buyer* and, consequently, triggers the frame **Commerce_pay**, whilst, in the second sentence, the frame **Commerce_collect** is evoked via the predicate **charge**, emphasizing the perspective of the *Seller*.

The original FrameNet was initially created for English, but, since then, many effort has been applied by different research groups to the creation of FrameNets for other languages. FrameNet frame hierarchy is considered mostly language-independent (Boas, 2005) and other proposals have followed thereafter aimed at the creation of FrameNets for other languages, in favor of applying the same theory, the same methodology and, sometimes, even the same annotation software. In this way, the initial project for English has evolved into a global, cooperative endeavor to cover other languages such as German (Burchardt et al., 2006; Boas, 2002), Japanese (Ohara et al., 2004), Spanish (Subirats and Sato, 2003) Dutch (Vossen et al., 2018) and many more. In that direction, Global FrameNet[9] is an effort to bring together all existing FrameNets in a common multilingual setting, focused on the development of collaborative research, shared tasks, and applications. Due to the growth of many projects creating FrameNets for other languages, the Multilingual FrameNet Project (Gilardi and Baker, 2018) was created as an attempt to find alignments between them all. Table 2.12 shows an overview of these international FrameNet initiatives.

Language	**name**	**URL**
English	FrameNet	`framenet.icsi.berkeley.edu`
German	German FrameNet	`www.laits.utexas.edu/gframenet/`
German	SALSA	`www.coli.uni-saarland.de/projects/ salsa/`
Spanish	Spanish FrameNet	`spanishfn.org/`
French	French FrameNet	`sites.google.com/site/anrasfalda/`
Brazilian Portuguese	FrameNet Brasil	`www.ufjf.br/framenetbr/`
Dutch	Dutch FrameNet	`dutchframenet.nl/`
Swedish	Swedish FrameNet	`spraakbanken.gu.se/eng/swefn`
Danish	Danish FrameNet	`framenet.dk/`
Japanese	Japanese FrameNet	`jfn.st.hc.keio.ac.jp/`
Chinese	Chinene FrameNet	`sccfn.sxu.edu.cn/portal-en/home.aspx/`
Korean	Korean FrameNet	`framenet.kaist.ac.kr/`

Table 2.12: Frame-semantic resources for different languages.

The website of the FrameNet project shows how this resource is highly used by hundreds of users worldwide in several areas of Natural Language Processing including sentiment analysis, building dialog systems, improving machine translation, teaching English as a second language, etc. FrameNet data has been mainly used for training and building different automatic Semantic Role Labeling (SRL) systems. Gildea and Jurafsky (2002) developed the first Semantic Role Labeling system based on FrameNet. After that, other SRL systems trained on the FrameNet data have been built (Erk and Pado, 2006; Johansson and Nugues, 2007a; Das et al., 2014). Along with the development of linguistic resources describing semantic role structures, new machine learning approaches were proposed that made use of the corpora annotated with such structures. The seminal work by Gildea and Jurafsky (2002)

[9] www.globalframenet.org

trained a maximum likelihood model to identify the semantic roles filled by constituents of a sentence following the semantic frames of FrameNet. They described a set features based on syntactic relations and proposed a two-step architecture, first to capture the frame-elements and second to label them. The task was subsequently disseminated by the Senseval-2004 shared task (Litkowski, 2004). Participants in this challenge, as well as later works, used as starting point the approach by Gildea and Jurafsky (2002). Later, Ruppenhofer et al. (2010) presented a task in SemEval-2010 on *Linking Events and Their Participants in Discourse* that, besides the traditional semantic role labelling, included an implicit role identification challenge based on FrameNet.

By means of Semantic Role Labeling tools FrameNet has been used in applications like Information extraction (Mohit and Narayanan, 2003), question answering (Shen and Lapata, 2007; Sinha, 2008) or Dialog Systems (Chen et al., 2013) .

2.2.6 SemLink

SemLink[10] (Palmer, 2009; Palmer et al., 2014) is a project working on the integration of the predicate information provided by different predicate resources via a rich set of manual mappings. Firstly, VerbNet semantic roles were assigned the corresponding numbered PropBank arguments by linking VerbNet semantic roles to a representative portion of the PropBank corpus (Loper et al., 2007). Currently, SemLink provides partial mappings between FrameNet (Baker et al., 1998b), VerbNet (Kipper, 2005), PropBank (Palmer et al., 2005), WordNet (Fellbaum, 1998b) and OntoNotes (Hovy et al., 2006; Pradhan et al., 2007) sense groupings, which provide coarse-grained sense representations based on manually created WordNet senses sets to guarantee a high level of agreement in the annotations.

PropBank, VerbNet and FrameNet offer diverse source of expert-curated knowledge; they vary in terms of the level and nature of semantic detail they describe. Palmer (2009) argue their complementarity and states that the redundancy caused by grouping the three resources together can be useful. Through their integration, a much richer and more complete resource is obtained, since PropBank provides the best coverage and the largest corpus

[10] `http://verbs.colorado.edu/semlink/`

PropBank	VerbNet	FrameNet
pay.01	pay-68-1	Commerce_pay
arg_0	*Agent*	-(*Buyer*)
arg_1	*Asset*	-(*Money*)
- (arg_2)	*Recipient*	-(*Seller*)
arg_3	*Theme*	-(*Goods*)

Table 2.13: SemLink role mappings for the verbal predicate **pay**.

that can be used as training data for supervised Machine Learning techniques, VerbNet contains the clearest links between syntax and semantics, and finally, the most fine-grained FrameNet provides the richest semantics and ontological knowledge. The example in Table 2.13 shows the full mapping between the different role structures of the same sense of the predicate **pay**.

SemLink aims at unifying these lexical resources at several different levels. First, it provides type-to-type mappings between the lexical units for each framework: coarse-grained rolesets in the case of PropBank, verbs that are members of VerbNet classes and lexical units associated with each semantic frame in FrameNet. The same lemma can have multiple PropBank rolesets and can be in several VerbNet classes and FrameNet frames, but always with different meanings. Second, SemLink also supplies a mapping between arguments and thematic-roles of PropBank and VerbNet respectively, as well as thematic-roles of VerbNet and frame-elements of FrameNet.

Manually mapping these resources is a very costly and difficult task due to their heterogeneous nature, different and changeable degree of coverage, and their different granularity. In the example of Table 2.13, the argument arg_2 of the predicate **pay.01** of PropBank is not connected to its corresponding thematic-role in VerbNet nor to its corresponding frame-element of FrameNet which is *Seller*. In addition, although the frame **Commerce_pay** is linked to the **pay.01** roleset and the VerbNet class ***pay-68-1***, none of the frame-elements of FrameNet are aligned to PropBank arguments and VerbNet thematic-roles. In Chapter 4 we will provide a complete description of the partial coverage of SemLink across the different predicate resources.

In Table 2.14, the example used in the previous sections is shown annotated according to the information provided by SemLink. In this case Sem-

	Theatres	**pay**	*movie producers*	*for showing their films*
VerbNet	AGENT		RECIPIENT	THEME
PropBank	ARG0 *buyer, payer*		-(ARG2) *the person being paid*	ARG3 *paid for what*
FrameNet	-(BUYER)		-(SELLER)	-(GOODS)

Table 2.14: Example of semantic annotations according to VerbNet, PropBank, and FrameNet in SemLink for the verb predicate **pay**.

Link does not contain information on the frame-elements of FrameNet and the mapping with PropBank is not complete either.

The Unified Verb Index (UVY)[11] provides a platform to to consult jointly the resources integrated in Semlink.

A more FrameNet-centered linked resource was developed by Palmer et al. (2014) called SemLink+ for its application in automatic event identification and extraction that encompasses their Event Ontology.

There are many works that exploit jointly the knowledge in the lexical resources integrated in SemLink stating that unifying information from different knowledge-bases is crucial for many complex language processing applications. For instance, Shi and Mihalcea (2005) integrate FrameNet, VerbNet, and WordNet, into a unified, richer knowledge-base, with the objective of enabling more robust semantic parsing. In the same direction, Giuglea and Moschitti (2006) interconnect FrameNet, VerbNet and PropBank for developing a robust semantic parser. Huang et al. (2016) combine Abstract Meaning Representation (AMR) (Banarescu et al., 2013), PropBank, FrameNet, VerbNet and OntoNotes knowledge for Event extraction.

[11] uvi.colorado.edu

2.2.7 AnCora

AnCora[12] (Taulé et al., 2008a) provides a multilevel corpus for Spanish and Catalan that includes annotations of lemmatization, syntactic constituents, WordNet senses, coreference, named entities and also semantic roles. Ancora also develops a semantic resource called *AnCora-Verb* (Juan Aparicio and Martí, 2008) that contains Spanish (*Ancora-Verb-Es*) and Catalan (*Ancora-Verb-Ca*) verbal predicates and their corresponding arguments structures (see Table 2.15). These two extensive verbal lexicons (*Ancora-Verb-Es* contains a total of 1965 different verbs corresponding to 3671 senses and *AnCora-Verb-Ca* contains 2151 verbs corresponding to 4513 senses), are the basis for the semantic role annotation of the *AnCora* corpora (*AnCora-Ca* and *AnCora-Es*). The lexicons and the annotated corpora constitute the richest linguistic resources of this kind freely available[13] for Spanish and Catalan.

In *AnCora-Verb* lexicons, the mapping between syntactic functions, arguments and thematic roles of each verbal predicate is established taking into account the verbal semantic class and the diatheses alternations in which the predicate can participate. Each verbal predicate is related to one or more semantic classes differentiated according to the four event classes - accomplishments, achievements, states and activities-, and on the diatheses alternations in which a verb can occur.

Table 2.16 shows an example sentence with the predicate **pagar** and its arguments annotated with *AnCora-Verb*.

AnCora-Verb is based on PropBank (see section 2.2.2) and both resources are linked by a wide set of mappings called *Ancora-Net*. *Ancora-Net* is a multilingual lexicon which combines syntactic-semantic and conceptual information from different sources: the *AnCora-Verb-Es* for Spanish, *AnCora-Verb-Ca* for Catalan lexicons, and information from the *Unified Verb Index* (UVI) for English predicates.

Additionally, for Spanish, *AnCora* includes the nominalizations of its verbal predicates in a resource called *AnCora-Nom*. For example, *pago.1.default*, the nominalization of the predicate *pagar.1.default*, is described in *AnCora-Nom* as shown in Table 2.17. Unless *Ancora-Nom* states otherwise, the correspondence between the arguments of the verbal and nominal predicates is direct.

[12] http://clic.ub.edu/corpus/es

[13] http://clic.ub.edu/corpus/

```
<lexentry lemma=”pagar” lng=”es” type=”verb”>
  <sense id=”1”>
    <frame lss=”A32.ditransitive-patient-benefactive”
    default=”yes” type=”default”>
      <argument argument=”arg0” function=”suj”
      thematicrole=”agt”/>
      <argument argument=”arg1” function=”cd”
      thematicrole=”pat”/>
      <argument argument=”arg2” function=”ci”
      thematicrole=”ben”>
      ...
```

Table 2.15: Example of the argument structure defined in *AnCora-Verb* for the predicate *“pagar.1.default”*.

[*agt* Los cines] **pagan** [*ben* a los productores de películas] [*pat* por mostrar sus pellículas].

Table 2.16: Example annotation for the verbal predicate **pagar** in *AnCora-Verb-Es*. The example sentence is the translation to Spanish of the sentence *“Theatres pay movie producers for showing their films”*.

```
<lexentry lemma=”pago” lng=”es” type=”verb”>
  <sense id=”1”>
    <frame lss=”A32.ditransitive-patient-benefactive”
    default=”yes” type=”default”>
      <argument argument=”arg0” function=”suj”
      thematicrole=”agt”/>
      <argument argument=”arg1” function=”cd”
      thematicrole=”pat”/>
      ...
```

Table 2.17: Example of the argument structure defined in AnCora-Nom for the nominal predicates *“pago.1.default”*.

Taulé et al. (2016) developed the first corpus annotated with implicit arguments for the Spanish language following the annotation scheme used for tagging the explicit arguments in Ancora. All these resources can be used by machine learning-based semantic role labeling systems as for other NLP applications.

2.2.8 Basque Verb Index

Estarrona et al. (2015) developed the **Basque Verb Index** (BVI)[14], a corpus-based lexicon for Basque. The lexicon is the result of semi-automatically annotating the EPEC-RolSem, a Basque corpus labeled at predicate level following the PropBank-VerbNet model. The predicates of the Basque Verb Index are mapped to PropBank and the roles are linked to VerbNet. For example, as shown in Table 2.18, the Basque predicate *ordaindu_1* is linked to the English predicate *pay.01* of PropBank and the arguments arg_0, arg_1, arg_2 and arg_3 of *ordaindu_1* are mapped to their corresponding VerbNet roles, *Agent*,*Asset*, *Beneficiary* and *Theme* respectively.

Table 2.19 shows an example sentence with the predicate **ordaindu** and its arguments annotated with *BVI*.

The *BVI* has been used to develop SRL systems for Basque like *bRol* (Salaberri et al., 2015), an automatic system for the parsing of syntactic and semantic dependencies in Basque.

[14] http://ixa2.si.ehu.eus/e-rolda/index.php

```
<aditz aditza=”ordaindu” zenb=””>
   <adiera zenb=”1”>
      <ordain zenb=”1” adiera=”pay.01” >
         <arg zenb=”0” rol=”agent” eadbrol=”” >
         <case grammcase=”erg”/>
         </arg>
         <arg zenb=”1” rol=”asset” eadbrol=”” >
         <case grammcase=”abs”/>
         </arg>
         <arg zenb=”2” rol=”beneficiary” eadbrol=”” >
         <case grammcase=”dat”/>
         </arg>
         <arg zenb=”3” rol=”theme” eadbrol=”” >
         <case grammcase=”abs/mot”/>
         </arg>
      </ordain>
   </adiera>
</aditz>
```

Table 2.18: Example of the argument structure defined in Basque Verb Index for the predicate *ordain_1* and its corresponding mappings to PropBank and VerbNet.

[*agent* Zinema aretoek] [*beneficiary* zinema ekoizleei] **ordaintzen** diete [*theme* beren filmak erakusteagatik].

Table 2.19: Example annotation for the verbal predicate **ordaindu** in *BVI*. The example sentence is the translation to Spanish of the sentence *“Theatres pay movie producers for showing their films”*.

2.3 Additional Resources

In this section, we provide a brief overview of some prominent additional resources that model event and predicate-argument structures but have not been integrated yet in the Predicate Matrix.

The **Event and Implied Situation Ontology (ESO)**[15] (Segers et al., 2015a) formalizes the pre and post situations of events and the roles of the entities affected by an event. It is manually built on top of existing resources such as WordNet, SUMO (Suggested Upper Merged Ontology) (Niles and Pease, 2001; Pease et al., 2002) and FrameNet. It is used to detect and abstract over dynamic and static events and their implications in text. Modeling of event implications allows for extracting sequences of states and changes over time regardless of this information being directly expressed in text, thus detecting implicit information. Events are interpreted as situations using RDF (Resource Description Framework) representation, taking all event components into account and expressing their relation with the event as triples. ESO includes mappings between ESO and FrameNet roles, and links from ESO classes to FrameNet frames and SUMO classes.

The ESO initial conceptual structure was derived from manually linking the SUMO ontology to those FrameNet frames that appeared most frequently in a large document collection. These links were the base to define the ESO event classes. Since the goal of the ontology is to model situation assertions, frames that refer to the same situational scenario were grouped together into the same class. Frame Elements perceived as key participants of the situations described by the classes were manually selected as ESO roles.

In order to make use of the interoperable capabilities offered by the Predicate Matrix, the classes and roles of ESO were also connected to the predicates and roles of the Predicate Matrix (Segers et al., 2016c, a) through FrameNet and SUMO labels.

The large-scale lexical-semantic resource **UBY**[16] (Gurevych et al., 2012) combines information from several widely used resources for English (WordNet, Wiktionary, Wikipedia, FrameNet and VerbNet) and German (Wikipedia, Wiktionary and GermaNet) and the multilingual OmegaWiki by linking them pairwise at the word sense level. This enables resource interoperability on the

[15] https://github.com/RoxaneSegers/ESO-Ontology

[16] http://www.ukp.tu-darmstadt.de/data/uby

sense level, e.g. by providing access to complementary information for a sense in different resources. All these resources have been modeled according to the Lexical Markup Framework (LMF) (Francopoulo et al., 2006), a standard model for modeling lexical resources. The alignments between the different resources are carried out by means of a flexible alignment framework based on the method of Niemann and Gurevych (2011).

In the project **Ontonotes** (Hovy et al., 2006; Weischedel et al., 2011) a large multigenre and multilingual (English, Chinese, and Arabic) corpus was annotated with syntactic dependencies, PropBank predicate-argument relations, nominal and verbal word senses, and coreference. The word senses in Ontonotes (also included in SemLink) where defined by grouping fine-grained WordNet senses into more coarse-grained semantic classes. In this way, the word sense granularity is tailored to achieve a high inter-annotator agreement (Palmer et al., 2004). In addition, some of the related Ontonotes senses are connected to concepts in the Omega ontology (Philpot et al., 2005).

The **SenSem** (Vàzquez et al., 2006; Alonso et al., 2007) project developed two resources for Spanish verbs. On the one hand, a corpus was manually annotated where sentences were analyzed at three different levels (the verb as a lexical item, the constituents of the sentence and the sentence as a whole) in order to associate them to their syntactic-semantic interpretation. On the other hand, a lexicon was built with verb senses linked to the corresponding annotated examples in the corpus. The resulting databank reflects the syntactic-semantic behavior of 250 frequent verbs of Spanish. The corpus was constituted with 100 occurrences of each verb, 25,000 sentences in total. The annotation process included verb sense disambiguation, syntactic structure analysis, interpretation of semantic roles and analysis of sentence semantics.

T-PAS[17] (Jezek et al., 2014) is a lexical resource of typed predicate-argument structures (T-PAS) for Italian, acquired by manually clustering distributional representations of verbs. T-PAS are corpus-derived verb-argument patterns where the arguments are specified by their semantic types. For example, a pattern for the verb *guida* (to drive) in T-PAS is [[Human]] *guida* [[Vehicle]]. The current T-PAS release includes typed patterns for 1,000 Italian verbs, and consists of three components: a repository of corpus-derived T-PAS linked to verbal lexical units, an inventory of about 200 corpus-derived semantic classes for nouns, relevant for the disambiguation of the verbs in

[17] http://tpas.fbk.eu/

context, and a corpus of sentences tagged with the lexical unit and the corresponding pattern number.

Im (2019) worked on linking the **Event Structure Frame** (ESF) (Im and Pustejovsky, 2009, 2010) to WordNet. Specifically, they developed a semi-automatic annotation of the ESF on the example sentences of each verb synset in WordNet. The semi-automatic annotation work is carried out using the Generator of the Event Structure Lexicon (GESL) tool by Im (2013) to automatically obtain ESF annotations that are manually corrected afterwards. The ESF models the event evoked by a verb defining its pre-state, process and post-state sub-events. This semantic description of the events is of great help for inferencing or reasoning tasks, and, by linking ESF to WordNet, the latter can be enriched with sub-eventual and argument structures.

Groningen Meaning Bank (GMB) (Basile et al., 2012; Bos et al., 2017) offers a large corpus for English semantically annotated with formal meaning representations following Discourse Representation Theory (Kamtl and Reyle, 1993). GMB integrates distinct phenomena (including predicate-argument structure, scope, tense, thematic roles, rhetorical relations and presuppositions) into an unified formalism, instead of covering single phenomena in a linguistically isolated way. The annotation units in the GMB are full texts, rather than isolated sentences, what allows to deal with ambiguities at the sentence level since they require the discourse context. The current version of the GMB contains more than 10,000 public domain texts aligned with Discourse Representation Structures, and is freely available for research purposes. It is semi-automatically created following a sophisticated bootstrapping approach. First, they generate automatic annotations making use of existing language technology tools as a starting point and then the final annotation is refined by both expert linguists (using wiki-like platform) and by non experts (using crowd-sourcing methods).

Word senses are expressed as WordNet 3.1 synset identifiers and for the annotation of the thematic roles they have mapped VerbNet roles to Combinatory Categorical Grammar (CCG) syntactic formalism encoded arguments.

The **Parallel Meaning Bank** (PMB) (Abzianidze et al., 2017) contains meaning banks for several languages (English, Dutch, German and Italian). Based on a cross-lingual projection approach, semantic information in PMB is projected from one sentence to its translated counterpart. The annotations for English are mapped through word-aligned translations to Dutch, German and Italian. PMB is based on the assumption that translations are meaning-preserving and, consequently, have equivalent meaning representation.

TRIPS [18] provides detailed lexical information integrated with an ontology [19]. It uses the set of semantic roles described in Allen and Teng (2018). The set of roles defined in TRIPS are built on different types of previous existing rolesets. For each semantic role, the ontology provides both selectional preferences and syntactic linking templates. TRIPS is connected to a publicly available domain-independent and broad-coverage semantic parser (Allen and Teng, 2017). As TRIPS contains mappings between its ontology types and WordNet synsets, the TRIPS parser can produce semantic representations of sentences even containing words not included in the TRIPS lexicon.

COLLIE (*Comprehensive OntoLogy and Lexicon In English*) extends the work carried out by the TRIPS project by extending its coverage and including logical axioms. Consequently, both resources are closely related. COLLIE it is an effort to build an extensive semantic lexicon for which they have been started from the verbal component **COLLIE-V** (Allen et al., 2020, 2022). The resource consists of two linked core components: an event ontology and a verbal lexicon. Thus, for each sense of a certain verb it contains ontological information about the concept expressed by that sense and lexical information about the behavior of the verb.

The lexical items of COLLIE-V are associated with linking templates (similar to the syntactic frames of VerbNet) that relate syntactic realizations to the semantic roles of the ontology concepts.

COLLIE-V is built semi-automatically, starting from the hand-built aforementioned TRIPS lexicon and ontology. First, the TRIPS parser is run on WordNet definitions obtaining *TRIPS*-based logical forms. In this way, sets of mappings between both resources are obtained that are manually cleaned up and extended. The process is run iteratively. The TRIPS ontology is used as the upper ontology and augmented with the WordNet hypernym hierarchy via the mappings obtained. Finally, the logical forms given by the TRIPS parser are also used to generate semantic types and axioms that are included in the COLLIE ontology.

[18] https://tripslab.github.io

[19] https://github.com/wdebeaum/trips-docs/wiki/Ontology

VerbAtlas[20] (Di Fabio et al., 2019b) is a verbal semantic resource where all verbal synsets from WordNet are clustered in 466 semantically-coherent frames. Verbal synsets expressing similar semantics are organized into a frame consisting of a prototypical argument structure which is applicable to all the synsets in a specific frame and concept-specific information. The frames in VerbAtlas are created via semantic similarity between synsets, that is, if two or more synsets share features like the purpose of the action and the participants in the action they are clustered together in the same frame. For example, the synsets <*pay* (Give money, usually in exchange for goods or services)>, <*accept, assume, bear* (Take on as one's own the expenses or debts of another person)> , <*finance* (Obtain or provide money for)> etc. are grouped into the *pay* frame because they express semantically-similar scenarios, i.e. they share similar participants (an Agent who pays/who assumes the expenses or debts/who finances; a Recipient who is paid/who gets rid of the expenses or debts/who is financed) and purpose (Agent gives or provide money to a Recipient). In order to achieve an argument structure capable of being applied to all the synsets in a frame, a common initial argument structure is created for the whole frame that is subsequently expanded for each synsets individually. The coverage of VerbAtlas is much higher than the most popular verbal resources such as PropBank, FrameNet and VerbNet since it covers all the verbs in WordNet. In addition to the resource coverage, VerbAtlas intends to overcome some other aspects such as the informativeness of the semantic roles. In contrast to the most underspecified PropBank enumerative arguments or, on the contrary, the most overspecified frame elements of FrameNet, the VerbAtlas role repository is inspired by VerbNet, but the number of roles was reduced in order to alleviate data sparsity issues. Moreover, the semantic roles are linked to selectional preferences expressed in terms of WordNet synsets. Nevertheless, VerbAtlas provides mappings to PropBank rolesets in order to make it suitable for NLP tasks that rely on PropBank. This mapping was done starting from the Predicate Matrix and then manually corrected and augmented.

SintagNet[21] (Maru et al., 2019) is a resource designed for its use in knowledge-based Word-sense Disambiguation systems. It consists of manually disambiguated lexical-semantic combinations. For example, the following verb-noun combinations are obtained for different senses of *run*:

[20] http://verbatlas.org

[21] http://syntagnet.org

- run (*carry out a process or program, as on a computer or a machine*) ↔ program (*a sequence of instructions that a computer can interpret and execute*).
- run (*compete in a race*) ↔ race (*a contest of speed*).
- run (*direct or control; projects, businesses, etc.*) ↔ farm (*workplace consisting of farm buildings*).

The verb-noun pairs are evoked from syntagmatic relations when the frequency of their co-occurrences is statistically significant. The lexical combinations are first extracted from the English Wikipedia and the British National Corpus, and then manually disambiguated according to WordNet 3.0. The resource contains 88,019 semantic combinations (61,249 noun-verb and 26,770 noun-noun semantic relations) and 20,629 links to WordNet 3.0 synsets (14,204 noun synsets and 6,422 verb synsets). The syntagnet verb-noun combinations could be used to "concretize" the roles of a predicate. For example, given the SyntagNet pair $\text{eat}_v^1 \leftrightarrow \text{bread}_n^1$, the VerbAtlas relation $\text{eat}_v^1 - \mathit{Theme} \rightarrow \text{food}_n^1$ could be "concretized" to $\text{eat}_v^1 - \mathit{Theme} \rightarrow \text{bread}_n^1$. SyntagNet's new lexical-semantic relations have been proven to be effective for knowledge-based WSD with the UKB disambiguation algorithm.

PASBio (Wattarujeekrit et al., 2004) is an extension of the predicate-argument structure of PropBank to the technical domain of molecular biology. Wattarujeekrit et al. (2004) semantically analysed the argument structure of 30 verbal predicates known to be used to describe biological events and proposed the construction of a domain-specific predicate-argument structure. They analyzed and annotated sentences from MEDLINE abstracts and full-text journal articles for building PASBio.

2.4 Automatic approaches for mapping lexical semantic resource

The various resources described in the previous sections contain complementary information. Linking them together make it possible to create new enhanced resources and to combine the different information types provided by the different lexical resources integrated into them (e.g Wordnet's synonymy and Framenet's frame relations) for tasks such as inferencing, consistency checking, interoperable semantic role labeling, word sense disambiguation, etc. Integrating several resources into a new, more complete resource has the advantage that all integrated information is easily accessible.

Furthermore, the coverage of resources are generally far from being complete, as building large and rich enough lexical resources takes a great deal of expensive manual effort. Therefore, the alignment of complementary lexical semantic resources have been used for improving the coverage and quality of specific resources. The integration of resources can be applied to overcome the lack of information of a specific resource by means of another resource.

The mapping of resources can be done manually, semi-automatically or fully automatically. For instance, EuroWordNet Interlingual Index (ILI, (Vossen, 1998b)) was one of the first linked resource created manually by aligning at sense-level eight different Wordnet-like resources. Although such manual mappings may be the most reliable and achieve the highest accuracy, they are the most expensive and time-consuming to obtain as they require the work of experts during long time. For instance, due to the distinct granularities on the resources, sense representations can vary considerably from one resource to another. It may occur that not all senses are represented in all resources or that a particular sense is represented by more than one sense in another resource, etc. These issues complicate the task of building manual alignments as the decisions to be made may not be trivial.

The effort required to create manual mappings can be considerably alleviated with automatic approaches. These automatic strategies exploit the information encoded in the resources to be integrated, such as textual information (glosses, definitions of the senses, examples of use of the senses etc.) or the relational structure of the resources (e.g. the semantic relations of Wordnet or FrameNet's frame hierarchy) and also obtain high precision. There are also semi-automatic approaches which consist in manually validating the mappings created automatically. This approach was followed by Henrich et al. (2014) for mapping German Wiktionary to GermaNet.

The alignment of resources can be done at different information type level. Specifically, we refer to two types of alignments in this document, on the one hand the alignments at word sense level or predicate level (e.g alignment of FrameNet frame lexical units and VerbNet class members) and on the other hand the alignments at the predicate argument or role level (e.g alignments of FrameNet frame-elements to VerbNet thematic roles). In both types of alignments, the granularity difference problem mentioned above complicates the task of integrating resources for either a manual or automatic approach. Nevertheless, another advantage of the automatic approaches is their systematisation that allows to easily recreate the mappings when, for example, an integrated resource is updated with new senses.

The development and the free availability of large collaborative resources such as Wikipedia, FreeBase (Bollacker et al., 2008) or DBpedia (Bizer et al., 2009) among others, and their possible application in NLP motivated the works on resource integration by developing advanced methods to link automatically different lexical resources at sense level. One of these pioneering works by Ruiz-Casado et al. (2005) started aligning WordNet synsets and Wikipedia entries. Similarly, most previous research efforts on the automatic integration of lexical-semantic resources targeted at knowledge about nouns and named entities at word sense level rather than predicate knowledge. Well known examples are YAGO (Suchanek et al., 2007), BabelNet (Navigli and Ponzetto, 2010) and UBY (Gurevych et al., 2012). YAGO is an ontology derived from Wikipedia which was unified with WordNet by means of rule-based and heuristic methods. Later, Hoffart et al. (2013) extended it with GeoNames[22] database by integrating spatial and temporal information for entities, facts, and events. Navigli and Ponzetto (2010) created the multilingual knowledge base BabelNet by mapping automatically WordNet synsets and Wikipedia entries. In order to link each Wikipedia page to a WordNet sense they first create textual representations for Wikipedia pages and WordNet senses exploiting different information provided by each resource. For WordNet they made use of the gloss, synonyms and hypernym information in order to create the disambiguation contexts and for Wikipedia pages the lemmas of the outgoing links jointly with the classification categories of its pages. UBY is a linking effort of nine resources in two languages, German and English. It offers pairwise sense alignments of a subset of expert-constructed resources such as the English WordNet, GermaNet, FrameNet and VerbNet and collaboratively constructed resources such as Wiktionary, Wikipedia and OmegaWiki which are previously mapped to a uniform representation. They define a standardized format for modeling lexical semantic resources, which is extensible to new languages, for helping to make the integration of resources smoother. They expand the approach for sense disambiguation used previously in (Niemann and Gurevych, 2011) where they combine a threshold-based Personalized PageRank proposed by

[22] http://www.geonames.org/

Agirre and Soroa (2009a) on the WordNet graph with a word overlap measure, for extracting a set of Wikipedia articles candidates for linking them to WordNet sysnsets. Similarly, an adaptation of the same approach is used to align WordNet and Wiktionary by Meyer and Gurevych (2011). Later, Miller and Gurevych (2014) described a method for automatically constructing n-way alignments from an arbitrary set of pairwise lexical semantic resources alignments. They apply their approach on existing WordNet-Wikipedia and WordNet-Wiktionary alignments to produce a three-way alignment of those resources. Henrich et al. (2014) calculate the overlap between Wiktionary to GermaNet glosses to map both resources together.

Some pioneering works on the integration of verbal resources were applied to improve performance on semantic parsing and SRL tasks. For instance, Shi and Mihalcea (2005) create semi-automatically mappings between FrameNet, VerbNet, and WordNet to reduce coverage problems and improve a rule-based semantic role labeling system. Nevertheless, many of predicate resources linking works are focused on aligning at sense level two specific lexical resources such as WordNet and FrameNet. The rich semantic information contained in FrameNet makes it a very interesting resource for application in various NLP tasks. But the lack of coverage of FrameNet has been mentioned several times (Padó and Lapata, 2007; Burchardt et al., 2009) as a reason for failing in the attempts to integrate it in different NLP tasks. As a consequence many authors have investigate the exploitation of WordNet to extend FrameNet, avoiding the high costs of manual annotation. Moreover, Burchardt et al. (2005) use WordNet-based word sense disambiguation to create linkings between WordNet and FrameNet to tackle the lack of senses in FrameNet and improve the frame assignment process in the task of automatic text annotation. Firstly, lexical units in unseen texts are annotated with their contextually determined WordNet synset and then synonyms and hypernyms relations are used to propose a set of frame candidates. Finally, the best frame is selected via a weighting scheme. With a similar purpose Johansson and Nugues (2007b) approach the task of assigning new unknown lexical units to existing frames as a machine learning problem. Using features derived from the WordNet hierarchy, a Support Vector Machine is trained on existing lexical units of FrameNet and applied to assign unknown lexical units to the correct frame. Similarly, (Pennacchiotti et al., 2008) is one of the pioneers in the task called Lexical Unit Induction that consists of automatically acquiring new lexical-units. They propose two unsupervised models one based on distributional techniques and one using WordNet as a support. The approach by Pennacchiotti et al. (2008) differs from previous work in that it leverages distributional properties to induce lexical units, instead of relying on pre-existing lexical resources as WordNet. They model existing frames and unknown lexical units as distributional co-occurrence vectors in the same semantic space. Likewise, De Cao et al. (2008) combine corpus-based distributional information and word sense information derived from WordNet for automatically expanding the English FrameNet, reducing lexical units polysemy by mapping them to WordNet synsets and also creating a new FrameNet for Italian.

Tonelli and Pianta (2009) create MapNet by exploiting the similarity between FrameNet definitions of lexical units and WordNet glosses. Given a lexical unit (a FrameNet predicate), they first look for the synsets containing it and then select the one with the highest similarity between its gloss and the FrameNet definition for that lexical unit. They try two similarity algorithms based respectively on stem overlap and on a modified version of the Levenshtein distance algorithm taking stems as comparison unit instead of characters. Their goal is twofold: to extend the coverage of the FrameNet lexicon with WordNet synonyms, and to obtain an Italian-FrameNet through English-Italian MultiWordNet.

Other works exploit information extracted from the relational structure of resources, sometimes also combined with some type of textual information like those mentioned above. Ferrández et al. (2010) model senses based on semantic relations of WordNet and frame relations of FrameNet and by comparing both relational contexts they create links between FrameNet lexical units and WordNet synsets. This structural information is used in addition to textual information of senses. Laparra and Rigau (2009a); Laparra et al. (2010a) exploit a graph-based Word Sense Disambiguation algorithm called SSI-Dijkstra (Cuadros Oller and Rigau Claramunt, 2008) to partially integrate FrameNet and WordNet in a new resource called eXtendedWord-FrameNet. The knowledge-based Word Sense Disambiguation algorithm is used for assigning the appropriate synset of WordNet to the semantically related lexical units of a given frame from FrameNet. This algorithm relies on the use of a large knowledge base derived from WordNet and eXtended WordNet (Mihalcea and Moldovan, 2001). Also, they achieve a multilingual extension of FrameNet by using the Multilingual Central Repository links to the Spanish, Italian, Basque and Catalan WordNets.

Matuschek and Gurevych (2013) introduce another graph-based algorithm for word sense alignment called Dijkstra-WSA and apply it, in combination with the gloss-based approach used in (Gurevych et al., 2012), for aligning Wikipedia and OmegaWiki. This strategy starts linking the resources through monosemous words and then calculating shortest paths between senses to decide which should be aligned, thus, it requires the resources to contain some degree of relational structure.

Pilehvar and Navigli (2014) propose a general approach that can be applied to any pair of lexical resources based on a combination of gloss and graph similarity. For those resources that do not follow a network structure, they design an algorithm to translate such resources into WordNet-like ontologies. Their methodology, that can be run in both supervised and unsupervised settings, achieved state-of-the-art results in the alignment of WordNet with Wikipedia, Wiktionary and OmegaWiki.

More recently, DCL-IBL (2019); Leseva and Stoyanova (2019); Leseva et al. (2020) work on the alignment of the verb inventory in WordNet and FrameNet. FrameNet semantic frames and frame elements are mapped with a set of WordNet verb and noun synsets respectively. The mapping is expanded to as many synsets as possible by exploiting the inheritance relation from a hypernym to a hyponym of WordNet. Frames and frame elements are associated with a particular semantic class that expresses the semantic properties of both. This is how the network of conceptual frames are created which enrich the WordNet structure with generalised verb predicate-argument semantic relations.

In addition to WordNet, there have also been attempts to integrate collaborative resources like Wikipedia and verbal resources like FrameNet. For example, Tonelli and Giuliano (2009) and Tonelli et al. (2013) aim to automatically enrich FrameNet by exploiting Wikipedia knowledge. They employ the WSD system of Gliozzo et al. (2005) to find the page of Wikipedia that best expresses the meaning of a particular lexical unit belonging to a specific frame in FrameNet. Then, the mapping is employed to extract new example sentences and acquire new lexical units, both for English and for all languages available in Wikipedia by exploiting the inter-lingual links. However, links to Wikipedia are usually restricted to nouns what rules out most of the predicates of resources like FrameNet. On the contrary, FrameNet role fillers are commonly nouns, so in (Tonelli et al., 2012) they focus on mapping them to Wikipedia and consequently model richer selectional preferences. In (Alonso Alemany et al., 2009) they connect semi-automatically FrameNet with Sensem (Alonso et al., 2007), another predicate model resource for Spanish with the aim of transfering semantic information from the former to the latter. The automatic mapping uses a Word Sense Disambiguation algorithm based on Structural Semantic Interconections (SSI) (Navigli and Velardi, 2005) to first connect FrameNet lexical units with WordNet synsets, and in a second step, FrameNet and Sensem are connected through the synsets associated with each sense of SenSem. They generate positive and negative frame-SenSem pair examples to train automatic classifiers to pre-validate mappings that have not yet been manually validated. Mousselly-Sergieh and Gurevych (2016) enrich Wikidata items with linguistic information from FrameNet by aligning both resources. The aligning method is based on labels and aliases of Wikidata. Hartmann (2017) propose a powerful approach to construct a FrameNet lexicon in other languages using Wiktionary as an interlingual representation. They also discuss and address the need of unification of different linguistic information types, and the different terminology used to represent these types in the various resources.

Finally, role-level alignments between different predicate argument models have been much less explored in the literature. Indeed, we have not found any references to automatic mappings for predicates and their semantic roles.

LEXICAL SEMANTICS - AUTOMATIC PREDICATE INFORMATION INTEGRATION

CHAPTER 3

A Framework for Predicate Information Integration

This chapter summarizes the complete framework of the research developed in this thesis. Section 3.1 introduces a general overview of the methodology proposed to construct the new lexical-semantic resource. Then, the details of these techniques for creating automatic mappings between lexical entries and roles of different predicate resources will be further explained in the following chapters. Section 3.2 introduces the new lexical-semantic resource with which this work contributes: the Predicate Matrix. Finally, in Section 3.3 we make a schematic summary of the characteristics of the different versions of our resource and the contributions of each chapter related to the construction of the Predicate Matrix.

3.1 Steps towards a new lexical resource

As it is mentioned in Section 2, one of the few projects working on the integration of different sources of predicate semantic information is SemLink. For each predicate, SemLink supplies a mapping between the semantic roles of VerbNet and PropBank, as well as the semantic roles of VerbNet and FrameNet. Moreover, SemLink provides mappings to WordNet senses for

VerbNet predicates. Figure 3.1 shows the resources that SemLink aims to connect at the predicate and role levels.

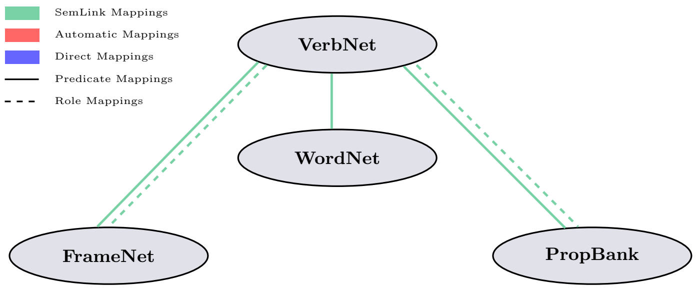


Figure 3.1: SemLink graph representation.

However, SemLink has some limitations. First, its coverage is still far from being complete. For instance, in some cases, a predicate from PropBank or FrameNet does not exist in VerbNet or some of its arguments may not have a corresponding role in VerbNet. In Chapter 4, we study and analyze these and other coverage issues in SemLink. A second limitation is that the mappings between the different resources have been manually developed, a very costly process which is also not systematic. Therefore, our proposal is to define automatic methods for mapping different predicate-argument models at predicate sense and role level. In this way, the resulting new resource will allow a more robust semantic interoperability between them. Furthermore, the automatic methodology makes easier to maintain updated the set of mappings when improved versions of the predicate knowledge resources (each one developed independently) are released.

In chapter 5, we describe and evaluate our proposed automatic methods to increase, complete and improve the semantic interoperability between VerbNet, PropBank, FrameNet and WordNet. The set of automatic methods work on the integration of predicate information both at the role and lexical level.

- **Automatic methods for the integration at lexical level**

  At the lexical level we use WordNet as a central resource in order to offer a wider coverage. Accordingly, at lexical level we work with three pairs of resources: WordNet-VerbNet, WordNet-FrameNet and

WordNet-PropBank, as shown in Figure 3.2. The methods for extending the mappings between lexical entries are based on a graph-based WSD approach which uses WordNet as a background knowledge base. Following (Laparra and Rigau, 2009b; Laparra et al., 2010b), we apply knowledge-based WSD algorithms that use a large-scale graph of concepts derived from WordNet to disambiguate the entries from the lexicons.

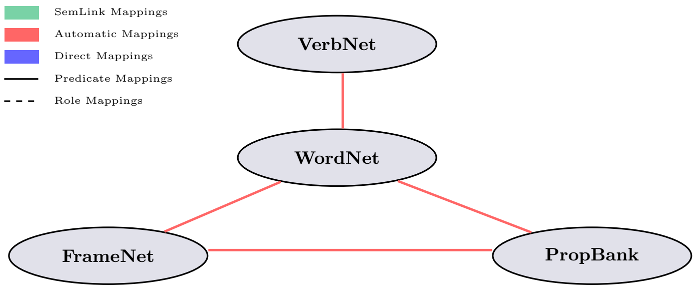


Figure 3.2: Predicate Matrix mappings at lexical level.

On the one hand, the lexical mappings from WordNet to FrameNet and from WordNet to VerbNet are obtained by applying graph-based WSD algorithms to semantically coherent groupings of verbal entries belonging to the same FrameNet frame or VerbNet class. The details of this method can be read in Section 5.2.1.1. On the other hand, for the lexical mappings from WordNet to PropBank, we cross the annotations obtained by a WordNet-based WSD algorithm on a corpus manually annotated with PropBank predicates. This method it is explained in detail in Section 5.2.1.2

In all cases the WSD strategies provide new links between predicates and WordNet senses. Consequently, we can connect verbs from different resources that are connected to the same WordNet sense.

In addition, the strategy followed to generate mappings between FrameNet and PropBank roles simultaneously generates lexical mappings between these two resources as well. This will be explained in more detail in Section 5.2.2.2.

Even though we do not propose any method to align PropBank and VerbNet directly, or VerbNet and FrameNet at lexical level, we obtain some new mappings between all those pairs of resources indirectly. For example, predicates from PropBank and VerbNet that are not linked obtain mappings to the same lexical unit of FrameNet. Table 5.4 in Chapter 5 shows the differences between SemLink and the Predicate Matrix in terms of mappings between lexicons.

- **Automatic methods for the integration at semantic role level**

  At role level, we work on the integration of the following two pairs of resources: VerbNet-FrameNet and FrameNet-PropBank. On the one hand, we propose an automatic method to increase the alignments between VerbNet thematic-roles and FrameNet frame elements and on the other hand, we focus on extending the mappings between FrameNet frame elements and PropBank arguments. This is represented in Figure 3.3. The details of the methodology applied to these two types of mappings are explained in Section 5.2.2.1 and Section 5.2.2.2, respectively.

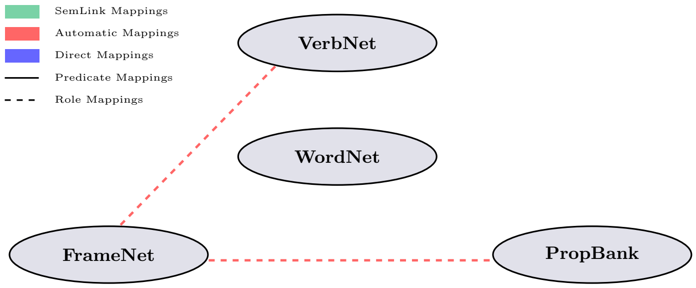


Figure 3.3: Predicate Matrix mappings at role level.

In short, the method to infer new semantic role mappings between VerbNet and FrameNet learns role patterns and frequencies from existing mappings. The method comprises three different steps that should be applied consecutively but we have set two alternative configurations. The first configuration of this three-step method requires the information contained in SemLink in order to first learn which alignments between VerbNet thematic-roles and FrameNet frame-elements are more frequent. However, the second configuration of this method is completely independent from SemLink. This configuration starts from the second step, where the thematic-roles and the frame-elements are aligned based on role pattern frequencies from the examples of use contained in VerbNet and the lexicographic annotations of FrameNet. In this way, when a verb from VerbNet is mapped to a frame of FrameNet, the most frequent thematic-role pattern for the class of the verb is aligned to the most frequent frame-element pattern for the frame (see Section 5.2.2.1).

In Section 5.2.2.2 we present the second role mapping method. In this case, we focus on PropBank and FrameNet, the two resources with the poorest role mapping coverage in SemLink. To obtain the role mappings between PropBank and FrameNet, our method acquires the most common correspondences between the annotations of both resources over the same sentences (see for Figure 3.4). The idea is to obtain first a corpus with gold FrameNet annotations and automatic PropBank annotations and a corpus with gold PropBank annotations and automatic FrameNet annotations. Then, we cross the annotations on both corpora to collect the coincidences.

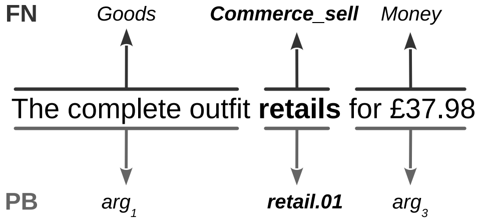


Figure 3.4: Example of matching annotations of FrameNet and PropBank at role level.

Figure 3.5 summarizes the types of direct alignments integrated in the Predicate Matrix thanks to these methods. As already have been mentioned, the integration of predicate information is focused on two levels: lexical and role level. This is represented by discontinuous and continuous lines respectively in Figure 3.5. For instance, WordNet does not contain information about roles, so its integration with other resources is at predicate level. However, all other resources are aligned at the two levels, predicates and roles. Figure 3.5 represents the mapping coming from SemLink or Predicate Matrix with different colors. For instance, as it is explained in the study of Semlink coverage in chapter 4, the mapping between PropBank and VerbNet is almost complete in Semlink so we do not propose any method to extend or complete the mappings between these two resources.

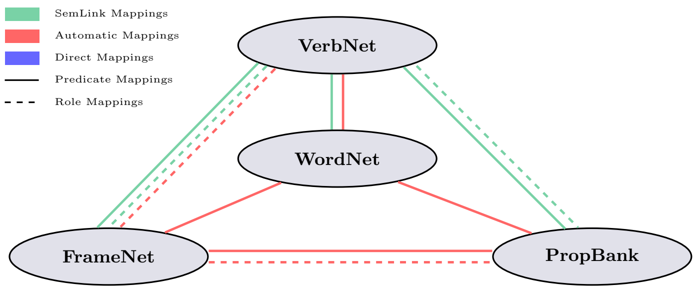


Figure 3.5: Predicate Matrix lexical and role mappings graph representation.

- **Extending the Predicate Matrix to cover nominalizations and multilingual predicates**

  In Chapter 6 we deal in a simple way with the problem of multilingualism and the nominalization of the Predicate Matrix. Firstly, we have extended the predicate information to languages other than English, turning it into a multilingual resource. Specifically, we have integrated resources in Spanish, Catalan, and Basque. In Figure 3.6 it is shown that the extension to Spanish and Catalan has been made integrating AnCora (Taulé et al., 2008b) corpus and the AnCoraVerb (Juan Aparicio and Martí, 2008). The Basque Verb Index (BVI) (Estarrona et al., 2015) corpus-based lexicon is used in the case of Basque. Note that the case of Basque is special. Unlike the others, where both predicates and roles are mapped between the same resources, for Basque, the predicates of the Basque Verb Index are mapped to PropBank and the roles are linked to VerbNet. As a result, the Predicate Matrix provides a multilingual lexicon to allow interoperable semantic analysis in multiple languages.

  Secondly, as Chapter 6 explains, the Predicate Matrix has been also extended to cover nominal predicates by adding mappings to NomBank (Meyers et al., 2004), which contains nominalizations of the PropBank predicates, and Spanish AncoraNom.

  The projection of the Predicate Matrix to a new language or extending it to nominal predicates follow the same strategy: if any other resource not included in the Predicate Matrix yet is linked to any of the resources included in it the projection to that language or new resource can be done straightforwardly. In Chapter 6 we demonstrate this feature.

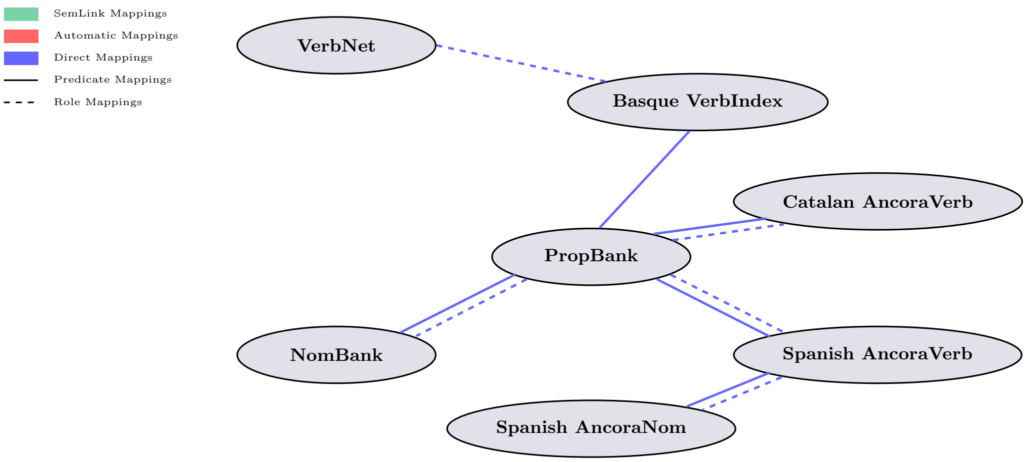


Figure 3.6: Nominal and multilingual Predicate Matrix graph representation.

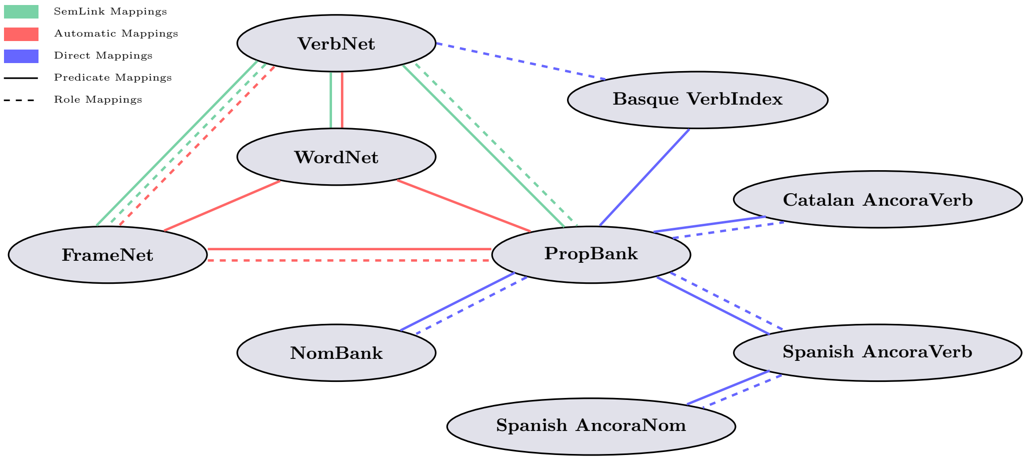


Figure 3.7: Predicate Matrix graph representation.

In summary, thanks to the new mappings obtained by our methodology, the predicate Matrix offers a large extension of the interoperable information contained by SemLink at both lexical and role level and in a cross-lingual manner.

For example, consider the verb **sell** that belongs to the VerbNet class ***give-13.1-1***. Given the sentence *"Tom sold Mary his car"*, an automatic semantic parser based on PropBank should annotate the sentence with the **sell.01** PropBank predicate and the roles arg_0, arg_1 and arg_2 , as shown in

Figure 3.8.

By means of SemLink, we know that the **sell.01** PropBank predicate belongs to the VerbNet class ***give-13.1-1*** and the arg_0, arg_1 and arg_2 PropBank arguments are aligned to the VerbNet *Agent*, *Theme* and *Recipient* thematic-roles respectively. The Predicate Matrix also offers information from FrameNet for this particular predicate. Thanks to the methodology followed in this work, the Predicate Matrix contains alignments for VerbNet and PropBank to the **Commerce_sell** frame and the **sell.v** lexical unit from FrameNet. *Agent* and arg_0 are equivalent to *Seller* frame element in FrameNet, *Theme* and arg_1 to *Goods*, and *Recipient* and arg_2 correspond to *Recipient* frame element. The Predicate Matrix also offers the mapping to the corresponding WordNet verb sense sell_v^1. Moreover, it allows to project the predicate information to Spanish, Catalan and Basque. For instance, the Spanish predicate **vender.1.default** and the Catalan predicate **vendre.1.default** shown in Figure 3.8 are mapped to the English PropBank predicate **sell.01**. The correspondence between the arguments in AncoraVerb and PropBank is direct. In the case of Basque, the predicate **saldu.1** is mapped to the English PropBank predicate **sell.01**. Instead, the arguments are mapped to VerbNet thematic-roles. In this way, argument 0 of the Basque verb **saldu.1** is equivalent to *Agent* in VerbNet, argument 1 to *Theme* and argument 2 to *Recipient*. Finally, the Predicate Matrix also includes nominalizations for English from NomBank (**sale.01**) and Spanish AncoraNom (**venta.1.default**).

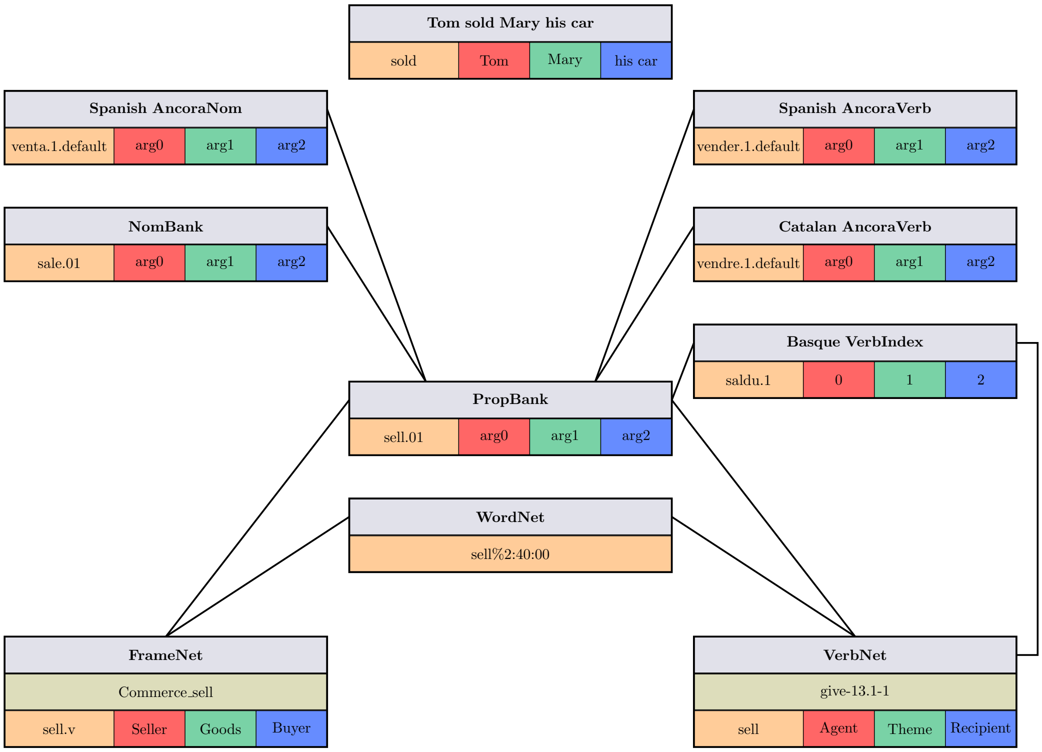


Figure 3.8: Example in Predicate Matrix graph representation.

3.2 Predicate Matrix format description

The Predicate Matrix for English is distributed as a tabulated file where each row represents the mapping of a role over the different resources and includes all the aligned knowledge about its corresponding verb sense. The file is structured in 21 columns, 13 of which concern the predicate models integrated as described in Table 3.1 and the remaining contains the additional semantic knowledge acquired from the MCR.

FIELD NAME	DESCRIPTION
1_VN_CLASS	Information of the VerbNet class
2_VN_CLASS_NUMBER	Information of the VerbNet class number
3_VN_SUBCLASS	Information of VerbNet subclass
4_VN_SUBCLASS_NUMBER	Information of the VerbNet subclass number
5_VN_LEMA	Information of the verb lemma
6_VN_ROLE	Information of the VerbNet thematic-role
7_WN_SENSE	Information of the word sense in WordNet
8_MCR_iliOffset	Information of the ILI number in the MCR3.0
9_FN_FRAME	Information of the frame in FrameNet
10_FN_LE	information of the corresponding lexical-entry in FrameNet
11_FN_FRAME_ELEMENT	Information of the frame-element in FrameNet
12_PB_ROLESET	Information of the predicate in PropBank
13_PB_ARG	Information of the predicate argument in PropBank

Table 3.1: Predicate Matrix main fields description.

The format of the multilingual Predicate Matrix is slightly different. An identifier is defined to distinguish between rows for English, Basque, Spanish and Catalan predicates, and between their verbal and nominal forms. This identifier is based on PropBank, AnCora, and the Basque Verb Index predicates and arguments, and is composed of 4 fields: language, form, predicate and argument. For example, following the example in Figure 3.8, the row in the Predicate Matrix corresponding to the argument "1" of the English nominal predicate "sale.01" is identified by "id:eng id:n id:sale.01 id:1". Similarly, the corresponding row for argument "arg0" of the Spanish verbal predicate "vender.1.default" is identified by "id:spa id:v id:vender.1.default id:arg0", as shown in Table 3.2.

id:eng id:n id:sale.01 id:1 vn:give-13.1 vn:Theme wn:ili-30-02244956-v fn:Commerce_sell fn:Goods pb:sell.01 pb:1
id:spa id:v id:vender.1.default id:arg0 vn:give-13.1 vn:Agent wn:ili-30-02244956-v fn:Commerce_sell fn:Seller pb:sell.01 pb:0
id:spa id:n id:venta.1.default id:arg2 vn:give-13.1 vn:Recipient wn:ili-30-02244956-v fn:Commerce_sell fn:Buyer pb:sell.01 pb:2
id:cat id:v id:vendre.1.default id:arg1 vn:give-13.1 vn:Theme wn:ili-30-02244956-v fn:Commerce_sell fn:Goods pb:sell.01 pb:1
id:eus id:v id:saldu.1 id:1 vn:give-13.1 vn:Theme wn:ili-30-02242464-v fn:Commerce_sell fn:Goods pb:sell.01 pb:1

Table 3.2: Some examples of mappings in the Multilingual and Nominal Predicate Matrix.

The Catalan and Basque rows are indexed by "id:cat" and "id:eus" respectively. Establishing such identifiers allows us to maintain the whole Predicate Matrix for all the languages in the same file.

3.3 Predicate Matrix evolution

The research carried out in this dissertation has resulted in different versions of the Predicate Matrix. However, the order of these versions does not correspond exactly to the order of the chapters in this document. While the versions of the Predicate Matrix arise from the resources mapped at each point in time, the structure of this thesis is based on the methodologies applied. For the sake of clarity, a road map is presented below to serve as a reference for knowing the content of each version of the Predicate Matrix and in which section of this dissertation it is described:

Predicate Matrix 1.0 Published in López de Lacalle et al. (2014a). It contains the mappings of SemLink in the format described in Section 3.3. This version also includes some tentative additional WordNet-VerbNet mappings for monosemous WordNet verbs (see Section 7.2.2) and synonyms of predicates already aligned to VerbNet (see Section 7.2.3).

Predicate Matrix 1.1 Published in López de Lacalle et al. (2014b). In this version, the tentative new mappings added in the previous version are replaced with automatic WordNet-VerbNet and WordNet-FrameNet lexical mappings (see Section 5.2.1.1) and VerbNet-FrameNet role mappings (see Section 5.2.2.1).

Predicate Matrix 1.2 Published in López de Lacalle et al. (2016b). It extends the previous version by adding automatic WordNet-PropBank lexical mappings (see Section 5.2.1.2) and PropBank-FrameNet lexical and role mappings (see Section 5.2.2.2).

Predicate Matrix 1.3 Published in López de Lacalle et al. (2016a). It incorporates mappings to multilingual resources (see Section 6.2). It expands the coverage to predicate nominalizations (see Section 6.3).

CHAPTER 4

A study of SemLink coverage

In this chapter we present a study of SemLink coverage. After a motivation of this work in Section 4.1, we present a detailed study of the coverage of the mappings between each resource included in SemLink. Semlink uses Verbnet as a central resource, so for the analysis of its coverage we analyze the mappings between the following resource pairs: first, in subsection 4.2.1 we analyze the alignments between WordNet and VerbNet, next, in subsection 4.2.2 the coverage between PropBank and VerbNet is examined and finally, the coverage between FrameNet and VerbNet in subsection 4.2.3. We describe the coverage and gaps of these mappings with respect to the lexical entries and the role structures of each resource. The chapter finishes with some concluding remarks in Section 4.3.

4.1 Introduction

The study we present in this chapter aims to motivate the need of the automatic mapping methods described in following chapters. As mentioned previously, SemLink is the main source available for links between predicative resources, but it has significant gaps in its coverage. In the following analysis, we detail numerically the magnitude of the missing information. Here, as in the rest of this dissertation, we work on the version 1.2.2 of Semlink that includes WordNet 3.0, FrameNet 1.3, VerbNet 3.2 and PropBank 2.1.

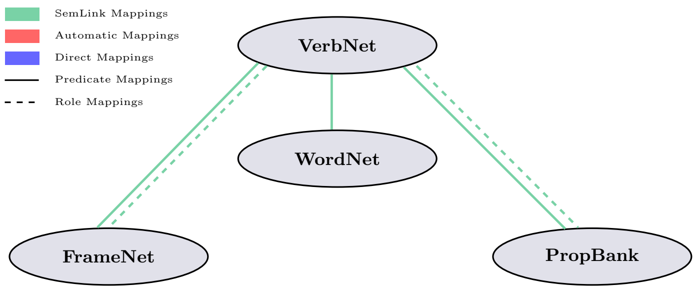


Figure 4.1: SemLink graph representation.

Our study includes all the direct mappings included in SemLink. As shown in Figure 4.1, VerbNet is the central resource in SemLink. Thus, we detail the coverage and gaps of the mappings between every other resource and VerbNet at both the lexical and role level.

4.2 A Study of SemLink Coverage

4.2.1 WordNet and VerbNet alignment

Although VerbNet is one of the largest verb lexicons available it does not reach the coverage of the verbal part of WordNet. While WordNet contains **25,047** different verb senses there are just **6,293** predicates in VerbNet classes. This means that the mapping between both resources is, obviously, incomplete. Specifically there are **18,764** verb senses of WordNet, corresponding to **9,995** different lemmas, that have not been assigned to any VerbNet predicate. In other words, the 74.92% of WordNet verb senses are not in VerbNet. Many of these cases appear because of the distinct granularities of both resources. In fact **6,120** WordNet senses (corresponding to **2,099** lemmas) that are not mapped to VerbNet belong to lemmas that have at least another WordNet sense properly mapped to VerbNet (this corresponds to the 32.62% of WordNet senses that are not mapped to VerbNet). For instance, Table 4.1 shows the mapping between the verb **drown** in WordNet and VerbNet. Note that only two of the six WordNet senses are assigned to VerbNet. Conversely, all the VerbNet senses of the verb **drown** are aligned to at least one WordNet sense. In addition, one of the verb senses of WordNet is aligned to more than one sense of VerbNet. This is the case of the sense drown$_v^4$. And, if we look in the opposite direction, more than one verb sense of WordNet have been added to the same VerbNet class. That is, a VerbNet predicate is aligned to more than one senses of WordNet. This is the case of the verb drown in class ***suffocate-40.7*** of VerbNet.

VerbNet		WordNet
class	member	sense
40.7	drown	drown$_v^3$
42.2 40.7	drown drown	drown$_v^4$
-	-	drown$_v^2$
-	-	drown$_v^1$
-	-	drown$_v^5$
-	-	drown$_v^6$

Table 4.1: WordNet to VerbNet alignment for the verb **drown**.

From the rest of missing senses, most correspond to those cases where the lemma does not exist in the VerbNet lexicon (**7,320** lemmas and **11,201** senses). For example the verb **abort** does not appear in VerbNet since its three WordNet senses are not part of SemLink. The remaining cases (**1,443** WordNet senses and **576** lemmas) correspond to lemmas that exist in both resources but there is no sense mapping between them. For instance, there is no mapping between the WordNet sense harm$_v^1$ and the VerbNet verb that belongs to the class ***amuse-31-1***. Summarizing, 32.62% of the verb senses of WordNet that have not been assigned to any VerbNet predicate belong to lemmas that have at least another WordNet sense properly mapped to VerbNet, a 59.69% belong to lemmas that does not exist in the VerbNet lexicon and finally, a 7.69% to lemmas that exist in both resources but there is no sense mapping between them.

Moreover, SemLink does not provide mappings to WordNet senses for **1,077** VerbNet predicates, the 17.11% of the total of VerbNet predicates. **304** of these VerbNet predicates share the same lemma with some other VerbNet sense that is already mapped to a WordNet sense. This is the case of the verb **reveal** as shown in Table 4.2. Note that only three of the five

VerbNet senses are assigned to WordNet. Moreover, these three senses of VerbNet have been aligned to the same sense of **reveal** in WordNet. The verb **reveal** has another two senses in WordNet that has not been aligned to VerbNet.

VerbNet		WordNet
class	member	sense
29.2	reveal	reveal_v^2
37.7	reveal	reveal_v^2
37.10	reveal	reveal_v^2
48.1.2	reveal	-
78	reveal	-

Table 4.2: VerbNet to WordNet alignment for the verb **reveal**.

From the rest of missing members, **574** correspond to those cases where the lemma of the predicate also exists in WordNet (like the example of **harm** explained previously). Finally, there are only **199** verb senses in VerbNet whose lemmas do not exist in WordNet. For example: **africanize**, **backfill** or **carbonify**. In percentage, 28.23% of the verb senses of VerbNet where SemLink does not provide mappings to WordNet belong to lemmas that have at least another VerbNet sense properly mapped to WordNet, a 53.30% belong to lemmas that exist in both resources but there is no sense mapping between them and a 18.48% belong to lemmas that does not exist in the WordNet lexicon.

4.2.2 PropBank and VerbNet alignment

The mapping between PropBank and VerbNet introduces additional complexity to the comparison of both resources. In this case, aligning the resources means that the arguments of the PropBank predicates must be aligned to the VerbNet thematic-roles.

First, regarding the lexicon mapping, once again, the differences in the coverages of the resources impede to obtain a complete alignment. From the

6,181 different PropBank predicates (comprising **4,552** lemmas), just **3,558** have their corresponding VerbNet predicate in SemLink. This is 57.56 % of the total of predicates in PropBank. That is, **2,623** PropBank predicates have no correspondences to VerbNet. However, all the lemmas of PropBank are contained within the VerbNet lexicon. This means that for each one of the **2,623** missing predicates from PropBank there exists at least another predicate with the same lemma that is mapped to VerbNet. That is the case of the PropBank predicate **abandon.02**, shown in Table 4.3.

VerbNet		PropBank
class	member	predicate
13.5.2	accept	accept.01
29.2-1-1	accept	
77	accept	
51.2	abandon	abandon.01
-	-	abandon.02

Table 4.3: PropBank to VerbNet alignments for **accept** and **abandon** verbs.

On the contrary, we found that the number of VerbNet predicates that are not aligned to PropBank is smaller than the number of PropBank predicates not aligned to VerbNet. That is, up to **4,736** of the **6,293** VerbNet predicates are aligned to PropBank while only **1,557** VerbNet predicates are not aligned to PropBank. The alignment of 24.74% of predicates from VerbNet to PropBank is missing in SemLink. Moreover, **298** of these VerbNet predicates do not exist in the PropBank lexicon. For instance, **arrogate**, **deconstruct**, **mewl** or **sprint** are some of the verb lemmas that do not belong to the lexicon of PropBank. Finally, there are **312** VerbNet predicates whose lemmas (**265** in total) are actually part of the PropBank lexicon but there is no alignment for them. For example, the predicate **offload** of the VerbNet class ***wipe_manner-10.4.1*** is not connected to the PropBank predicate **offload.01**. For the rest, there exists the lemma in PropBank and there is some other alignment for that lemma (for some other VerbNet predicate with that lemma). Table 4.4 shows some alignments from VerbNet to PropBank. In VerbNet, a single sense for the verb **laugh** is considered. It belongs to the ***nonverbal_expression-40.2*** class and it is mapped in SemLink to its corresponding PropBank predicate **laugh.01**. The verb **flow** can be found in both resources, but only one of the two senses of this verb in VerbNet, belonging to the ***entity_specific_modes_being-47.2*** class, is aligned to PropBank.

VerbNet		PropBank
class	member	predicate
40.2	laugh	laugh.01
47.2	flow	flow.01
48.1.1	flow	-

Table 4.4: VerbNet to PropBank alignments for *laugh*$_v$ and *flow*$_v$.

Regarding the PropBank arguments and the VerbNet thematic-roles, **7,915** out of **15,871** arguments from PropBank[1] are mapped to a thematic-role from VerbNet[2]. That is, around a half of the total PropBank arguments, leaving out the remaining **7,956** arguments. From the opposite point of view, **9,682** out of **17,382** thematic-roles from VerbNet are included in the SemLink mapping. This means that **7,700** thematic-roles are not aligned to any PropBank argument. Table 4.5 contains some examples of existing and also missing mappings between PropBank arguments and VerbNet thematic-roles. For instance, the first example, the one concerning to the verb **paint**, shows a fully complete mapping at lexicon and role level between these two resources. Conversely, the verb **plant** has three senses in VerbNet, but for the sense that belongs to class ***spray-9.7***, the mapping to PropBank is non-existent at the lexicon and role level. Instead, the other two senses of the verb **plant** in VerbNet, belonging to the classes ***establish-55.5*** and ***put-9.1-1***, are aligned to PropBank's **plant.01** predicate and its arguments. Finally, the last example belongs to a case where the mapping is partial. In this particular case, the predicate **abandon** of VerbNet class ***leave-51.2*** is aligned to the predicate **abandon.01** of ProbBank, but at the role level, there is only mapping for the thematic-role *Theme*. Contrarily, the thematic-role *Initial_Location* lacks alignment to a PropBank argument, and in turn, arguments arg_1 and arg_2 of PropBank also have no correspondence in VerbNet.

[1] Arguments of particular PropBank predicates. For instance, arg_0 of *paint.01*.

[2] Thematic-roles of particular VerbNet predicates. For instance,*Agent* of the class member *paint*

VerbNet			PropBank	
class	member	thematic-role	predicate	argument
9.9	paint	Agent	paint.01	arg_0
9.9	paint	Destination	paint.01	arg_1
9.9	paint	Theme	paint.01	arg_2
9.7	plant	Agent	-	-
9.7	plant	Destination	-	-
9.7	plant	Theme	-	-
51.2	abandon	Theme	abandon.01	arg_0
51.2	abandon	Initial_Location	abandon.01	-
-	-	-	abandon.01	arg_1
-	-	-	abandon.01	arg_2

Table 4.5: Some alignments between VerbNet thematic-roles and PropBank arguments.

4.2.3 FrameNet and VerbNet alignment

The mapping between FrameNet and VerbNet means that, on the one hand, VerbNet classes are aligned to FrameNet frames and on the other hand, the frame-elements of the FrameNet frames must be aligned to the VerbNet thematic-roles. Ultimately, VerbNet predicates[3] are aligned to its corresponding lexical-units from FrameNet[4]. The alignment between FrameNet and VerbNet proves to be very incomplete. For example, only **1,730** lexical-units from FrameNet are aligned to, at least, one VerbNet predicate. This number represents only 16% out of the total **10,195** lexical-units of FrameNet. Table 4.6 presents some alignments between VerbNet predicates and FrameNet lexical-units. For instance, the verbs **sell** and **buy** are aligned at predicate level. But, for the case of the verb **delay**, the alignment of the two resources in SemLink is not complete. The lexical-unit **delay.v** belongs to two different semantic frames in FrameNet, but only the lexical-unit from the frame **Hin-**

[3] Predicate of a particular VerbNet class. For instance, class member *sell* from *give-13.1-1* VerbNet class.

[4] Lexical-units of particular FrameNet frames. For instance, *sell.v* from the frame *Commerce_sell*.

dering has a mapping to the predicate **delay** of the VerbNet class ***linger-53.1-1***. The mapping is missing for the lexical-unit **delay.v** of the frame **Change_event_time**, although this frame is mapped to the ***linger-53.1-1*** class. Semlink also lacks for the alignments for the **employ** predicate, although it is part of both VerbNet and FrameNet and the classes ***hire-13.5.3*** and ***use-105*** are mapped to the frames **Employing** and **Using** respectively.

VerbNet		FrameNet	
class	member	frame	lexical-unit
13.1-1	sell	Commerce_sell	sell.v
13.5.1	buy	Commerce_buy	buy.v
53.1-1	delay	Hindering	delay.v
53.1-1	delay	Change_event_time	-
53.1-1	-	Change_event_time	delay.v
13.5.3	employ	Employing	-
13.5.3	-	Employing	employ.v
105	employ	Using	-
105	-	Using	employ.v

Table 4.6: Some alignments between VerbNet predicates and FrameNet lexical-units.

SemLink also includes the alignment between the semantic roles of both resources. However, unlike PropBank, the roles of FrameNet, that are called frame-elements, are defined at frame-level and not at predicate level. Therefore, the mapping of the VerbNet thematic-roles and the frame-elements of FrameNet is defined between VerbNet classes and FrameNet frames. Table 4.7 presents an example of the alignment of some roles from both resources for the VerbNet class ***register-54.1***. This class in particular, groups verbs such as **clock**, **time** and **mistime**. It is aligned to **Adding_up** frame and its thematic-roles *Agent, Theme* and *Value* are aligned to *Cognizer*, *Numbers* and *Result* frame-elements respectively.

Once again, the mapping between VerbNet and FrameNet presents significant gaps and mismatches. For instance, at role level, just **825** of the

VerbNet		FrameNet	
class	thematic-role	frame	frame-element
54.1	Agent	Adding_up	Cognizer
54.1	Theme	Adding_up	Numbers
54.1	Value	Adding_up	Result

Table 4.7: Some alignments between VerbNet thematic-roles and FrameNet frame-elements.

7,124 frame-elements of FrameNet[5] are linked to a VerbNet thematic-role. That is, **88%** of the frame-elements from FrameNet are not aligned to any VerbNet thematic-role. Moreover, only **262** frames out of **795** (33%) have at least one frame-element aligned to a VerbNet thematic-role. That is, just a few frames are used in the mapping. However, it also seems that, at a class level, most of the VerbNet thematic-roles appear to be aligned to at least one frame-element. VerbNet covers **787** different thematic-roles[6]. From these, **541** appear to be aligned to a FrameNet frame-element. This means that around 69% of the thematic-roles are aligned to at least one FrameNet frame-element. In other words, it seems that just **246** thematic-roles, are missing from the mapping provided by SemLink. Table 4.8 presents some class level alignments between VerbNet thematic-roles and FrameNet frame-elements.

The VerbNet class ***fulfilling-13.4.1*** is not aligned to any frame, and, obviously, its thematic-roles to any frame-elements either. It can be said that at class level there is no alignment to FrameNet.

The class ***occurrence-48.3*** of VerbNet is aligned to the frame **Catastrophe** and its two thematic-roles, *Theme* and *Location*, are aligned to a frame-element of the mentioned frame. Curiously, the thematic-role *Location* of VerbNet class ***occurrence-48.3*** is mapped to two different frame-elements of FrameNet frame *Catastrophe*: *Place* and *Time*. The VerbNet class ***occurrence-48.3*** only has the *Theme* and *Location* thematic-roles but in VerbNet there is a thematic-role for *Time*, so, the mapping between Verb-

[5] Frame-elements of a particular FrameNet frame. For instance, the frame-element Cognizer for the *Adding_up* frame

[6] Role of a particular VerbNet class. For instance, Agent of VerbNet class fire-10.10

VerbNet		FrameNet	
class	thematic-role	frame	frame-element
13.4.1	Agent	-	-
13.4.1	Theme	-	-
13.4.1	Recipient	-	-
48.3	Theme	Catastrophe	Undesirable_Event
48.3	Location	Catastrophe	Place
48.3	Location	Catastrophe	Time
48.3	-	Catastrophe	Cause
48.3	-	Catastrophe	Circumstances
48.3	-	Catastrophe	Degree
48.3	-	Catastrophe	Manner
48.3	-	Catastrophe	Undergoer
-	-	Addiction	Addict
-	-	Addiction	Addictant
-	-	Addiction	Compeller
-	-	Addiction	Degree
-	-	Addiction	State

Table 4.8: Some alignments between VerbNet thematic-roles and FrameNet frame-elements.

Net *Location* thematic-role and FrameNet *Time* frame-element seems to be a mistake. The rest of the frame-elements of the frame **Catastrophe** do not have alignments to VerbNet thematic-roles.

Finally, there are frames which are not part of the Semlink mapping. For example, the frame **Addiction** and its frame-elements are not included in the set of alignments.

4.3 Conclusions

In this chapter, we have presented a complete study of the coverage of the mappings encoded in SemLink. We have seen that the mapping between the different sources of predicate information is far from being complete. For instance, the alignment between VerbNet and FrameNet proves to be the least complete one. Only **1,730** lexical-units of FrameNet are aligned to, at least, one VerbNet predicate. This number represents only the **16%** of the total **10,195** lexical units of FrameNet. Moreover, not only the lexicon but the role sets of both resources are weakly connected. For instance, just **825** of the **7,124** existing frame-elements of FrameNet are linked to a VerbNet thematic-role. That is, **88%** of the frame-elements of FrameNet are not aligned to any VerbNet thematic-role.

The mapping at lexicon level between PropBank and VerbNet is also incomplete. From the **6,181** different PropBank predicates, **2,623** have no connection to VerbNet. This means that only the **57%** of the total PropBank predicates are aligned to, at least, one VerbNet predicate. Therefore, half of PropBank's predicates remain to be aligned. Regarding the PropBank arguments and the VerbNet thematic-roles, around a half of the total PropBank arguments (**7,915** out of **15,871** arguments) are mapped to a thematic-role from VerbNet. From the opposite point of view, 9,682 out of 17,382 thematic-roles from VerbNet are included in the SemLink mapping. This means that 7,700 thematic-roles are not aligned to any PropBank argument. As with predicates, around half of the roles of both resources are missing to align.

Moreover, SemLink does not provide a complete alignment between VerbNet and WordNet. Specifically there are **18,559** verbal senses of WordNet, corresponding to **9,995** different lemmas, that have not been assigned to any VerbNet predicate. In other words, this means that the 74.92% of WordNet verb senses are not in VerbNet.

The gaps found in this analysis shows the difficulties in manually mapping the predicative information from different resources and motivates the development of the automatic techniques presented in the following chapter.

CHAPTER 5

Automatically extending the semantic interoperability between predicate resources

In this chapter we present an approach to improve the interoperability between four semantic resources that incorporate predicate information. After a motivation of this work in Section 5.1, we describe our proposal for mapping the semantic knowledge included in WordNet, VerbNet, PropBank and FrameNet and prove that our approach provides productive and reliable mappings in Section 5.2. Next, in Section 5.3 we introduce the new lexical-semantic resource built applying the methodology described in the previous section. Finally, we present some concluding remarks about this approach in Section 5.4.

5.1 Introduction

In chapter 4 we describe the gaps in the SemLink coverage. We show that the mapping between the different sources of predicate information integrated in SemLink is far from being complete. In addition, as we mentioned in Chapter 2, the mappings between resources have been manually developed. Building or manually integrating large and rich enough predicate models for new languages and domains is resource intensive and thus expensive. In this chapter, we propose a set of automatic methods in order to alleviate this problem at both the lexical and role level.

The lexical mappings are centralized in WordNet in order to offer a wider coverage. We apply graph-based algorithms on three different resource-pairs: WordNet and VerbNet, WordNet and FrameNet, and WordNet and PropBank. Regarding the roles, we propose two different approaches to infer new mappings between the following resource-pairs: VerbNet and FrameNet, and PropBank and FrameNet.

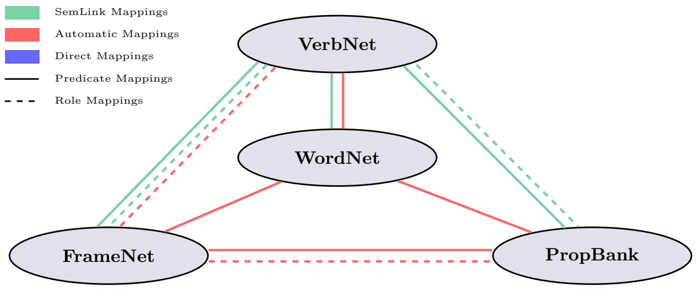


Figure 5.1: Predicate Matrix lexical and role mappings graph representation.

The new set of mappings obtained by our automatic methods are integrated with those in SemLink into the Predicate Matrix as shown in Figure 5.1. This way, we provide a more robust and complete interoperability between the predicative resources.

For example, consider the verb **struggle** that belongs to the VerbNet class ***battle-36-4-1***. Given the sentence "John struggled with Mary for the last piece of cake.", an automatic semantic parser based on PropBank should annotate the sentence with the **struggle.01** PropBank predicate and the roles arg_0, arg_1 and arg_2, as shown in Figure 5.2.

By means of SemLink, we know that the **struggle.01** PropBank predicate belongs to the VerbNet class ***battle-36-4-1*** and the arg_0, arg_1 and arg_2 PropBank arguments are aligned to the VerbNet *Agent*, *Co-Agent* and *Topic* thematic-roles respectively. SemLink also offers the mapping to WordNet but for this particular predicate, it lacks information from FrameNet.

Thanks to the methodology followed in this chapter, the Predicate Matrix contains mappings to VerbNet and PropBank for the **Hostile_encounter** frame and the **struggle.v** lexical-unit. It is also defined that *Agent* and arg_0 are equivalent to *Side1* frame-element in FrameNet, *Co-Agent* and arg_1 to

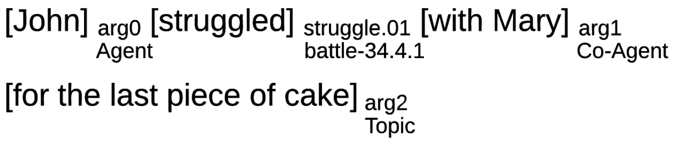

[John] arg0 Agent
[struggled] struggle.01 battle-34.4.1
[with Mary] arg1 Co-Agent
[for the last piece of cake] arg2 Topic

Figure 5.2: PropBank information obtained from an automatic Semantic Role Labeling and the corresponding VerbNet mappings obtained from SemLink.

Side2, and *Topic* and arg_2 correspond to *Issue* and *Purpose* frame elements. In that way, the annotations of one particular semantic resource can be projected to any other of the resources integrated into the Predicate Matrix, as shown in Figure 5.3. Moreover, now the rich predicate information encoded in FrameNet is also available for further semantic processing.

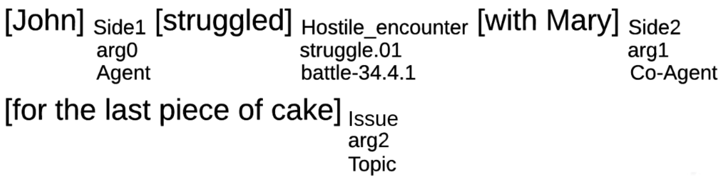

[John] Side1 arg0 Agent
[struggled] Hostile_encounter struggle.01 battle-34.4.1
[with Mary] Side2 arg1 Co-Agent
[for the last piece of cake] Issue arg2 Topic

Figure 5.3: FrameNet information obtained from an automatic Semantic Role Labeling and the corresponding PropBank and VerbNet mappings obtained from the Predicate Matrix.

In the following section, we describe our methodology to automatically integrate predicate information.

5.2 Automatic mappings between lexical entries and roles

This section presents the set of automatic methods based on advanced graph-based Word Sense Disambiguation (WSD) algorithms and corpus alignments to automatically establish the appropriate mappings among lexical entries and roles of semantic resources that incorporate predicate information.

The integration of predicate information is performed at two levels: lexical and role levels. Table 5.1 summarizes the type of mappings we present per section. Each section describes a method to obtain the mappings between resources as well as the evaluation results of the proposed method.

All the mappings obtained at the **lexical** level are based on graph-based WSD algorithms. The lexical mappings from WordNet to FrameNet and VerbNet are obtained by applying WSD algorithms to semantically coherent groupings of verbal entries (see Section 5.2.1.1). The lexical mappings from WordNet to PropBank are obtained by applying WSD to a corpus annotated with PropBank predicates (see Section 5.2.1.2). We have not created new mappings between PropBank and VerbNet because PropBank already offers this information and its coverage is nearly complete.

As it happens with the lexical mappings, PropBank also offers quite complete role mappings between PropBank and VerbNet. Thus, we concentrate our efforts on finding new **role** mappings between FrameNet and VerbNet and between FrameNet and PropBank. The mappings between FrameNet frame-elements and VerbNet thematic-roles are obtained following a three-step methodology (see Section 5.2.2.1). A corpus-based method is used to automatically create new role mappings between FrameNet and PropBank (see Section 5.2.2.2). This method obtains mappings between predicates and roles at the same time.

All these methods are described in the following sections in detail.

5.2.1 Lexical mappings

The methods for extending the mappings between lexical entries are based on a graph-based WSD approach which uses WordNet as a background knowledge base.

Following (Laparra and Rigau, 2009b; Laparra et al., 2010b), we apply knowledge-based WSD algorithms that use a large-scale graph of concepts derived from WordNet to disambiguate the entries from the lexicons.

In the case of FrameNet and VerbNet, the graph-based WSD algorithms are applied to coherent groupings of words belonging to the same FrameNet

	Section	Mappings	Method
Lexical Mappings	5.2.1.1	WN - FN WN - VN	Lexicon disambiguation
	5.2.1.2	WN - PB	Crossing SRL (predicates) and WSD corpus annotations
Role Mappings	5.2.2.1	FN - VN	Learning role patterns and frequencies
	5.2.2.2	FN - PB	Crossing SRL corpus annotations

Table 5.1: Summary of lexical and role mappings. WN: WordNet; FN: FrameNet; VN: VerbNet; PB: PropBank.

frame or VerbNet class. For PropBank, the WSD approach is applied to a corpus annotated with PropBank predicates. In all cases, the disambiguation provides new links between those verbal entries and the WordNet senses. Thus, we can connect verbs from different resources that are connected to the same WordNet sense.

We tested two different graph-based WSD algorithms. An advanced version of the Structural Semantic Interconnections algorithm (SSI) (Navigli and Velardi, 2005) called SSI-Dijkstra+ (SSID+) (Cuadros and Rigau, 2008; Laparra and Rigau, 2009b; Laparra et al., 2010b) and UKB (Agirre and Soroa, 2009b). SSI-Dijkstra+ is a greedy graph algorithm that disambiguates a set of words by calculating the shortest path distances between word senses. UKB applies the Personalized PageRank (Page et al., 1999) on a graph to rank the possible senses and perform disambiguation. Both algorithms use the graph formed by the senses and the semantic relations of WordNet.

5.2.1.1 WordNet-FrameNet and WordNet-VerbNet

We extend the lexical mappings from VerbNet and FrameNet to WordNet taking advantage of the fact that both resources group semantically related lemmas in coherent semantic classes or frames. Our strategy is to apply a WSD algorithm using those groupings as contexts. For that, we have used UKB (Agirre and Soroa, 2009b) and SSID+ (Laparra et al., 2010b).

Although FrameNet covers more than 10,000 lexical-units and 795 frames, only 721 frames have at least a lexical unit associated. From those, 10,086 lexical-units (word-frame pairs) are recognized by WordNet (out of 92%) corresponding to 708 frames and 2,867 verbs.

In FrameNet, the lexical units of a frame can be nouns, verbs, adjectives and adverbs representing a coherent and closely related set of meanings that can be viewed as a small semantic field. For example, the frame **Education_teaching** contains lexical units referring to the educational activity and their participants. It is evoked by lexical units like **cram.v**, **instruction.n**, **instruct.v**, **learn.v**, **lecturer.n**, **study.v**, etc. The frame also defines core frame-elements such as *Student* or *Subject* that are semantic participants of the frame and their corresponding lexical-units.

VerbNet also groups semantically related verbs. It groups 4,403 verbs in 386 classes and subclasses. From those, 6,078 verbal senses (verb-class pairs) are recognized by WordNet (out of 97%).

For instance, the VerbNet class ***learn-14*** groups together verbs like **assimilate**, **cram**, **glean**, **learn**, **memorize** or **read**. This VerbNet class also defines a set of thematic-roles: *Agent*, *Source* and *Topic*.

Evaluation As SemLink includes some manual assignments of WordNet senses to VerbNet and FrameNet, we can use them to evaluate the accuracy of the automatic mappings. For the evaluation, we used as gold-standard 272 VerbNet classes and their associated verbs and 214 FrameNet frames having at least one WordNet sense manually assigned to a verb. The average length of the contexts or coherent semantic groupings is 23.30 verbs for VerbNet and 19.38 lexical units for FrameNet. For comparison, we built a baseline system which assigns to each verb the most frequent sense according to WordNet.

Table 5.2 presents the precision (P), recall (R) and F1 measure (harmonic mean of recall and precision) of the different methods and knowledge resources when mapping WordNet to VerbNet and FrameNet. *WN* stands for the Lexical Knowledge Base (LKB) built using only the relations from WordNet while *WN+G* refers to the LKB also integrating the relations from

VerbNet	Method	LKB	P	R	F1
	baseline	-	18.7	15.4	16.9
	UKB	WN	84.2	84.2	84.2
	UKB	WN+G	**85.3**	**85.3**	**85.3**
	SSID+	WN	83.8	83.5	83.7
	SSID+	WN+G	83.8	83.5	83.7
FrameNet	Method	LKB	P	R	F1
	baseline	-	72.5	70.4	71.4
	UKB	WN	79.0	79.0	79.0
	UKB	WN+G	79.4	79.4	79.4
	SSID+	WN	82.5	81.3	81.9
	SSID+	WN+G	**82.9**	**81.8**	**82.4**

Table 5.2: Results of the disambiguation process when mapping WordNet to VerbNet and FrameNet.

the semantically tagged glosses.[1] Table 5.2 also presents the baseline system results. We observe very high results and robust behavior independently of the WSD algorithm and LKB, and in every case the *baseline* is widely outperformed. We could expect even higher results when also including the gold-standard cases from SemLink in the WSD process.

5.2.1.2 WordNet-PropBank

In PropBank, each predicate, which has no relation with any other predicate, has its own unique role structure. For this reason, we propose a slightly different method to extend the lexical mappings between PropBank and WordNet. We use the WordNet based WSD algorithms to disambiguate a corpus annotated with PropBank predicates. Then, the method obtains the most common matches between the annotations of both resources over the same verbs, as in Figure 5.4.

We use two different sources of contexts.[2] First, the annotated subset of

[1] `https://wordnetcode.princeton.edu/glosstag.shtml`

[2] We obtain better results combining all sources of contexts than exploiting them separately.

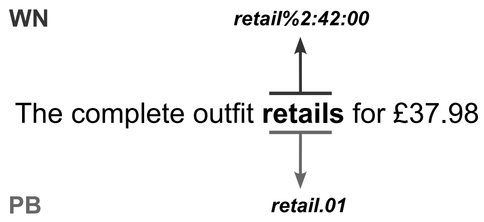


Figure 5.4: Example of matching annotations of WordNet (WN) and PropBank (PB).

the PropBank corpus distributed by the CoNLL shared task, that assures a fully reliable SRL annotation for 500 documents. Second, the FrameNet corpus that includes 99 documents with continuous text and 168,519 sample sentences for the 64% of the lexical units. This corpus does not contain PropBank annotations, so the annotations must be obtained from an automatic SRL processing. To disambiguate the corpora with WordNet senses we use UKB and SSID+. To tag PropBank predicates on the FrameNet documents we apply the mate-tools[3] (Bohnet, 2010) pipeline. The pipeline includes a highly accurate SRL module that obtains 95.59% F1 performance identifying the appropriate PropBank predicates (Björkelund and Hafdell, 2009). In this way, we obtain a full set of documents containing both PropBank (some of them manually annotated and others predicted) and WordNet annotations. By crossing both annotations we obtain PropBank predicates and WordNet senses for some words. Then, for each predicate we select its most frequent corresponding sense obtaining a set of mappings between the lexicon of both resources.

Evaluation In the case of PropBank, we build a gold-standard by recovering from SemLink the set of predicates manually connected to WordNet senses. We also built a baseline system which matches the most frequent predicate in the PropBank corpus with the most frequent sense according to WordNet. For instance, in the case of the verb **sell**, the baseline system matches **sell.01** and sell_v^1

Table 5.3 presents the precision (P), recall (R), and F1 measure of the different methods and knowledge resources when mapping WordNet to PropBank. It also presents the baseline system results. All the strategies outper-

[3] `https://code.google.com/p/mate-tools/`

PropBank	Method	LKB	P	R	F1
	baseline	-	**74.9**	24.0	36.4
	UKB	WN	71.3	**58.0**	**64.0**
	UKB	WN+G	70.7	57.2	63.2
	SSID+	WN	67.2	54.7	60.3
	SSID+	WN+G	68.3	55.3	61.1

Table 5.3: Results of the disambiguation process when mapping WordNet to PropBank.

form the *baseline* in terms of *F1 measure* and, in general, the precision shows that our method generates quite reliable mappings.

5.2.1.3 Comparison with SemLink

Table 5.4 shows the number of mappings between WordNet and VerbNet, FrameNet, and PropBank in SemLink and the number of mappings obtained by the best configuration of our automatic methods. It also compares the number of new and common mappings obtained automatically with respect to SemLink. In all the cases, the number of predicates mapped automatically is higher than the predicates mapped in SemLink. Note that our methods only connect each predicate to a single synset of WordNet while SemLink includes several possible links. For example, SemLink takes into account 3,137 predicates of PropBank but they add up to 5,489 mappings to WordNet. On the other hand, we automatically obtain 4,484 links corresponding to exactly the same number of predicates. From these, 2,924 automatic mappings are completely new.

Both results show the appropriateness of our methodology to obtain mappings between WordNet and FrameNet, VerbNet and PropBank. The automatic methods obtain good results in general and the number of mappings is higher compared to the number of mapping offered by SemLink.

	SemLink	Automatic	Intersection	New
VN-WN	7,665 (5,255)	6,081	4,131	1,950
FN-WN	4,851 (2,419)	3,877	1,842	2,035
PB-WN	5,489 (3,137)	4,848	1,924	2,924

Table 5.4: Links between VerbNet, FrameNet, PropBank predicates and WordNet synsets. In parentheses the number of predicates covered by the corresponding set of mappings in SemLink.

5.2.2 Role mappings

In order to infer new role mappings among different predicate schemas, we have defined two methods. Section 5.2.2.1 presents a three-step process to increase the alignments between VerbNet thematic-roles and FrameNet frame-elements. Section 5.2.2.2 explains the corpus-based method used to extend the mappings between FrameNet and PropBank.

5.2.2.1 FrameNet - VerbNet

This method focuses on obtaining the missing correspondences between the semantic roles from VerbNet and FrameNet. The missing links can belong to verbs already included in SemLink or to the verb senses obtained applying the methods presented in Section 5.2.1. The method comprises three different steps that should be applied consecutively. We have set two alternative configurations:

- **Configuration 1-2-3**: it runs Step 1 to 3 and it uses information contained in SemLink;
- **Configuration 2-3**: it runs Step 2 and 3 and it is completely independent from SemLink.

Step 1: The first step learns from SemLink which alignments between VerbNet thematic-roles and FrameNet frame-element names are more frequent independently of the FrameNet frame. For example, Table 5.5 shows the frequencies of the alignments for the thematic-role *Location*.

Thematic-Role	Frame-Element	Frequency
Location	Area	383
Location	Goal	322
Location	Path	177
Location	Ground	78
Location	Sound_source	76
Location	Fixed_location	50
Location	Source	49
Location	Place	41
Location	Location	25
Location	Body_part	21

Table 5.5: Frequencies of the frame-element names mapped to the thematic-role *Location* in SemLink.

For every verb of VerbNet aligned to a frame of FrameNet, we obtain the thematic-roles that have not been assigned to any frame-element. Then, we link each of these roles with the most frequently aligned frame-element in the whole set of frames. For example, the verb **paddle** of the VerbNet class ***spank-18.3*** is mapped to the frame **Corporal_punishment** of FrameNet. However, the thematic-role *Location* of this verb is not linked to any frame-element. The frame **Corporal_punishment** contains frame-elements like *Agent*, *Evaluee*, *Reason*, *Instrument*, *Degree* and *Body_part*. According to the data showed in Table 5.5, *Body_part* is the frame-element of the frame *Corporal_punishment* that is mapped to the thematic-role *Location* in a greater number of times. Thus, we map *Location* to *Body_part*.

In Table 5.6 we present this new mapping and some other *Location* cases obtained by this method.

lemma	VN-class	Thematic-Role	FN-frame	FE
sit	spatial_configuration-47.6	Location	Placing	Area
spew	substance_emission-43.4	Location	Excreting	Goal
move	roll-51.3.1	Location	Change_position_on_a_scale	Path
paddle	spank-18.3	Location	Corporal_punishment	Body_part

Table 5.6: Examples of new frame-elements (FE) mapped to the thematic-role *Location*.

Step 2: For those verbs from VerbNet that are mapped to one particular frame of FrameNet, but none of their thematic-roles are linked to any frame-element, this step aligns the thematic-roles and the frame-elements based on pattern frequencies. This step looks into the examples of use contained in VerbNet to acquire patterns of thematic-roles for each class. Given the following sentence:

I$_{Experiencer}$ **saw** the play$_{Stimulus}$

This step obtains the pattern *Experiencer* - *verb* - *Stimulus* for the VerbNet class ***see-30.1***.

The same process is performed looking into the lexicographic annotations of FrameNet to obtain patterns of *core* frame-elements for each frame, like in the following example:

... she$_{Cognizer\_agent}$ **felt** for it$_{Sought\_entity}$ with her right hand ...

In this case, the pattern *Cognizer_agent* - *verb* - *Sought_entity* for the frame **Seeking** is acquired.

Then, when a verb from VerbNet is mapped to a frame of FrameNet, the most frequent thematic-role pattern for the class of the verb is aligned to the most frequent frame-element pattern for the frame. In this way, the thematic-roles and the frame-elements that share the same positions are mapped.

For instance, the verb **feel** of the class ***see-30.1*** is mapped to the frame **Seeking**, but none of its thematic-roles (*Experiencer* and *Stimulus*) are linked to any of the frame-elements of the frame **Seeking**. Table 5.7 presents the pattern frequencies obtained for the class ***see-30.1*** and the frame **Seeking**.

In this particular case, the step just finds examples that follow the pattern *Experiencer - verb - Stimulus* for the class ***see-30.1*** and two patterns for the frame **Seeking**.

Source	Class/Frame	Pattern			Freq.
VerbNet	see-30.1	Experiencer	v	Stimulus	100%
FrameNet	Seeking	Cognizer_agent	v	Sought_entity	68.6%
		Sought_entity	v	Cognizer_agent	31.4%

Table 5.7: Frequencies of the role patterns in VerbNet class *see-30.1* and frame *Seeking*.

After comparing the most frequent ones, the method aligns the thematic-roles and the frame-elements that share the same positions. According to Table 5.7, the most frequent pattern for the frame **Seeking** is *Cognizer_agent - verb - Sought_entity*. Thus, as Table 5.8 shows, the method links the thematic-role *Experiencer* with the frame-element *Cognizer_agent* and *Stimulus* with *Sought_entity* because they appear in the same relative position with respect to the verb.

Unlike step 1, this step is completely independent of the knowledge that SemLink can provide. As already explained, this second step only makes use of examples of use contained in VerbNet and the lexicographic annotations of FrameNet to learn role patterns and obtain frequencies of each one. Apart from this, this step can be executed independently of the first step. That is, it is not mandatory to execute step 1 in order to execute step 2.

lemma	VN-class	Thematic-Role	FN-frame	Frame-element
feel	see-30.1	Experiencer	Seeking	Cognizer_agent
feel	see-30.1	Stimulus	Seeking	Sought_entity
listen	peer-30.3	Experiencer	Seeking	Cognizer_agent
listen	peer-30.3	Stimulus	Seeking	Sought_entity

Table 5.8: Examples of new mappings between thematic-roles and frame-elements of the frame *Seeking*.

Step 3: This last step follows the same strategy as **Step 1**, but it includes the role mappings obtained automatically. As it is presented in Table 5.9, if we include the automatic links from Steps 1 and 2,[4] the frequencies of the mappings between frame-elements and thematic-roles are different to those obtained in Step 1 (see Table 5.5).

Thematic-Role	Frame-Element	Frequency
Location	Area	341
Location	Goal	213
Location	Place	148
Location	Path	145
Location	Ground	111
Location	Source	83
Location	Sound_source	78
Location	Location	71

Table 5.9: Frequencies of frame-elements mapped to the thematic-role *Location* including the automatic links obtained in Step 1 and Step 2.

Evaluation For this evaluation, we have used as a testing set the existing 6,934 SemLink role alignments between FrameNet and VerbNet. The evaluation process has been the same as the one used for the lexical mappings (cf. Section 5.2.1). For each role mapping, we apply a leave-one-out evaluation process. We learn the frequencies from the whole SemLink except the one we are evaluating. This process allows using the full set of role mappings from SemLink as a gold-standard. Thanks to this process, we have evaluated the method with two different configurations: Configuration 1-2-3 and Configuration 2-3. We have also compared the configurations with a baseline system. For each verb, the baseline matches the most frequent thematic-role in the examples of use of VerbNet with the most frequent frame-element in the lexicographic annotations contained in FrameNet.

[4] As explained before, to discover new alignments, it is possible to start from Step 1 or Step 2.

Table 5.10 contains the number of alignments when executing Configuration 1-2-3. The table shows how each step increments the number of cases covered by the previous method and it also includes the individual evaluation of the methods. It also presents the evaluation results of the baseline system. As it can be seen, the three methods outperform the baseline by more than 30 points in terms of *F1 measure*.

Method	New	Total	P	R	F1
SemLink	-	6,934	-	-	-
baseline	-	10,189	39.9	21.6	28.0
Step 1	4,611	11,545	89.0	88.3	88.6
Step 2	407	11,952	72.3	49.0	58.4
Step 3	523	12,475	81.7	81.1	81.4

Table 5.10: Number of new role alignments and performance when executing Configuration 1-2-3.

The results show that the majority of the new mappings are obtained by Step 1. Steps 2 and 3 are less productive when exploiting SemLink frequencies.

Table 5.11 presents the results when executing Configuration 2-3. This configuration does not require any manual mapping so this configuration provides a fully automatic set of new mappings. The number of final mappings is similar to those obtained by the previous method (see Table 5.10). In this case, the *baseline* is also widely outperformed.

Method	New	Total	P	R	F1
SemLink	-	6,934	-	-	-
baseline	-	10,189	39.9	21.6	28.0
Step 2	7,132	7,132	72.3	49.0	58.4
Step 3	4,137	11,269	63.9	62.0	62.9

Table 5.11: Number of new role alignments and performance when executing Configuration 2-3.

As expected, according to the evaluations shown in Table 5.10 and Table 5.11, the most reliable set of mappings is obtained when using previous manual information, that is Configuration 1-2-3. The influence of this knowledge is more evident comparing the results of the Step 1 and Step 3. Note that these two steps are fundamentally the same but they work with different sets of role-mapping frequencies. The frequencies used in Step 1 are learned directly from SemLink, while in Step 3, the frequencies are calculated adding the new mappings discovered by the previous steps. Obviously, this introduces some noise into the process. In our second configuration, the frequencies for Step 3 are obtained without taking into account SemLink. For that reason, the results in this case are lower.

5.2.2.2 FrameNet - PropBank

To obtain the role mappings between PropBank and FrameNet, our method acquires the most common correspondences between the annotations of both resources over the same sentences (cf. Figure 5.5). The idea is to obtain first a corpus with gold FrameNet annotations and automatic PropBank annotations and a corpus with gold PropBank annotations and automatic FrameNet annotations. Then, we cross the annotations on both corpora to collect the coincidences. This way, we obtain pairs <*PropBank-argument,FrameNet-frame-element*> when the filler of one PropBank argument matches a FrameNet frame-element or vice versa.

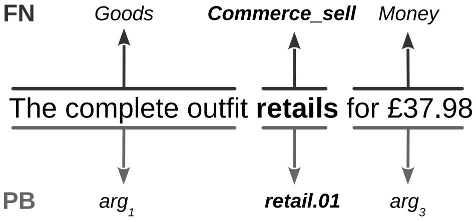


Figure 5.5: Example of matching annotations of FrameNet (FN) and PropBank (PB).

To assure a fully reliable annotation, we exploit existing manually annotated FrameNet and PropBank corpora. The FrameNet corpus can be divided in two different sets. On the one hand, FrameNet version 1.3 includes 168,519 sample sentences for the 64% of the lexical units. On the other hand, it contains continuous text annotations for 99 documents from different sources as WikiNews or the American National Corpus. In the PropBank corpus, the syntactic trees of the Penn Treebank Wall Street Journal data are enriched with PropBank predicate-argument relations. In this work, we use a subset of 500 different documents distributed by the CoNLL shared-task.

To automatically obtain the corresponding counterparts of the data presented above we have made use of two available tools that offer state-of-the-art results on SRL using FrameNet and PropBank. For the FrameNet based annotations we use SEMAFOR[5] (Chen et al., 2010). The parser provides both frame and frame-element identification with an overall performance of **62.76%** precision and **41.89%** recall. The SEMAFOR package includes a modified version of the MST Parser (McDonald et al., 2005) to obtain the required syntactic dependencies. The PropBank based annotation has been done using the mate-tools[6] (Bohnet, 2010). It is a complete multilingual NLP pipeline that includes a highly accurate SRL module that obtains 79.29% F1 performance labeling arguments (Björkelund and Hafdell, 2009).

In this way, we obtain one corpus with manual FrameNet annotations and predicted PropBank annotations using mate-tools. Similarly, we also generate another corpus with manual PropBank annotations and predicted FrameNet annotations using SEMAFOR. We cross both annotations and then we follow two different strategies to obtain different sets of mappings.

The first strategy filters out the cases we consider too infrequent by setting a *threshold* of more than **T** cases per pair <*PropBank-argument,FrameNet-frame-element*>. We apply different values of **T** obtaining different sets of mappings. Finally, we select the most common ones for each predicate. For example, for the predicate **retail.01** we obtain that the arg_1 and the arg_3 match most frequently the frame-elements *Goods* and *Money* of the frame **Commerce_sell** respectively. However, following this strategy the arguments arg_1 and arg_3 of **retail.01** could be also assigned to other frame-elements of other frames, as long as they overcome the threshold **T**.

The second strategy selects for each PropBank argument only its most frequent mapping to a FrameNet frame-element. We first calculate the most common coincidences between PropBank predicates and FrameNet frames. Then, for each predicate we establish a mapping with only one frame. After that, we obtain the most frequent <*PropBank-argument,FrameNet-frame-element*> pair that fits that mapping. As a result, for each argument of each

[5] `http://www.ark.cs.cmu.edu/SEMAFOR/`

[6] `https://code.google.com/p/mate-tools/`

predicate we gather a single mapping with a frame-element. For example, with this strategy we also map the arguments arg_1 and the arg_3 of the predicate **retail.01** with the frame-elements *Goods* and *Money* of the frame **Commerce_sell** respectively, but unlike the previous strategy, no more mappings can be produced for these arguments.

Note that following these **cross-annotation strategies**, we generate mappings between predicates and roles at the same time because the pairs obtained crossing the annotations contain both types of information. The previous example, <*retail.01* - arg_1,*Commerce_sell* - *Goods*>, contains a relation between the predicate **retail.01** and the frame **Commerce_sell** and also a relation between the argument arg_1 of that predicate and the frame-element *Goods* of that frame.

Evaluation We perform two different evaluations. On the one hand, we evaluate the mappings between PropBank predicates and FrameNet frames. For this, we use the set of 2,562 manual mappings of SemLink. On the other hand, we evaluate the mappings between arguments of PropBank and frame-elements of FrameNet. Similarly, we use as the testing set the 4,394 mappings existing in SemLink. We have implemented a baseline that matches the most frequent <*predicate* - *argument*> pair in the manual PropBank annotation with the most frequent <*frame* - *frame-element*> pair in the manual FrameNet annotations.

The results in Table 5.12 contains the performances of both strategies and the baseline. For the first strategy we provide the evaluation with different threshold **T** values. The performance of our second strategy is showed in the *Only-one* row. The results show that the *baseline* is outperformed except when we map predicates using our first strategy with a threshold equal to 7. According to Table 5.12, our second strategy provides the automatic mappings with the highest precision, both for predicates and roles. Obviously, the best recall is obtained by our first strategy with the lowest threshold values, specially for **T**=0, because they are the least restrictive methods.

Table 5.13 shows, for different values of **T**, the number of mappings obtained from the first **cross-annotation strategy** and the number of mappings given by our second strategy (*Only-one*). The table presents, in the *New* columns, how many of these automatic mappings are new. Note that our method obtains mappings for the core arguments of PropBank (arg_0,

	Predicates			Roles		
Method	P	R	F1	P	R	F1
baseline	78.2	47.8	59.3	13.3	22.5	16.7
T=0	76.4	71.3	73.8	60.7	51.5	55.7
T=1	81.3	64.4	71.9	65.5	47.1	54.8
T=4	85.4	52.8	65.3	70.7	38.6	49.9
T=7	86.9	44.7	59.0	73.2	31.8	44.4
Only-one	89.8	52.4	66.2	75.0	41.2	53.2

Table 5.12: Results of the evaluation of the cross-annotation process between PropBank and FrameNet.

arg_1,...) and also for non-core arguments like arg_{loc} or arg_{tmp}. The latter are not considered by SemLink. The last column (*Core*) presents the number of new mappings involving core arguments.

	Predicates		Roles		
	Total	New	Total	New	Core
SemLink	2,562	-	4,394	-	4,394
T=0	3,865	2,038	13,582	11,321	6,095
T=1	3,061	1,411	8,892	6,820	4,282
T=4	2,255	901	5,156	3,462	2,679
T=7	1,845	701	3,667	2,268	1,941
Only-one	2,584	1,242	9,820	8,011	4,117

Table 5.13: Number of mappings obtained with different values of **T** compared to SemLink.

As it can be seen, both configurations obtain a substantial number of new accurate mappings for predicates and roles. As expected our first strategy with **T**=0 is the configuration that provides the highest number of mappings. However, it is remarkable the high number of mappings obtained by the *Only-one* method, the configuration having the highest precision.

5.3 Resulting Predicate Matrix

In Section 5.2 we have presented a set of methods and techniques to automatically integrate different knowledge bases that contain predicate and role information.

Since the methods presented in Section 5.2 can be applied in different ways, in order to generate the **Predicate Matrix**, we select the settings we consider the most appropriate. In most of the cases, we prioritize precision over recall. That is, we give preference to more reliable sets even if they are smaller. Table 5.14 presents the settings used to obtain the automatic mappings.[7]

Lexical entries		
VN-WN	UKB	WN+G
FN-WN	SSID+	WN+G
PB-WN	UKB	WN

Roles	
VN-FN	Steps 1-3
PB-FN	Only-one strategy

Table 5.14: Settings used to obtain the automatic mappings to build the Predicate Matrix 1.2.

Although it is possible to build a new complete resource starting from scratch using these methods, the Predicate Matrix keeps the original mappings provided by SemLink since they are manually created.

Tables 5.15 and 5.16 compare the size of the original SemLink with the result of combining SemLink with the automatically obtained mappings. Table 5.15 presents the differences in terms of mappings between lexicons and Table 5.16 the differences among roles. Although both sets of mappings overlap in many cases, the Predicate Matrix widely outnumbers the set of original mappings in SemLink. First, it provides more verb alignments between VerbNet and FrameNet (from 3,709 to 5,462 in Table 5.15). Second, it also enlarges the WordNet verb sense alignments (from 7,665 to 10,832 VerbNet verb senses and from 4,851 to 8,583 FrameNet verb senses in Table 5.15). Third, the Predicate Matrix doubles the role alignments between VerbNet

[7] Note that the Only-one strategy applied to obtain mappings between PropBank and FrameNet deals with lexical and role mappings.

	WN-VN	WN-PB	WN-FN	VN-PB	VN-FN	PB-FN
SemLink	7,665	5,489	4,851	4,503	3,709	2,562
Predicate Matrix	10,832	9,516	8,583	4,947	5,462	4,163

Table 5.15: Differences between SemLink and the Predicate Matrix: Mappings between lexicons

	PB-VN	FN-VN	FN-PB
SemLink	9,950	6,934	4,384
Predicate Matrix	11,749	14,258	14,195

Table 5.16: Differences between SemLink and the Predicate Matrix: Mappings between roles

and FrameNet (from 6,934 to 14,258 in Table 5.16) and multiplies almost by three the number of role alignments between PropBank and FrameNet (from 4,384 to 14,195 in Table 5.16).

As we explain in Section 5.2, we do not propose any method to map PropBank and VerbNet directly. However, we obtain some new mappings between both resources indirectly. For example, PropBank predicates and VerbNet verbs that are not linked obtain mappings to the same lexical unit of FrameNet. Note that the sets of mappings to FrameNet are highly enlarged by our methods, both for predicates and roles. Recall that FrameNet is the resource with the poorest coverage in SemLink. Specially remarkable is the new set of mappings between PropBank and FrameNet. Unlike SemLink, we provide direct links between those resources. The Predicate Matrix also includes mappings for modifiers (non-core arguments) of PropBank resulting in connections between roles that describe *time*, *location*, *manner*, etc. Moreover, the mappings comprising just core arguments are highly extended. From the **14,195** mappings between PropBank and FrameNet roles in the Predicate Matrix, **10,320** correspond to core arguments. This figure is three times higher than the **4,384** mappings existing in SemLink.

5.4 Conclusions

Building large and rich predicate models takes a great deal of expensive manual effort. Predicate resources such as VerbNet, FrameNet, PropBank and WordNet offer individually interesting characteristics not provided by their alternatives. Unfortunately, these semantic resources are developed independently and they are not fully integrated in a common platform. Thus, a common semantic framework will allow the interoperability among all these resources.

In this Chapter we have presented a proposal which defines a set of automatic methods for mapping the semantic knowledge included in WordNet, VerbNet, PropBank and FrameNet in order to construct the common semantic framework mentioned above. The integration of predicate information is performed first at a lexical level, and second at a role level. After studying different settings for each method using SemLink as a gold-standard for evaluation, we prioritize precision over recall so that we give preference to more reliable alignments. Then, we integrate the mappings automatically obtained with those existing in SemLink.

In summary, all the mappings obtained at the lexical level are based on graph-based Word Sense Disambiguation (WSD) algorithms. More specifically, the lexical mappings from WordNet to FrameNet and also to VerbNet are obtained by applying WSD algorithms to semantically coherent groupings of verbal entries whereas the lexical mappings from WordNet to PropBank are obtained by applying the WSD to a corpus annotated with PropBank predicates. A corpus-based approach crossing different annotations is used to create automatic predicate mappings between FrameNet and PropBank.

We have not created new mappings between PropBank and VerbNet because PropBank already offers this information and its coverage is quite complete. As it happens with the lexical mappings, PropBank also offers quite complete role mappings between PropBank and VerbNet. Thus, we concentrate our efforts on finding new role mappings between FrameNet and VerbNet and between FrameNet and PropBank. The mappings between FrameNet frame-elements and VerbNet thematic-roles are obtained following a three-step process whereas the same corpus-based approach used previously for predicates is applied to automatically create new role mappings between FrameNet and PropBank.

Thanks to the methodology proposed for creating automatic mappings between lexical entries and roles, the resource obtained is much larger than SemLink. Note that the Predicate Matrix arises from the union of SemLink and the set of mappings obtained by our automatic methods. Although the sets of mappings obtained by the Predicate Matrix and SemLink overlap in many cases, our approach obtains a wider coverage compared to the original set of mappings in SemLink. Additionally, these methods also increase the number of mappings that are included in those resources manually aligned. Furthermore, the automatic methodology makes it easier to maintain updated the set of mappings when improved releases of the predicate resources integrated are developed.

Nevertheless, these methods only cover verbal predicates in English. In the next chapter we will describe how to collect other resources to expand the Predicate Matrix to nominalizations and multilingual predicates.

CHAPTER 6

Nominalization and Multilingualism: extending to nominal predicates and to other languages

This chapter describes a straightforward process to incorporate nominal and multilingual resources into the Predicate Matrix. First, we motivate the work presented in this chapter in Section 6.1. In Sections 6.2 and 6.3, we explain how we implement the nominalization of the Predicate Matrix and how we integrate multiple languages respectively. Next, in 6.4 we present the resulting version of the Predicate Matrix and we finalize with some conclusions in Section 6.5.

6.1 Introduction

In the previous chapter, we present the set of automatic methods that we propose for the integration of different models of predicates. However, these methods only cover verbal predicates in English. This imposes some limitations on the interoperability that the Predicate Matrix can provide. For example, to address cross-lingual event detection, a common multilingual framework for event representation is needed. A SRL module enriched with such a framework could help to discover which lexical-semantic units refer to the same events or roles for different languages.

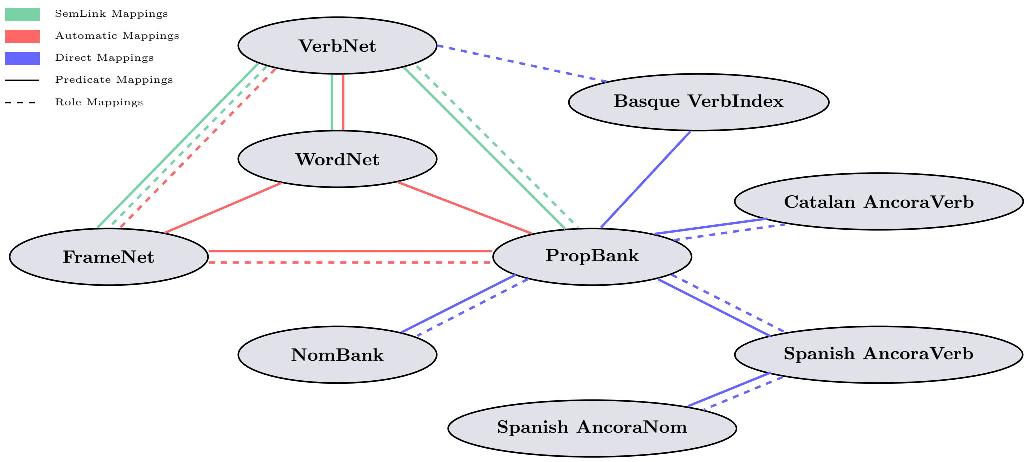


Figure 6.1: Predicate Matrix with multilingual and nominal mappings.

The approaches described in Chapter 5 do not take this issue into consideration, however, overcoming this limitation can be a straightforward process by just gathering other resources linked to any component of the Predicate Matrix.

In the ensuing sections of this chapter, we will explain how if any other resource not included in the Predicate Matrix yet is linked to any of the resources included in the Predicate Matrix the projection to that language or new resource can be done directly. We demonstrate this feature by extending the Predicate Matrix to Spanish, Basque, and Catalan languages and to nominal predicates (English and Spanish), as shown in Figure 6.1. For this purpose, we have made use of the mappings existing between the following resources:

- English nominal predicates:
  PropBank(PB)-NomBank(NB)

- Spanish verbal predicates:
  PropBank(PB)-Spanish AnCora-Verb(SAV)

- Spanish nominal predicates:
  Spanish AnCora-Verb(SAV)-Spanish AnCora-Nom(SAN)

- Catalan verbal predicates:
  PropBank(PB)-Catalan AnCora-Verb(CAV)

- Basque verbal predicates:
  PropBank(PB)/VerbNet(VN)-Basque Verb Index(BVI)

In the case of the English nominal predicates, we use NomBank (Meyers et al., 2004) which contains nominalizations of the PropBank predicates. The projection to Spanish and Catalan is possible thanks to the AnCora (Taulé et al., 2008b) corpus and the AnCora-Verb (Juan Aparicio and Martí, 2008) and AnCora-Nom semantic resources which contain verbal predicates and the nominalizations. Finally, the Basque Verb Index (BVI) (Estarrona et al., 2015) corpus-based lexicon is used in the case of Basque.

Table 6.1 contains the number of mappings listed above both for the lexicon and the roles. Note that the case of Basque is special. Unlike the others, where both predicates and roles are mapped between the same resources, for Basque, the predicates of the Basque Verb Index are mapped to PropBank and the roles are linked to VerbNet.

	PB-NB	PB-SAV	SAV-SAN	PB-CAV	PB-BVI	VN-BVI
lexicon	3,494	7,966	2,966	6,493	506	-
roles	10,307	22,544	9,132	18,660	-	1,408

Table 6.1: Number of mappings between different resources. PB: PropBank; VN: VerbNet; NB:NomBank; SAV: Spanish AnCora-Verb; SAN: Spanish AnCora-Nom; CAV: Catalan AnCora-Verb; BVI: Basque Verb Index.

6.2 Multilingual Predicate Matrix

The strategy to project the Predicate Matrix to new languages is very simple, we *only* need a resource in that language linked to any of the resources included in the Predicate Matrix. This is the case of AnCora (Taulé et al., 2008b)[1], a multilevel corpus that includes both for Spanish and Catalan, annotations of lemmatization, syntactic constituents, WordNet senses, coreference, named entities and also semantic roles. AnCora also develops a semantic resource called AnCora-Verb (Juan Aparicio and Martí, 2008) that contains Spanish and Catalan verbal predicates and their corresponding arguments structures (see Table 6.2).

[1] `http://clic.ub.edu/corpus/ancora`

```
<lexentry lemma=”vender” lng=”es” type=”verb”>
    <sense id=”1”>
        <frame lss=”A32.ditransitive-patient-benefactive” default=”yes” type=”default”>
            <argument argument=”arg0” function=”suj” thematicrole=”agt”/>
            <argument argument=”arg1” function=”cd” thematicrole=”pat”/>
            <argument argument=”arg2” function=”ci” thematicrole=”ben”/>
            ...
<lexentry lemma=”dialogar” lng=”ca” type=”verb”>
    <sense id=”1”>
        <frame lss=”A22.transitive-agentive-theme” default=”yes” type=”default”>
            <argument argument=”arg0” function=”suj” thematicrole=”agt”/>
            <argument argument=”arg1” function=”creg” thematicrole=”tem”/>
            ...
```

Table 6.2: Examples of argument structures defined in AnCora-Verb for the predicates **vender.1.default** and **dialogar.1.default**.

AnCora-Verb is based on PropBank and both resources are linked by a wide set of mappings called AncoraNet. The Spanish predicate **verb.vender.1. default** shown in Table 6.3 is for instance mapped to the English predicate **sell.01** and the Catalan predicate **verb.dialogar.1.default** is mapped to the English **speak.01**.

```
<link ancoralexid=”verb.vender.1.default” propbankid=”sell.01”>
</link>

<link ancoralexid=”verb.dialogar.1.default” propbankid=”speak.01”>
        <arglink ancoralexarg=”arg1” propbankarg=”2”/>
</link>
```

Table 6.3: Examples of mappings between AnCora-Verb and PropBank predicates. Notice that Ancora refers to the ProBank arguments only with the corresponding numbers (i.e. 0 for arg_0 or 1 for arg_1).

Unless AncoraNet states otherwise, the correspondence between the arguments in Ancora-Verb and PropBank is direct. For the Spanish predicate

verb.vender.1.default, the arguments *“arg0”*, *“arg1”* and *“arg2”* correspond respectively to the arguments *“0”*, *“1”* and *“2”* of the English predicate **sell.01**. The Catalan predicate **verb.dialogar.1.default** has its *“arg0”* linked directly to PropBank argument *“0”*, but AncoraNet also explicitly establishes (Table 6.3) that the *“arg1”* corresponds to the argument *“2”* of PropBank. We use these alignments to duplicate the entries in the Predicate Matrix. For example, all the mappings involving the argument *“0”* of the predicate **sell.01** are projected to argument *“arg0”* of the Spanish predicate **verb.vender.1.default**.

In the case of Basque, the mappings are defined in a slightly different way. The predicates of the Basque Verb Index are mapped to PropBank and the roles are linked to VerbNet. For example, as shown in Table 6.4, the Basque predicate **saldu_1** is linked to the English predicate **sell.01** of PropBank and the arguments *“0”*, *“1”* and *“2”* of **saldu_1** are mapped to their corresponding VerbNet roles, *Agent*, *Theme* and *Recipient* respectively.

```
<aditz aditza=”saldu” zenb=”222” >
        <adiera zenb=”1” >
                <ordain zenb=”1” adiera=”sell_01” >
                        <arg zenb=”0” rol=”agent” eadbrol=”abiapuntua” >
                                <case grammcase=”erg”/>
                        </arg>
                        <arg zenb=”1” rol=”theme” eadbrol=”gaia”>
                                <case grammcase=”abs”/>
                        </arg>
                        <arg zenb=”2” rol=”recipient” eadbrol=”helburua”>
                                <case grammcase=”dat”/>
                        </arg>
                </ordain>
        </adiera>
</aditz>
```

Table 6.4: Example of the argument structure defined in Basque Verb Index for the predicate **saldu_1** and its corresponding mappings to PropBank and VerbNet.

Nevertheless, the projection to Basque can be performed because both PropBank and VerbNet are part of the Predicate Matrix.

6.3 Nominal Predicate Matrix

Extending the Predicate Matrix to nominal predicates follows the same strategy as the one previously explained for the projection to a new language, provided the existence of semantic resources containing the argument structures for the nominalizations of the verbal predicates. In the case of English predicates, this knowledge can be obtained from NomBank (Meyers et al., 2004) [2] that includes nominalizations of the PropBank predicates, like **sale.01** aligned to the source verbal predicate **sell.01** (see Table 6.5).

```
<roleset id=”sale.01” name=”commerce: seller” source=”verb-sell.01” vncls=”13.1-1”>
        <roles>
                <role descr=”seller” n=”0”>
                        <vnrole vncls=”13.1-1” vntheta=”Agent”/>
                </role>
                <role descr=”thing sold” n=”1”>
                        <vnrole vncls=”13.1-1” vntheta=”Theme”/>
                </role>
                <role descr=”buyer” n=”2”>
                        <vnrole vncls=”13.1-1” vntheta=”Recipient”/>
                </role>
```

Table 6.5: Examples of argument structures defined in NomBank for the predicate **sale.01**.

For Spanish, AnCora also includes the nominalizations of its verbal predicates in a resource called AnCora-Nom. For example, **venta.1.default**, the nominalization of the predicate **vender.1.default**, is described in AnCora-Nom as shown in Table 6.6.

Once again, unless these resources state otherwise, the correspondence between the arguments of the verbal and nominal predicates is direct. Hence, the lines we showed previously for **sell.01** can be replicated for its nominalization.

[2] `http://nlp.cs.nyu.edu/meyers/NomBank.html`

```
<lexentry lemma=”venta” lng=”es” origin=”deverbal” type=”noun”>
        <sense originlemma=”vender” id=”1” cousin=”no” originlink=”verb.vender.1”
        denotation=”result” lexicalized=”no” wordnetsynset=”16:00721968”>
                <frame appearsinplural=”yes” type=”default”>
                        <argument argument=”arg0” thematicrole=”agt”>
                        </argument>
                        <argument argument=”arg1” foundincorporated=”yes”
                        thematicrole=”pat”>
                        </argument>
                        <argument argument=”arg2” thematicrole=”ben”>
                        </argument>
```

Table 6.6: Description of the nominal predicate **venta.1.default** in AnCora-Nom.

6.4 Resulting Predicate Matrix

Following the strategy presented in the previous sections, we are able to extend the Predicate Matrix to include nominalizations and multilingual predicates. As a result of including NomBank, AnCora-Verb, AnCora-Nom, and the Basque Verb Index into the Predicate Matrix, we have obtained new mappings between these resources and VerbNet, FrameNet, and WordNet. In Tables 6.7 and 6.8, we show the number of new mappings we obtain.

	PB	VN	FN	WN
NB	2,963	3,923	3,911	7,430
SAV	6,745	9,092	8,777	15,310
SAN	4,469	6,190	6,157	10,747
CAV	5,529	7,567	7,347	13,109
BVI	415	652	745	1,330

Table 6.7: Number of lexicon Mappings in the multilingual Predicate Matrix. NB: NomBank; SAV: Spanish AnCora-Verb; SAN: Spanish AnCora-Nom; CAV: Catalan AnCora-Verb; BVI: Basque Verb Index; WN: WordNet; FN: FrameNet; VN: VerbNet; PB: PropBank.

	PB	VN	FN
NB	7,699	9,699	10,351
SAV	17,152	19,173	20,296
SAN	11,752	13,177	14,439
CAV	14,307	16,174	17,204
BVI	1,048	1,275	1,629

Table 6.8: Number of role Mappings in the multilingual Predicate Matrix. NB: NomBank; SAV: Spanish AnCora-Verb; SAN: Spanish AnCora-Nom; CAV: Catalan AnCora-Verb; BVI: Basque Verb Index; WN: WordNet; FN: FrameNet; VN: VerbNet; PB: PropBank.

Table 6.9 shows some examples of the resulting records in the new version of the Predicate Matrix. Note that we have defined an identifier to distinguish between lines for English, Basque, Spanish and Catalan predicates, and between their verbal and nominal forms. This identifier is based on PropBank, AnCora, and the Basque Verb Index predicates and arguments, and is composed of 4 fields: language, form, predicate, and argument. For example, according to Table 6.9, the line that corresponds to the argument *"1"* of the English nominal predicate **sale.01** is identified by **id:eng id:n id:sale.01 id:1**. Similarly, the corresponding line for argument *"arg0"* of the Spanish verbal predicate **vender.1.default** is identified by *"id:spa id:v id:vender.1.default id:arg0"*, as shown in Table 6.9. The Catalan and Basque lines are indexed by *"id:cat"* and *"id:eus"*, respectively. Establishing such identifiers allows us to maintain the whole Predicate Matrix for all the languages in the same file.

id:eng id:n id:sale.01 id:1 vn:give-13.1 vn:Theme wn:ili-30-02244956-v fn:Commerce_sell fn:Goods pb:sell.01 pb:1
id:spa id:v id:vender.1.default id:arg0 vn:give-13.1 vn:Agent wn:ili-30-02244956-v fn:Commerce_sell fn:Seller pb:sell.01 pb:0
id:spa id:n id:venta.1.default id:arg2 vn:give-13.1 vn:Recipient wn:ili-30-02244956-v fn:Commerce_sell fn:Buyer pb:sell.01 pb:2
id:cat id:v id:dialogar.1.default id:arg0 vn:talk-37.11 vn:Agent wn:ili-30-00941990-v fn:Chatting fn:Interlocutor_1 pb:speak.01 pb:0
id:eus id:v id:saldu.1 id:1 vn:give-13.1 vn:Theme wn:ili-30-02242464-v fn:Commerce_sell fn:Goods pb:sell.01 pb:1

Table 6.9: Some examples of mappings in the multilingual Predicate Matrix.

6.5 Conclusions

In this chapter, we have described the extension of the Predicate Matrix to cover nominal and cross-lingual predicates. By gathering other resources that contain links to the components we automatically mapped in Chapter 5, we have yielded a new version of the Predicate Matrix that extends its interoperable capabilities and allows to integrate annotations coming, not only from different predicate schemes but also from different languages. First, we have extended the predicate information to languages other than English, turning it into a multilingual resource. In particular, we have integrated resources in Spanish and Catalan (Ancora-Verb), and Basque (BVI). Secondly, the Predicate Matrix has been extended also to cover nominal predicates by adding mappings to NomBank which contains nominalizations of the PropBank predicates and Spanish Ancora-Nom.

In the next chapter, we present a study on the feasibility of further extending the knowledge contained in the Predicate Matrix through WordNet.

CHAPTER 7

WordNet as a leverage point for knowledge expansion

This chapter gives some insight into WordNet's exploitability to extend the coverage of the Predicate Matrix by exploiting its semantic relations hierarchy. First we introduce this chapter in Section 7.1. After that, in Section 7.2.1, we first analyze the coverage of WordNet in the Predicate Matrix for its different types of verbs. Next, in Section 7.2.2 and Section 7.2.3, we propose straightforward methods to extend mappings coverage and enlarge the knowledge included in the Predicate Matrix dealing with monosemous verbs included in WordNet and the synonymy relation type, respectively. In Section 7.2.4, we analyze shallowly the inheritance of the semantic information included in the Predicate Matrix through the hypernymy/hyponymy relation type of WordNet. In Section 7.3, we explain how we easily include semantic knowledge from the Multilingual Central Repository (MCR) and the positive side-effect of including it on local WordNets linked to the Predicate Matrix. The chapter is finished in Section 7.4 with some concluding remarks.

7.1 Introduction

Although with the set of automatic methods explained in Chapter 5 the level of integration of the resources in the Predicate Matrix increase significantly, three-quarters of WordNet's verbal concepts are still outside the scope of our resource. In this chapter we give just a few hints on how WordNet structure can be exploited to project the knowledge of the predicates already present in the Predicate Matrix to new verb senses.

WordNet offers a large lexical knowledge base for English organized through several semantic relations, such as synonymy and hypernymy, that can be potentially exploited to extend the coverage of any other lexical resource linked to it.

The analysis presented in this chapter, starts from the intuition that verbs belonging to the same WordNet synset, and therefore representing the same sense, should share the same predicative information (e.g. semantic roles, classes). Consequently, WordNet verb synonyms should belong to the same VerbNet class or FrameNet frame. In a similar way, we could expect semantically related verbs to be closely connected in all the resources. Our goal is to explore to what extent this approach can be applied to extend the knowledge in the Predicate Matrix and discover potential limitations and inconsistencies.

In this chapter, we also present the integration of WordNet in the Multilingual Central Repository as a potential source of additional knowledge that can be easily incorporated into the Predicate Matrix.

In the next section, we start analyzing the coverage of the different WordNet lexical files in the Predicate Matrix, which reveals the large amount of verbal lexicon that is not yet integrated.

7.2 Mapping propagation through WordNet

7.2.1 WordNet coverage in the Predicate Matrix

Prior to the analysis of WordNet as a foothold for the extension of our resource, we present a survey on its coverage compared to VerbNet, FrameNet and PropBank. For instance, from the total number of 25,051 WordNet verbal senses, the Predicate Matrix obtained by the automatic methods explained in Chapter 5 only contains 6,443 WordNet verb senses aligned to VerbNet classes. That is, there are 18,608 WordNet verb senses still without mappings to VerbNet classes. In percentage values, only the 23% of the verb senses, a quarter of the total, has an alignment to VerbNet. Similarly, the Predicate Matrix only contains 3,648 WordNet senses aligned to FrameNet frames (around the 14%). Thus, there are 21,500 WordNet word senses without mappings to FrameNet frames. Finally, in relation to PropBank, just a fifth, the 20% of its verb senses are aligned to WordNet's senses, exactly 5,039 verb senses.

LF	WN senses	not in VN (%)	not in FN (%)	not in PB (%)	LF name
29	1,130	871 (77.08)	957 (84.69)	914 (80.88)	body
30	4,171	3,321 (79.63)	4,634 (88.90)	3524 (84.49)	change
31	1,404	1,139 (81.13)	1,225 (87.50)	1154 (82.19)	cognition
32	3,120	2,466 (79.04)	2,667 (85.48)	2482 (79.55)	communication
33	733	620 (84.58)	672 (91.68)	653 (89.09)	competition
34	476	366 (76.89)	406 (85.29)	389 (81.72)	consumption
35	3,698	2,584 (69.88)	3,041 (82.23)	2789 (75.42)	contact
36	1,151	882 (76.63)	982 (85.32)	928 (80.63)	creation
37	763	525 (68.81)	595 (77.98)	475 (62.25)	emotion
38	2,491	1,829 (73.42)	2,020 (81.09)	1919 (77.04)	motion
39	820	579 (70.61)	651 (79.39)	603 (73.54)	perception
40	1,431	1,112 (77.71)	1251 (87.42)	1120 (78.27)	possession
41	2,202	1,835 (83.33)	1,978 (89.83)	1879 (85.33)	social
42	1,409	1,119 (79.42)	1,239 (87.93)	1164 (82.61)	stative
43	146	94 (64.38)	105 (71.92)	113 (77.40)	weather

Table 7.1: WordNet verbal senses not covered by VerbNet classes, FrameNet frames and PropBank framesets in the Predicate Matrix. From left to right: lexicographic file number, number of verb senses pertaining to the lexicographic file, number (and percentage) of verb senses not aligned to a VerbNet classes, number (and percentage) of verb senses not aligned to FrameNet frames, number (and percentage) of verb senses not aligned to PropBank and lexicographic file name.

As an example of the coverage of the Predicate Matrix, table 7.1 shows the distribution according to the lexicographic files from WordNet of the verbal senses not covered by VerbNet classes, FrameNet frames and PropBank framesets in the Predicate Matrix. Interestingly, the coverage of the resources are quite different depending on the area of WordNet selected. The VerbNet coverage ranges from *weather* verbs (it remains 64.38% of WordNet verb senses to be complete) up to *competition* verbs (84.58%) whereas FrameNet coverage ranges from *weather* (71.92%) up to *competition* (91.68%). Once again, the verbs related to the area of *competition* are the most under-covered in PropBank. In contrast to the other resources, *emotion* is the best-covered area.

In general, the lack of coverage is quite large for all areas of WordNet for all the resources. In the case of VerbNet, the percentage of non-aligned verbs ranges from 65% to 85%, for PropBank ranges from 62% to 89% and in the case of FrameNet from 72% to 92%. Hence, the poor coverage is even more remarkable for FrameNet. For example, the *competition* verbs are only covered in an 8.32%. Additionally, the ones that has the greatest coverage are *weather* verbs with an 28.08% of coverage, still far from being complete.

As mentioned, most verb senses belonging to different areas or verb types of WordNet are not covered by the resources integrated in the Predicate Matrix. Exploiting semantic relations connecting predicate senses in WordNet, such as synonymy and hypernym-hyponym relation types, is explored in the following sections with the goal of proposing possible methods to increase the coverage.

7.2.2 Monosemous predicates

In this section, we briefly deal with the monosemous verbs of WordNet. Monosemous verbs from WordNet can be directly assigned to VerbNet predicates still without a WordNet alignment. By adding them, the coverage of the alignment between VerbNet predicates and WordNet senses is incremented. In particular, this very simple strategy solves **240** alignments. In this way, VerbNet predicates such as **divulge**, **exhume**, **mutate**, or **upload** obtain a corresponding WordNet word sense. Obviously, these alignments can be considered just as suggestions to be revised later on manually. In the SemLink coverage analysis presented in Chapter 4 we saw that only **576** lemmas from VerbNet were not aligned to WordNet. Specifically, these cases correspond to lemmas that exist in both resources but there is no sense mapping between them. By just incorporating the monosemous verbs from WordNet, almost 42% are solved.

7.2.3 Synonyms

In this section, we analyze WordNet's synonymy as a way to extend the coverage of the lexicons of the resources included in the Predicate Matrix. Starting from the strong assumption that WordNet synomyms share the same predicate information, we explore what results are obtained by projecting the Predicate Matrix information associated to a particular WordNet sense through its synonymy relations. For instance, the predicate **desert**, member of the VerbNet class ***leave-51.2-1***, is assigned to desert$_v^1$ (*"leave someone who needs or count on you"*) WordNet verbal sense. In WordNet, this word sense also has three synonyms, abandon$_v^5$, forsake$_v^1$, and desolate$_v^1$. According to the previous assumption, these three verbs can also be aligned to the same VerbNet class. Similarly, the semantic frame **Departing** linked to desert$_v^1$ can be extended to its synonyms as well. Table 7.2 shows some productive examples.

VerbNet	WordNet	FrameNet	New
leave-51.2.1	desert$_v^1$	Departing	abandon$_v^5$ forsake$_v^1$ desolate$_v^1$
remove-10.1	retract$_v^1$	—	abjure$_v^1$ recant$_v^1$ forswear$_v^1$ resile$_v^3$
correspond-36.1-1	disagree$_v^1$	Be_in_agreement_on_assesment	dissent$_v^3$ take_issue$_v^1$

Table 7.2: New WordNet senses aligned to VerbNet and FrameNet.

In order to study the effect of the resulting mappings, we conducted a simple experiment using as gold-standard those synsets in SemLink that have more than one lemma aligned to VerbNet or FrameNet. For each lemma in the gold-standard, we remove all its mappings and apply the synonymy strategy. Then, we compare the new assignments with the original ones in the gold standard.

Table 7.3 presents the results of this evaluation in terms of precision (P). The threshold establishes that a lemma of a synset is included as a new lexical member of a class (or a frame), only if more than **T** lemmas of that synset are assigned to that class or frame in SemLink. For example, desert^1_v belongs to the class ***leave-51.2.1*** but none of its synonyms are linked to it. If we set **T**=0 the verbs *abandon*, *forsake* and *desolate* would be included as new members of the class ***leave-51.2-1***. For this synset, using **T**=1 we would not project any information to the rest of the synonyms. As expected, increasing the threshold reduces the number of synonym projections while augmenting the precision.

	VerbNet		FrameNet	
Threshold	P	New Members	P	New LUs
T=0	32.6	12,186	32.2	11,254
T=1	59.6	4,158	62.4	3,680
T=2	74.8	1,988	72.4	1,834

Table 7.3: Results of extending the lexicon of VerbNet and FrameNet with different thresholds **T**.

The results in Table 7.3 show that this method can obtain quite reliable new lexical members depending on the threshold. Surprisingly, the predicate information assigned to different WordNet synonyms seems to be inconsistent. That is, we were expecting verbal synonyms to share their predicate information. Interestingly, this is not the case in the vast majority of cases. In fact, according to these results, predicate information is not shared between synonyms in WordNet.

For example, consider the following WordNet synset <*understand*, *read*, *interpret*, *translate*> with the gloss *"make sense of a language"* and the example sentences *"She understands French; Can you read Greek?"*. As synonyms, these verbs denote the same concept and are interchangeable in many contexts. However, in SemLink, read$^{1}1_v$ is aligned with the VerbNet class ***learn-14-1***[1] while one of its synonyms understand^3_v is aligned with the VerbNet class ***comprehend-87.2***.[2] Moreover, the thematic-roles of both classes are different. **textitLearn-14-1** has the *Agent* (with semantic type [*+animate*]),

[1] http://verbs.colorado.edu/verb-index/vn/learn-14.php#learn-14-1

[2] http://verbs.colorado.edu/verb-index/vn/comprehend-87.2.php#comprehend-87.2-1

Topic and *Source* thematic-roles while ***comprehend-87.2*** has *Experiencer* (with semantic type [*+animate* or *+organization*]), *Attribute* and *Stimulus.*

As we have observed in this analysis, WordNet synonymy cannot be consistently exploited to project the mappings included in SemLink. However, it is not clear why the assumption that synonyms share the same predicative information is not met. The inconsistencies that we found could be due to errors in the SemLink mappings, or be an effect of differences in how synonymy is defined across resources. Answering these questions is a matter for future research for which this experiment may be a starting point.

7.2.4 Hypernymy-hyponymy hierarchy

Hyponymy is a transitive type of relation between two senses where the child sense (hyponym) inherits the semantics of the parent sense (hypernym). Based on this definition, in this section, we analyze whether predicate information such as the semantic frame and roles can be inherited for a specific predicate from its hypernym.

Specifically, we focus on analyzing the inheritability of the VerbNet class and the FrameNet semantic frame. For this purpose, we will take into account the existing mappings in SemLink and analyze the level of compatibility of the VerbNet class and the FrameNet frame between verbs that are hypernyms and hyponyms in WordNet. In total, our analysis cover 7,825 pairs of predicates from VerbNet and 2,222 pairs from FrameNet that are linked in SemLink to WordNet verbs holding a hypernymy relation. For example, the verb sense abdicate$_v^1$ *"give up, such as power, as of monarchs and emperors, or duties and obligations"* is hyponym of the verbs renounce$_v^2$, resign$_v^1$ and vacate$_v^1$ all belonging to the same synset *"leave (a job, post, or position) voluntarily"*. For this case, we check if these three verbs share compatible semantic information with abdicate$_v^1$.

On the one hand, we consider that the VerbNet classes of the two predicates are compatible (and therefore inheritable) when both predicates are members of the same class. Alternatively, we also evaluate a case as compatible if the hyponym belongs to a subclass of the hypernym's class (hyper → hypo) or the other way around (hypo → hyper). On the other hand, we consider that the FrameNet frame of the two predicates are compatible if the two verb senses belong to the same frame or, as we do for VerbNet, when there is any kind of frame relation between the frames of both predicates. In any other case, the inheritance between the predicates is considered incompatible.

Table 7.4 summarises the results of our study and shows the number of compatible and incompatible pairs both for VerbNet and FrameNet.

	VerbNet		FrameNet	
Same class/frame	2567	32.8%	1027	46.2%
Sub-class/frame (hyper → hypo)	246	3.1%	173	7.8%
Sub-class/frame (hypo → hyper)	593	7.6%	27	1.2%
Incompatible	4419	56.5%	995	44.8%
Total	7825	100%	2222	100%

Table 7.4: Compatibility cases of semantic information between hypernyms and hyponyms.

In the case of VerNet, from the 7,825 pairs analyzed 43.5% of the cases are compatible in the terms described above. For the vast majority of these cases (2,567) where the inheritance of the VerbNet class would be possible, both verbs belong to the same class. For instance, abdicate$_v^1$ *"give up, such as power, as of monarchs and emperors, or duties and obligations"* and his hypernym renounce$_v^2$ *"leave (a job, post, or position) voluntarily"*, and the synonyms of the latter, resign$_v^1$ and vacate$_v^1$, are members of the same VerbNet class: ***resign-10.11***. In the remaining 839 pairs, either the hypernym or the hyponym belongs to a subclass. Interestingly, in most of them, it is the hypernym that is mapped to the subclass (593 cases out of 839). Inheritance of the Verbnet class would also be possible between the hyponym waddle$_v^1$ *"walk unsteadily"* and the hypernym walk$_v^1$ *"use one's feet to advance; advance by steps"*, although in this case, the hyponym belongs to the Verbnet class ***run-51.3.2*** and the hypernym to the subclass ***run-51.3.2-1***. Finally, 46.5% of the verb pairs do not share the VerbNet class, for example, the verb sense abandon$_v^3$ *"leave behind empty; move out of"* mapped to the class ***leave-51.2*** and its hypernym leave$_v^1$ *"go away from a place"* mapped to the class ***escape-51.1***.

In the case of FrameNet, 1,027 hypernym-hyponym pairs share the same frame (46% of the cases). For example, the verb sense accompany$_v^2$ *"go or travel along with"* is a hypernym of the verb sense escort$_v^2$ *"accompany or* *escort"*, and, they both are mapped to the same semantic frame **Cotheme**. Moreover, in 200 cases, although both verbs do not belong to the same frame, they are linked through some semantic relation. This is the case of the verb senses murder$_v^1$ *"kill intentionally and with premeditation"* and its hyponym execute$_v^2$ *"murder in a planned fashion"*. While the former belongs to the FrameNet frame **Killing** and the latter to **Execution**, both frames are connected with the "Inheritance" frame relation type. Table 7.5 shows the number of verb pairs per relation type we found in our analysis. It can be seen, that the "Inheritance" relation is, by far, the most frequent. However, there is a large number of pairs (44.8%) where the verbs belong to frames that are not directly related and for which a compatible inheritance could not be assumed. For example, the verb accumulate$_v^1$ *"get or gather together"* belongs to the **Amassing** semantic frame whilst its direct hypernym store$_v^1$ *"keep or lay aside for future use"* belongs to **Storing**, for which there is no associated frame relation.

Frame-relation	hyper → hypo	hypo → hyper
Inheritance	112	26
SubFrame	15	0
Using	28	0
See_also	8	0
Causative_of	9	0
Precedes	1	0
Perspective_on	0	1

Table 7.5: Frecuency of the different frame relation types that share the frames of the hypernym and the hyponym.

From the analysis presented in this section, we can conclude that, although WordNet hypernym shows a promising line of study, it cannot be directly applied without manual supervision. Otherwise, a large amount of semantic information could be incorrectly inherited.

7.3 Fetching additional knowledge from the Multilingual Central Repository

Although the main focus of the Predicate Matrix is the predicates and their roles, the resource can be easily enriched by incorporating additional ontological knowledge linked to any of the resources included in it. WordNet is particularly suitable for this purpose thanks to its integration into the MCR (Gonzalez-Agirre et al., 2012b). The MCR relates senses of WordNets in different languages through the InterLingual Index (ILI). For example, the ili *ili-30-00007739-v* connects the English synset *eng-30-00007739-v* $blink_v^1$ $wink_v^3$ $nictitate_v^1$ $nictate_v^1$ and the Spanish synset *spa-00007739-v* $pestañear_v^1$. The ILIs also link the senses with several ontologies and external references such as Adimen-SUMO (Álvez et al., 2012), the new WordNet domains (González-Agirre et al., 2012) and the Base Level Concept (Izquierdo et al., 2007).

Thus, once a predicate is mapped to a WordNet synset, we can also include the corresponding ontological knowledge into the Predicate Matrix as shown in the examples in Table 7.6. The verbs **stutter** and **squeak** of the VerbNet class ***manner_speaking-37.3*** are mapped to the WordNet senses stutter_v^1 (*stutter%2:32:00*) and squeak_v^1 (*squeak%2:39:00*). In the MCR, the sense stutter_v^1 (*stutter%2:32:00*) is aligned to the ILI *ili-30-00981544-v*, that belongs to the SUMO class *Communication*, to the domain *factotum* and its BLC is the sense speak_v^1 (*speak%2:32:00*). Also, the sense squeak_v^1 (*squeak%2:39:00*) is aligned to the ILI *ili-30-02171664-v*, that belongs to the SUMO class *SoundAttribute*, to the domain *factotum* and, this time, its BLC is the sense sound_v^2 (*sound%2:39:00*).

Additionally, the mappings to the MCR ILIs provide as a side-effect the possibility of extending the lexicons of the local WordNets linked to the Predicate Matrix. For example, in the multilingual Predicate Matrix the Spanish synset *spa-00007739-v pestañear_1* also has associated the predicates **parpadear** and **guiñar** that are not included in the Spanish WordNet. Table 7.7 presents the total number of new senses that can be obtained for different local WordNets. Interestingly, some additional word senses are also created for the English WordNet. This new word sense alignments could be included in future releases of the MCR.

VN_LEMA	**VN_CLASS**	**WN_SENSE**	**FN_FRAME**	**PB_ROLESET**
stutter	37.3	stutter%2:32:00	Communication_manner	stutter.01
	MCR_iliOffset	**MCR_SUMO**	**MCR_Domain**	**WN_BLC**
	ili-30-00981544-v	Communication	factotum	speak%2:32:00
VN_LEMA	**VN_CLASS**	**WN_SENSE**	**FN_FRAME**	**PB_ROLESET**
squeak	37.3	squeak%2:39:00	Communication_noise	squeak.01
	MCR_iliOffset	**MCR_SUMO**	**MCR_Domain**	**WN_BLC**
	ili-30-02171664-v	SoundAttribute	factotum	sound%2:39:00

Table 7.6: MCR knowledge projected to the Predicate Matrix.

	new word senses
English WN	53
Spanish WN	6,092
Catalan WN	5,182
Basque WN	855

Table 7.7: Number of new word senses created for the different WordNets.

7.4 Conclusions

In this chapter, we have studied the exploitability of some of the semantic relations offered by Wordnet in order to extend mapping coverage and enlarge the knowledge in the Predicate Matrix.

One of the straightforward methods studied to extend the mapping between WordNet and VerbNet consists on including synonyms of already aligned WordNet senses as new members of the corresponding VerbNet class and FrameNet frame. Thus, the Predicate Matrix information associated to a particular WordNet sense is projected to its synonyms. This method assumes that synonyms should belong to the same VerbNet class and FrameNet frame. However, as shown in Section 7.2.3, this is not always the case. In order to study the effect of these phenomena, we conducted a simple experiment using as gold standard those synsets in SemLink that have more than one lemma aligned to VerbNet or FrameNet. The results show that the predicate information assigned to different WordNet synonyms seems to be quite inconsistent. That is, according to these results, predicate information is not shared between synonyms in WordNet.

In addition to the synonymy relation, we have studied the hypernymy relation type offered by WordNet between verb senses, considering it, intuitively, the most exploitable for applying the inheritance of semantic information between related predicates. The study has specifically focused on analysing broadly the possible inheritance of the VerbNet class and the FrameNet semantic frame between hypernyms and hyponyms. For this purpose, we have analysed the level of compatibility of the VerbNet class and the semantic frame of FrameNet between pairs of verb senses that are hypernym and hyponym. The VerbNet class is inheritable in 43% of the cases and the semantic framework in 55% of the cases analyzed. Consequently, they would require manual supervision. In addition to semantic relations such as those studied in this chapter, WordNet also includes other types of relations, such as morphosyntactic relations[3] that can be analysed in the future in order to expand the predicative information from verbs to nouns (e.g from govern to government).

As the mappings offered in the Predicate Matrix are also aligned to WordNet, it is possible to use WordNet to enlarge the knowledge included in the Predicate Matrix. For instance, we have easily included semantic knowledge from the MCR.

Finally, we have mentioned a positive side effect of including multilingual resources into the Predicate Matrix. New word sense alignments are created, producing the enrichment of the WordNets integrated into the MCR. This new word sense alignments could be included in future releases of the MCR.

[3] http://wordnetcode.princeton.edu/standoff-files/morphosemantic-links.xls

CONCLUSION AND FURTHER WORK

CHAPTER 8

Conclusion and further work

This last chapter presents a summary (Section 8.1) that reviews the goals we have reached during our research on automatically mapping predicate resources on lexical and role levels. In Section 8.3 we list the research papers we have published that are related to this work. After that, we review the different usages of the Predicate Matrix and the works in which it has been included. Finally, Section 8.4 proposes some possible future lines of research.

8.1 Summary

Building large and rich enough predicate models takes a great deal of expensive manual effort. Furthermore, the same effort should be invested for each different language. Predicate resources such as VerbNet, FrameNet, PropBank, and WordNet offer individually some interesting characteristics not provided by their alternatives. Unfortunately, these semantic resources are developed independently and they are not fully integrated into a common platform. Obviously, a common semantic framework allows interoperability between all these resources.

One of the few projects working on the integration of the predicate information is SemLink (Palmer, 2009). It is an interesting approach but it has some limitations. First, the mapping has been manually developed. A very costly process that is also not systematic. Second, its coverage is still far from being complete. A study on the coverage of SemLink is included as part of the research developed in this work.

In this research, we have focused on defining automatic methods for mapping in a systematic way different semantic resources containing predicate information. The aim is to allow more complete semantic interoperability between them. For that, we have worked on the integration of predicate information at lexical and role levels.

As a result of the research, we have also developed the Predicate Matrix, a new lexical-semantic resource resulting from the union of the mappings obtained by our automatic methods and SemLink. Although the sets of mappings obtained by the Predicate Matrix and SemLink overlap in many cases, our approach obtains a wider coverage compared to the original set of mappings in SemLink.

Our lexical mappings are centralized through WordNet in order to offer wider coverage. For that, we apply graph-based WSD algorithms that use WordNet as the background knowledge base in three different scenarios: a) mappings between WordNet and VerbNet lexicons; b)mappings between WordNet and FrameNet lexicons; and c) mappings between WordNet and PropBank lexicons. On the one hand, the lexical mappings from WordNet to FrameNet and from WordNet to VerbNet are obtained by applying graph-based WSD algorithms to semantically coherent groupings of verbal entries belonging to the same FrameNet frame or VerbNet class. On the other hand, for the lexical mappings from WordNet to PropBank, the WSD approach is applied to a corpus annotated with PropBank predicates (SRL at predicate level). We cross the annotations obtained by the WordNet-based WSD algorithm in a corpus annotated with PropBank predicates. All these strategies provide new mappings between the verbal entries in the resources and the WordNet senses. Consequently, we can connect predicates from different resources that are connected to the same WordNet sense.

Regarding the role mappings, we have proposed two approaches to infer new role mappings between two pairs of resources: a) VerbNet and FrameNet; and b) FrameNet and PropBank. First, we have defined a three-step method to increase the alignments between VerbNet thematic-roles and FrameNet frame-elements. This method exploits the current content of SemLink and the examples of use contained in VerbNet and the lexicographic annotations of FrameNet. Second, we also present a corpus-based approach to extend the mappings between FrameNet and PropBank. To obtain the role mappings between PropBank and FrameNet, our method acquires the most common correspondences between the annotations of both resources over the same sentences. We cross SRL corpus annotations.

Moreover, we have dealt in a simple way with the problem of multilingualism and the nominalization of the Predicate Matrix. Firstly, we have extended the predicate information to languages other than English, turning it into a multilingual resource. Specifically, we have integrated resources in Spanish, Catalan, and Basque. The extension to Spanish and Catalan has been made thanks to AnCora (Taulé et al., 2008b) corpus and the Spanish AnCora-Verb and Catalan AnCora-Verb verbal lexicons (Juan Aparicio and Martí, 2008) and the Basque Verb Index (BVI) (Estarrona et al., 2015) corpus-based lexicon is used in the case of Basque. As a result, the Predicate Matrix provides a multilingual lexicon to allow interoperable semantic analysis in multiple languages.

Secondly, the Predicate Matrix has been extended also to cover English and Spanish nominal predicates by adding mappings to NomBank (Meyers et al., 2004) which contains nominalizations of the PropBank predicates and to Spanish Ancora-Nom.

In addition, the Predicate Matrix has been enriched with knowledge coming from additional semantic resources that use WordNet as a backbone. Specifically, has been added the knowledge associated with the sense of WordNet in the MCR, such as the Adimen-SUMO (Álvez et al., 2012) and the WordNet domain aligned to WordNet 3.0 (Gonzalez-Agirre et al., 2012b) features as well as the Base Level Concept (Izquierdo et al., 2007) of the WordNet sense. Also, each line of the Predicate Matrix also includes the frequency and the number of relations of the WordNet word sense.

The Predicate Matrix is publicly available[1]

In summary, this research work presents a novel approach to improve the interoperability between various semantic resources that incorporate predicate information. Our proposal defines a set of automatic methods for mapping the semantic knowledge included in WordNet, VerbNet, PropBank, and FrameNet, which allows more complete semantic interoperability between them. As mentioned, this has resulted in a new lexical-semantic resource called Predicate Matrix.

[1] https://adimen.si.ehu.es/web/PredicateMatrix

As a proof of the applications it may have, in the following section are summarized the works in which the Predicate Matrix has been exploited.

8.2 Predicate Matrix in Action

The Predicate Matrix has been exploited in many other research works either to project the disambiguation of an event based on a particular resource to other predicative models such as in the *Pikes*[2] project, to identify that information in texts written in different languages are referring to the same event (Vossen et al., 2016), to exploit the extended mapping between two resources to transfer linguistic information from the first to the second (Laparra, 2015) or to build versions of the predicate Matrix in other languages through WordNet connections to predicate resources in those languages (Vossen et al., 2016). Also, the predicate Matrix has been integrated into linked resources that integrate many other linguistic resources (Gangemi et al., 2016). Additionally, the navigational interface of *Semantikos*[3] allows to browse comfortably through the information contained in the Predicate Matrix, as it has been integrated into it.

One of the earliest uses of the Predicate Matrix was in the project *XLike* (Padró et al., 2014), where the aim was to develop technology for enabling the extraction of language-independent knowledge from documents in multiple languages and genres. Specifically, they use the very first version of the Predicate Matrix to project the WordNet-based concepts they obtain to PropBank predicates and FrameNet diathesis structures. In this way, they also manage to normalise the semantic roles produced by the SRL, since the SRL they use produces treebank-dependent roles and these are not the same for all languages.

Laparra (2015) exploits semantic and ontological relations between predicates and semantic roles in FrameNet for Implicit Semantic Role Labelling based on PropBank/NomBank. The relations of FrameNet form a huge graph where the participants of different events are interconnected. These links express the implications among the roles and can facilitate inferencing when the participants that are part of different events are actually the same. But,

[2] http://pikes.fbk.eu/

[3] https://play.google.com/store/apps/details?id=org.sqlunet.browser

each semantic relation between FrameNet frames and their frame-elements has been defined strictly from a parent frame to their direct children. So, Laparra (2015) propose a set of rules for inferring new FrameNet relations among frames and frame-elements based on the descriptions included in its technical documentation and in this way extend the number of direct relations. Then, by means of different sets of mappings between FrameNet frame-elements and PropBank/NomBank arguments transfer the semantic relations from the first to the second. They evaluated the use of three different sets of mappings to project the frame-element relations acquired from FrameNet to PropBank/NomBank arguments and analysed empirically how the inclusion of this information affects an existing system for ISRL. One of the sets of mappings applied for the projection of the semantic relations from FrameNet to PropBank/NomBank was the one from the Predicate Matrix available in that moment. They also use SemLink and a set of mappings generated by them. The use of any of the three sets of mappings improved the result of the basic configuration of the ISRL system and interestingly, the Predicate Matrix offered quite comparable results with their best system.

In (Segers et al., 2016b) the Event and Implied Situation Ontology (ESO) (Segers et al., 2015b) is injected into the Predicate Matrix and demonstrated how these resources are used to detect information in large sets of documents that otherwise would have remained implicit.

In other dissertation about applicating concept-based and relation-based corpus in digital humanities (Fabo, 2017) also uses *Ixa-pipes-srl* which in turn has the Predicate Matrix integrated.

In (Rospocher et al., 2016) is introduced a system that automatically builds event-centric knowledge graphs (ECKG) from news articles. They use a cross-lingual framework based on different modular pipelines for different languages. These pipelines integrate modules for basic NLP processing as well as more advanced tasks such as cross-lingual named entity linking, semantic role labeling, and time normalization. Thus, the modular event extractor system allows for the interpretation of events, participants, locations, and time, as well as the relations between them for different languages. Then, the output of each individual pipeline is intended to be used as input for a system that obtains event-centric knowledge graphs. Their semantic role labeler annotates the events with the PropBank concepts. By using the Predicate Matrix the SRL module can add many more classes that are available in the Predicate Matrix. The enrichment with concepts from the Predicate Matrix provides semantic interoperability across different predicate models but also across different languages. These ECKGs are used in the NewsReader project (Vossen et al., 2016) in which the Predicate Matrix is part of its multilingual event detection system. Again, the NewsReader's pipelines, composed of different NLP tasks modules, guarantee interoperability across language and predicate resources by integrating the Predicate Matrix within the SRL modules. The Event and Situation Ontology (Segers et al., 2015b) was also developed in the background of the Newsreader project, which is designed to formalize implications of events, in other words, the pre and post conditions of events and the roles of the entities affected by an event. These event implications are mapped to Wordnet, SUMO, and Framenet. Then, in order to interoperate with PropBank the Predicate Matrix is used.

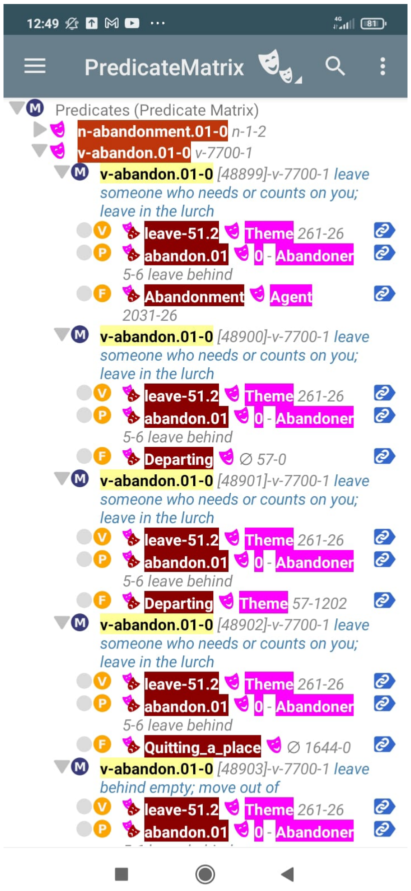
In (Laparra et al., 2017) presented an approach to extract ordered timelines of events, their participants, locations, and times from a set of multilingual and cross-lingual data sources for which they effectively leverage several multilingual resources such as the Predicate Matrix and DBpedia to improve the performance of building cross-lingual timelines in a setting where no parallel data is available as input. They make use of the Predicate Matrix in order to obtain interoperability across languages and semantic role labeling annotations. The event representation provided by their SRL systems is based on PropBank, for English, and AnCora for Spanish. As the Predicate Matrix gathers knowledge bases that contain predicate and semantic role information in different languages, including links between PropBank and AnCora, they can exploit these mappings for establishing, for example, that the role *arg0* of the Spanish predicate *vender.1* is aligned to the role *A0* of the PropBank predicate *sell.01*.

In the *CRF4TimeML* end-to-end Temporal Processing system (Caselli and Morante, 2018) was added lexical semantic information by using not only WordNet synsets, but also VerbNet classes and FrameNet frames, obtained from the alignments in the Predicate Matrix.

In the thesis on Semantic Role Labeling for Basque (Izko et al., 2017) the Predicate Matrix has been used to carry out the disambiguation of predicate senses, in this case, by exploiting the mapping between WordNet and PropBank.

In addition, the Predicate Matrix has been integrated into a freely available application called *Semantikos* released by Bernard Bou. *Semantikos* is an application based on SqlUNet that unifies WordNet, VerbNet, Prop-

Bank, FrameNet, and Predicate Matrix through SQL into a single relational database. It offers a friendly navigation interface to browse through English word senses and their semantic roles. It obviously brings benefits such as the possibility of using SQL language to build sophisticated queries. As mentioned, it permits exploring the Predicate Matrix's semantic role alignments in a user-friendly interface. For instance, Figure 8.1 shows the results provided by the *Semantikos* application for the search of the English verb *abandon*.



Predicates (Predicate Matrix)
n-abandonment.01-0 n-1-2
v-abandon.01-0 v-7700-1
v-abandon.01-0 [48899]-v-7700-1 leave someone who needs or counts on you; leave in the lurch
V leave-51.2 Theme 261-26
P abandon.01 0 - Abandoner 5-6 leave behind
F Abandonment Agent 2031-26
v-abandon.01-0 [48900]-v-7700-1 leave someone who needs or counts on you; leave in the lurch
V leave-51.2 Theme 261-26
P abandon.01 0 - Abandoner 5-6 leave behind
F Departing ∅ 57-0
v-abandon.01-0 [48901]-v-7700-1 leave someone who needs or counts on you; leave in the lurch
V leave-51.2 Theme 261-26
P abandon.01 0 - Abandoner 5-6 leave behind
F Departing Theme 57-1202
v-abandon.01-0 [48902]-v-7700-1 leave someone who needs or counts on you; leave in the lurch
V leave-51.2 Theme 261-26
P abandon.01 0 - Abandoner 5-6 leave behind
F Quitting_a_place ∅ 1644-0
v-abandon.01-0 [48903]-v-7700-1 leave behind empty; move out of
V leave-51.2 Theme 261-26
P abandon.01 0 - Abandoner

Figure 8.1: Example of exploring the Predicate Matrix via Semantikos.

Moreover, the Predicate Matrix is exposed in RDF in PreMOn (Corcoglioniti et al., 2016b), and the same authors exploit it in their work focused on populating FrameBase, the FrameNet's frame-based Semantic Web ontology (Corcoglioniti et al., 2016a).

The Predicate Matrix also was integrated into the Framester project[4]. Framester (Gangemi et al., 2016) is a large FrameNet frame-based ontological RDF knowledge graph that links together a wealth of linguistic resources, among which is the Predicate Matrix. In order to add the Predicate Matrix to the Framester linked data cloud a RDF version of the Predicate Matrix was created. In turn, Alam et al. (2021) presents the *TakeFive* SRL algorithm that leverages Framester. Consequently, they use the Predicate Matrix through the Framester.

The tool PIKES[5] also exploits the Predicate Matrix to get the disambiguation of events from the resources integrated in it. PIKES processes documents in order to extract different type of knowledge and represent it in graphs. Different instances are identified such as time expressions, locations, or events. The events are disambiguated with respect to different linguistic resources such as FrameNet, VerbNet, and PropBank by exploiting the Predicate Matrix.

The predicate matrix has also made it possible to evaluate the work of others. In a study about the translatability of VerbNet classes from English to typologically diverse languages by means of a particular manual methodology proposed by Majewska et al. (2018), in order to verify they obtain accurate enough gold standard classes, they compare the results obtained using the presented manual method with those potentially obtainable (semi-)automatically by using the mappings between WordNet senses and VerbNet in the Predicate Matrix. They chose Mandarin as the test language and use the Predicate Matrix in order to obtain candidate verbs for all of the 17 English VerbNet classes used in the study. Starting from pairings of English verbs and VerbNet classes, they looked up corresponding WordNet synsets in the Predicate Matrix, and subsequently used the links between Princeton WordNet and the Chinese Open WordNet to obtain Mandarin candidate verbs. They conclude that, although the output of the automatic method is noisy and it misses out over half of the gold standard candidates identified manually, the automatic method using the Predicate Matrix picked up 46% of the candidate verbs manually identified by the Mandarin translator and suggest that the Predicate Matrix can serve as a useful auxiliary tool. According to the feedback provided by the evaluators, the verbs identified by both methods were particularly good candidates for each class. The Predicate

[4] `http://etna.istc.cnr.it/framester_web/`

[5] `http://pikes.fbk.eu/`

Matrix could therefore be used to identify prototypical class members within the manually obtained sets of translations that would carry more weight in machine evaluation.

Popov and Sikos (2019) explore how to exploit structural information from WordNet and FrameNet in the frame identification task based on neural networks using graph embeddings. In order to extend and make denser the FrameNet-based graph embedding they map through the Predicate Matrix FrameNet predicates to WordNet synsets, so they can have access to the dense semantic network of WordNet and incorporate the lexico-semantic relations and other relations expressing relatedness into their graph embedding. The same author uses the Predicate Matrix in other work (Popov et al., 2019) about knowledge-based word sense disambiguation. They modify a knowledge base created originally on the basis of WordNet by enriching it with the addition of new relations to it. In particular, they extract new relations from VerbNet and FrameNet for which they have first made use of the Predicate Matrix to obtain most of the cross-mappings between WordNet and VerbNet and FrameNet.

(Gruzitis et al., 2018, 2020) convert semantic roles from FrameNet to PropBank via Predicate Matrix in the creation of a Multilayer Latvian Corpus for NLU.

Recently, Aceta et al. (2021) has used the Predicate Matrix lexicon to enrich the ontology of its dialogue system.

8.3 Publications

Below, we present chronologically the list of publications related with the research described in this document:

- López de Lacalle M., Laparra E. and Rigau G. *First steps towards a predicate matrix.* 7th Global Wordnet Conference (GWC'07). Tartu, Estonia. 2014.

  The contributions of the previous publication are described in Chapter 4.

- López de Lacalle M., Laparra E. and Rigau G. *Predicate Matrix: extending SemLink through WordNet mappings.* 9th International Conference on Language Resources and Evaluation. Reykjavik, Iceland. 2014.

  This study and its corresponding experiments are presented in Chapter 5.

- López de Lacalle M.,Laparra E., Aldabe I. and Rigau G. *Predicate Matrix: Automatically extending the interoperability between predicative resources.* Language Resources and Evaluation. 2016.

  This publication contains part of the contributions presented in Chapter 5.

- López de Lacalle M.,Laparra E., Aldabe I. and Rigau G. *A multilingual predicate matrix.* 10th International Conference on Language Resources and Evaluation. Portoroz, Eslovenia. 2016.

  This publication contains part of the contributions presented in Chapter 6.

The following references are not covered but are very closely related to this thesis:

- Agerri R., Agirre E., Aldabe I., Altuna B., Beloki Z., Laparra E., López de Lacalle M., Rigau G., Soroa A., Urizar R. *The NewsReader project.* Proceedings of the 30th Annual Meeting of Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN'14). Girona, Spain. Procesamiento del Lenguaje Natural. Vol. 53 pp. pp 215-218. ISSN: 1135-5948. 2014.

8.4 Future work

Although we have taken Semlink as a starting point for the construction of the Predicate Matrix and therefore we mainly have focused our research on further extending the mapping coverage among the different predicate schemas integrated in it, we are aware of the many interesting resources including predicate information available, such as those mentioned in Chapter 2, that it might be worth be integrated as additional resources to the

Predicate Matrix. By linking some of the existing resources in the literature, interesting benefits could be obtained. For example, by joining the Predicate Matrix and VerbAtlas, the Predicate Matrix could be used to enrich VerbAtlas with information from FrameNet, or, in turn, VerbAtlas could be used to increase the coverage of the Predicate Matrix and fix the incorrect automatic mappings.

The techniques used in this research for the automatic mapping of predicates and their semantic roles between different resources are prior to the advent of Deep Learning-based techniques, and, some of them are quite basic and straightforward. Although it has been proven in this thesis that with Pre Deep Learning Era techniques we obtain successful results, there is still great room for improvement and new deep neural-network offer new ways of addressing NLP tasks, including the automatic linking of resources. For example, by means of sentence embedding methods such as the one provided by sentence-BERT(SBERT) (Reimers and Gurevych, 2019) we could create meaningful text embeddings for representing the predicates of different resources making use of different textual information such as glosses, descriptions or examples of use provided by the resources themselves, and then apply Semantic Textual Similarity (STS) techniques for comparison of the embeddings for disambiguation (e.g using cosine-similarity).

Another interesting work we may carry out in the future is representing the Predicate Matrix as Knowledge Graph (KG). With TransE (Bordes et al., 2013) type systems, the knowledge behind the mappings between the different predicate-argument models that the Predicate Matrix is composed of could be encoded in structured data models such as knowledge graphs or semantic networks. We could define the Predicate Matrix knowledge graph as a graph where nodes are the specific predicates and roles of the different resources included in the Predicate Matrix and each edge is a relation between the predicates and roles. This would include semantic relations between predicates within the same resource and the relations or mappings of predicates between different resources.

For example, as shown in Figure 8.2 *Giving#give.v* would represent the node of the lexical unit or predicate *give* of the frame **Giving** in FrameNet for which two types of relations have been described. On the one hand, the semantic relation of *inheritace* it maintains with the predicate **sell** of the frame **Commerce_Sell** in FrameNet itself, and on the other hand, the relation that reflects the mapping with the same predicate in VerbNet (called *fn-mapping* in the image). Specifically, the node *13.1-1#give* represents the verb **give** of the verb class ***give-13.1-1*** of VerbNet. In the same way that the semantic relations between the predicates of the FrameNet hierarchy are represented, for verbs belonging to the same VerbNet class the relation textitclass will be defined, as in the case of the verbs **give** and **sell**, the nodes *13.1-1#give* and *13.1-1#sell*, of the VerbNet verb class ***give-13.1-1*** .

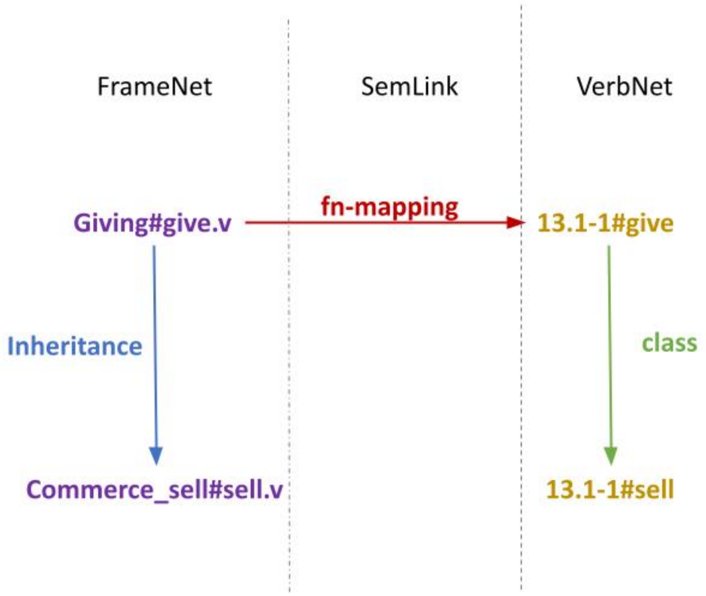


Figure 8.2: Example of mappings in the Predicate Matrix represented as a Knowledge Graph.

Models such as TransE learn a vector of relations (Rn) for each entity and each relation such that it is satisfied and this, in turn, allows us to infer new relations or mappings and get a more complete knowledge base. In other words, it would allow us to infer new mappings based on the Predicate Matrix knowledge graph. Following the example in Figure 8.2, the relations already established in the knowledge graph would lead us to infer a new relation of the type *fn-mapping* between the predicate **sell** of the frame **Commerce_Sell** in FrameNet, the node *Commerce_sell#sell.v*, and the verb **sell** of the VerbNet verb class ***give-13.1-1*** , the node *13.1-1#sell*.

Finally, we consider that the mapping between PropBank and FrameNet can be exploited to expand the QA-SRL dataset generated in (Pyatkin et al., 2021), where questions are based on PropBank, by creating more complex questions that internalize FrameNet knowledge (e.g by means of its structural information between roles) and then evaluate models such as T0 (Sanh et al., 2021) or T5 (Raffel et al., 2020a), checking if they are able to answer that kind of questions. Question-answer driven semantic role labeling (QA-SRL) (He et al., 2015) was proposed as a natural, easily attainable formulation of SRL. QA-SRL labels each predicate-argument relation with a question-answer pair, where natural language questions represent semantic roles (e.g. arg_0, arg_2, $ArgM - TMP$ in PropBank), and answers correspond to arguments of the predicate in a specific sentence. In the sentence *"Thomas has proved that God exists"*, the semantic role arg_0 of PropBank is labeled with *Who has proved something? - Thomas* question-answer pair. Originally, in (He et al., 2015) the questions for QA-SRL were generated following general templates that were not based on any resource, in a more recent work Pyatkin et al. (2021) base on PropBank to know the roles of each predicate and create a question for each one. For instance, they generate 6 role questions for the predicate *arrive* in the sentence *"The plane took off in Los Angeles. The tourists will arrive in Mexico at noon."*. Some questions are for explicit arguments: *entity in motion- "Who will arrive in Mexico?"*, *end point- "Where will the tourists arrive?"*, *temporal- "When will the tourists arrive?"*. Some other questions are for implicit arguments: *start point- "Where will the tourists arrive from?"*, *manner- "How will the tourists arrive?"*. Finally, some questions are for arguments that do not appear at all *cause- "Why will the tourists arrive?"*. In this work, use the dataset (https://github.com/uwnlp/qasrl-bank) to fine-tune a BART (Lewis et al., 2019) neural model and obtain an SRL model that is based on answering questions for each predicate. In addition to this, it would be interesting to test whether a T0/T5-like model, obviously without fine-tuning it in this dataset, is able to answer these questions. If it does well, it means that the model has learned the knowledge described in PropBank. In addition, this dataset could be extended through SemLink or the Predicate Matrix to include questions derived from the knowledge described in VerbNet or FrameNet. Thus, in addition to the questions designed to elicit the roles of the predicate *arrive* in the context described above, we could also generate the question *"Where did the tourist departed from?"* because the frame *Arriving* is related to *Departing* with the relation *Is_Preceded_by* in FrameNet. If T0/T5 model does well in these questions, it means that it has also learned this kind of knowledge, and, if not, then we would have a dataset to fine-tune and obtain a quite interesting model.

Bibliography

Abend, O., Reichart, R., and Rappoport, A. (2008). A supervised algorithm for verb disambiguation into verbnet classes. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 9–16.

Abzianidze, L., Bjerva, J., Evang, K., Haagsma, H., Van Noord, R., Ludmann, P., Nguyen, D.-D., and Bos, J. (2017). The parallel meaning bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. *arXiv preprint arXiv:1702.03964*.

Aceta, C., Fernández, I., and Soroa, A. (2021). Ontology population reusing resources for dialogue intent detection: Generic and multilingual approach. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 10–18.

Agirre, E. and Edmonds, P. (2007). *Word sense disambiguation: Algorithms and applications*, volume 33. Springer Science & Business Media.

Agirre, E., Lopez De Lacalle, O., and Soroa, A. (2009). Knowledge-based wsd on specific domains: performing better than generic supervised wsd. In *Twenty-First International Joint Conference on Artificial Intelligence*.

Agirre, E., Otegi, A., and Rigau, G. (2008). Ixa at clef 2008 robust-wsd task: Using word sense disambiguation for (cross lingual) information retrieval. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 118–125. Springer.

Agirre, E. and Soroa, A. (2009a). Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 33–41.

Agirre, E. and Soroa, A. (2009b). Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (EACL-2009)*, Athens, Greece. Eurpean Association for Computational Linguistics.

Akbik, A. and Li, Y. (2016). Polyglot: Multilingual semantic role labeling with unified labels. In *Proceedings of ACL-2016 System Demonstrations*, pages 1–6.

Alam, M., Gangemi, A., Presutti, V., and Reforgiato Recupero, D. (2021). Semantic role labeling for knowledge graph extraction from text. *Progress in Artificial Intelligence*, 10(3):309–320.

Allen, J., An, H., Bose, R., de Beaumont, W., and Teng, C. M. (2020). A broad-coverage deep semantic lexicon for verbs. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3243–3251.

Allen, J., An, H., Bose, R., de Beaumont, W., and Teng, C. M. (2022). Collie: a broad-coverage ontology and lexicon of verbs in english. *Language Resources and Evaluation*, pages 1–30.

Allen, J. and Teng, C. M. (2018). Putting semantics into semantic roles. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 235–244.

Allen, J. F. and Teng, C. M. (2017). Broad coverage, domain-generic deep semantic parsing. In *2017 AAAI Spring Symposium Series*.

Alonso, L., Capilla, J. A., Castellón, I., Fernández-Montraveta, A., and Vázquez, G. (2007). The sensem project: Syntactico-semantic annotation of sentences in spanish. *Amsterdam studies in the theory and history of linguistic science series 4*, 292:89.

Alonso Alemany, L., Castellón Masalles, I., Laparra Martín, E., and Riga Claramunt, G. (2009). Evaluación de métodos semi-automáticos para la conexión entre framenet y sensem. Sociedad Española para el Procesamiento del Lenguaje Natural.

Álvez, J., Atserias, J., Carrera, J., Climent, S., Oliver, A., and Rigau, G. (2008). Consistent annotation of eurowordnet with the top concept ontology. In *Proceedings of Fourth International WordNet Conference (GWC'08)*.

Álvez, J., Lucio, P., and Rigau, G. (2012). Adimen-sumo: Reengineering an ontology for first-order reasoning. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 8(4):80–116.

Atserias, J., Villarejo, L., Rigau, G., Agirre, E., Carroll, J., Magnini, B., and Vossen, P. (2004). The meaning multilingual central repository. In *2nd International Global Wordnet Conference, January 20-23, 2004: proceedings*, pages 23–30. Masaryk University.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998a). The berkeley framenet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, ACL '98, pages 86–90, Montreal, Quebec, Canada.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998b). The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, ACL '98, pages 86–90, Montreal, Quebec, Canada.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N. (2013). Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186.

Basile, V., Bos, J., Evang, K., and Venhuizen, N. (2012). Developing a large semantically annotated corpus. In *Eighth International Conference on Language Resources and Evaluation*, pages 3196–3200. EUROPEAN LANGUAGE RESOURCES ASSOC-ELRA.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. (2009). Dbpedia-a crystallization point for the web of data. *Journal of web semantics*, 7(3):154–165.

Björkelund, A. and Hafdell, L. (2009). *High-performance multilingual semantic role labeling.* MSc thesis, Lund University.

Boas, H. C. (2002). Bilingual framenet dictionaries for machine translation. In *LREC*.

Boas, H. C. (2005). Semantic frames as interlingual representations for multilingual lexical databases. *International Journal of Lexicography*, 18(4):445–478.

Bohnet, B. (2010). Very high accuracy and fast dependency parsing is not a contradiction. In *The 23rd International Conference on Computational Linguistics (COLING 2010)*, Beijing, China.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

Bos, J., Basile, V., Evang, K., Venhuizen, N. J., and Bjerva, J. (2017). The groningen meaning bank. In *Handbook of linguistic annotation*, pages 463–496. Springer.

Brown, S. W., Dligach, D., and Palmer, M. (2014). Verbnet class assignment as a wsd task. In *Computing Meaning*, pages 203–216. Springer.

Burchardt, A., Erk, K., and Frank, A. (2005). A wordnet detour to framenet. *Sprachtechnologie, mobile Kommunikation und linguistische Resourcen*, 8:408–421.

Burchardt, A., Erk, K., Frank, A., Kowalski, A., Padó, S., and Pinkal, M. (2006). The salsa corpus: a german corpus resource for lexical semantics. In *LREC*, pages 969–974.

Burchardt, A., Pennacchiotti, M., Thater, S., and Pinkal, M. (2009). Assessing the impact of frame semantics on textual entailment. *Natural Language Engineering*, 15(4):527.

Busso, L. and Lenci, A. (2016). Italian verbnet: A construction-based approach to italian verb classification. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2633–2642.

Carreras, X. and Màrquez, L. (2004). Introduction to the CoNLL-2004 Shared Task: Semantic Role Labeling. In *Proceedings of the Eigth Conference on Computational Natural Language Learning (CoNLL-2004)*, pages 89–97, Boston, MA, USA.

Carreras, X. and Màrquez, L. (2005). Introduction to the conll-2005 shared task: Semantic role labeling. In *Proceedings of the 9th Conference on Computational Natural Language Learning*, CoNLL '05, pages 152–164, Ann Arbor, Michigan, USA.

Caselli, T. and Morante, R. (2018). Agreements and disagreements in temporal processing: An extensive error analysis of the tempeval-3 systems. In *Proceedings of Language Resources and Evaluation Conference*.

Chen, D., Schneider, N., Das, D., and Smith, N. A. (2010). Semafor: Frame argument resolution with log-linear models. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 264–267, Uppsala, Sweden. Association for Computational Linguistics.

Chen, Y.-N., Wang, W. Y., and Rudnicky, A. I. (2013). Unsupervised induction and filling of semantic slots for spoken dialogue systems using frame-semantic parsing. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 120–125. IEEE.

Clark, P., Dalvi, B., and Tandon, N. (2018). What happened? leveraging verbnet to predict the effects of actions in procedural text. *arXiv preprint arXiv:1804.05435*.

Conia, S., Bacciu, A., and Navigli, R. (2021). Unifying cross-lingual semantic role labeling with heterogeneous linguistic resources. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 338–351.

Corcoglioniti, F., Rospocher, M., and Aprosio, A. P. (2016a). Frame-based ontology population with pikes. *IEEE Transactions on Knowledge and Data Engineering*, 28(12):3261–3275.

Corcoglioniti, F., Rospocher, M., Aprosio, A. P., and Tonelli, S. (2016b). Premon: a lemon extension for exposing predicate models as linked data. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 877–884.

Cuadros, M. and Rigau, G. (2008). Knownet: Building a large net of knowledge from the web. In *22nd International Conference on Computational Linguistics (COLING'08)*, Manchester, UK.

Cuadros Oller, M. and Rigau Claramunt, G. (2008). Knownet: Building a large net of knowledge from the web. In *22nd International Conference on Computational Linguistics*, pages 1–8.

Dang, H. T. (2004). *Investigations into the role of lexical semantics in word sense disambiguation*. University of Pennsylvania.

Das, D., Chen, D., Martins, A. F., Schneider, N., and Smith, N. A. (2014). Frame-semantic parsing. *Computational linguistics*, 40(1):9–56.

Daza, A. and Frank, A. (2020). X-SRL: A parallel cross-lingual semantic role labeling dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3904–3914, Online. Association for Computational Linguistics.

DCL-IBL, B. (2019). Enhancing conceptual description through resource linking and exploration of semantic relations. In *Wordnet Conference*, page 280.

De Cao, D., Croce, D., Pennacchiotti, M., and Basili, R. (2008). Combining word sense and usage for modeling frame semantics. In *Semantics in Text Processing. STEP 2008 Conference Proceedings*, pages 85–101.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Di Fabio, A., Conia, S., and Navigli, R. (2019a). Verbatlas: a novel large-scale verbal semantic resource and its application to semantic role labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 627–637.

Di Fabio, A., Conia, S., and Navigli, R. (2019b). Verbatlas: a novel large-scale verbal semantic resource and its application to semantic role labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 627–637.

Duran, M. S. and Aluísio, S. M. (2012). Propbank-br: a brazilian treebank annotated with semantic role labels. In *LREC*, pages 1862–1867.

Elhadad, M. K., Badran, K. M., and Salama, G. I. (2017). A novel approach for ontology-based dimensionality reduction for web text document classification. *International Journal of Software Innovation (IJSI)*, 5(4):44–58.

Erk, K. and Pado, S. (2006). Shalmaneser - a flexible toolbox for semantic role assignment. In *Proceedings of LREC*, volume 6, page 73.

Estarrona, A., Aldezabal, I., and Díaz de Ilarraza, A. (2015). A methodology for the semiautomatic annotation of epec-rolsem, a basque corpus labeled at predicate level following the propbank-verbnet model. *Digital Scholarship in the Humanities*, pages 1–23.

Exner, P. and Nugues, P. (2011). Using semantic role labeling to extract events from wikipedia. In *DeRiVE@ ISWC*, pages 38–47.

Fabo, P. R. (2017). *Concept-based and relation-based corpus navigation: applications of natural language processing in digital humanities.* PhD thesis, Université Paris sciences et lettres.

Falk, I., Gardent, C., and Lamirel, J.-C. (2012). Classifying french verbs using french and english lexical resources. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 854–863.

Fellbaum, C., editor (1998a). *WordNet. An Electronic Lexical Database.* The MIT Press.

Fellbaum, C. (1998b). *WordNet: an electronic lexical database.* MIT Press.

Ferrández, O., Ellsworth, M., Munoz, R., and Baker, C. F. (2010). Aligning framenet and wordnet based on semantic neighborhoods. In *LREC*, volume 10, pages 310–314.

Fillmore, C. J. (1967). The case for case. In *Universals in Linguistic Theory*. ERIC.

Fillmore, C. J. (1976a). Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, volume 280, pages 20–32, New York.

Fillmore, C. J. (1976b). Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, volume 280, pages 20–32, New York.

Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M., and Soria, C. (2006). Lexical markup framework (lmf).

Gangemi, A., Alam, M., Asprino, L., Presutti, V., and Recupero, D. R. (2016). Framester: A wide coverage linguistic linked data hub. In *European knowledge acquisition workshop*, pages 239–254. Springer.

Gerber, M. and Chai, J. (2008). Class-based nominal semantic role labeling: a preliminary investigation.

Gilardi, L. and Baker, C. (2018). Learning to align across languages: Toward multilingual framenet. In Torrent, T. T., Borin, L., and Baker, C. F., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Gildea, D. and Jurafsky, D. (2000). Automatic labeling of semantic roles. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL ’00, pages 512–520, Hong Kong.

Gildea, D. and Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Giuglea, A.-M. and Moschitti, A. (2006). Semantic role labeling via framenet, verbnet and propbank. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 929–936.

Gliozzo, A., Giuliano, C., and Strapparava, C. (2005). Domain kernels for word sense disambiguation. In *Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL'05)*, pages 403–410.

Gonzalez-Agirre, A., Laparra, E., and Rigau, G. (2012a). Multilingual central repository version 3.0. In *LREC*, pages 2525–2529.

Gonzalez-Agirre, A., Laparra, E., and Rigau, G. (2012b). Multilingual central repository version 3.0. In *LREC*, pages 2525–2529.

González-Agirre, A., Rigau, G., and Castillo, M. (2012). A graph-based method to improve wordnet domains. In *CICLING*, pages 17–28. Springer.

Gonzalo, J., Verdejo, F., Chugur, I., and Cigarran, J. (1998). Indexing with wordnet synsets can improve text retrieval. *arXiv preprint cmp-lg/9808002*.

Griesel, M., Bosch, S., and Mojapelo, M. L. (2019). Thinking globally, acting locally–progress in the african wordnet project. In *Wordnet Conference*, page 191.

Gruber, J. S. (1965). *Studies in lexical relations*. PhD thesis, Massachusetts Institute of Technology.

Gruzitis, N., Darĝis, R., Rituma, L., Nešpore-Bērzkalne, G., and Saulīte, B. (2020). Deriving a propbank corpus from parallel framenet and ud corpora. In *Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet*, pages 63–69.

Gruzitis, N., Pretkalniņa, L., Saulīte, B., Rituma, L., Nešpore, G., Znotins, A., and Paikens, P. (2018). Creation of a balanced state-of-the-art multilayer corpus for nlu. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Gurevych, I., Eckle-Kohler, J., Hartmann, S., Matuschek, M., Meyer, C. M., and Wirth, C. (2012). Uby - a large-scale unified lexical-semantic resource based on lmf. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 580–590.

Hajič, J., Ciaramita, M., Johansson, R., Kawahara, D., Martí, M. A., Màrquez, L., Meyers, A., Nivre, J., Padó, S., Štěpánek, J., Straňák, P., Surdeanu, M., Xue, N., and Zhang, Y. (2009). The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, CoNLL '09, pages 1–18, Boulder, Colorado, USA.

Hartmann, S. (2017). *Knowledge-based Supervision for Domain-adaptive Semantic Role Labeling*. PhD thesis, Technische Universität.

Hautli, A., King, T. H., and Ramchand, G. (2015). Encoding event structure in urdu/hindi verbnet. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 25–33.

Haverinen, K., Nyblom, J., Viljanen, T., Laippala, V., Kohonen, S., Missilä, A., Ojala, S., Salakoski, T., and Ginter, F. (2014). Building the essential resources for finnish: the turku dependency treebank. *Language Resources and Evaluation*, 48(3):493–531.

He, L., Lee, K., Lewis, M., and Zettlemoyer, L. (2017). Deep semantic role labeling: What works and what's next. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

He, L., Lewis, M., and Zettlemoyer, L. (2015). Question-answer driven semantic role labeling: Using natural language to annotate natural language. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 643–653.

Henrich, V., Hinrichs, E., and Vodolazova, T. (2014). Aligning germanet senses with wiktionary sense definitions. In *Language and Technology Conference*, pages 329–342. Springer.

Hensman, S. and Dunnion, J. (2004). Automatically building conceptual graphs using verbnet and wordnet. In *Proceedings of the 2004 international symposium on Information and communication technologies*, pages 115–120.

Hoffart, J., Suchanek, F. M., Berberich, K., and Weikum, G. (2013). Yago2: A spatially and temporally enhanced knowledge base from wikipedia. *Artificial Intelligence*, 194:28–61.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60.

Huang, L., Cassidy, T., Feng, X., Ji, H., Voss, C., Han, J., and Sil, A. (2016). Liberal event extraction and event schema induction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 258–268.

Im, S. (2013). *The generator of the event structure lexicon (GESL): automatic annotation of event structure for textual inference tasks*. Brandeis University.

Im, S. (2019). Semi-automatic annotation of event structure, argument structure, and opposition structure to wordnet by using event structure frame. In *Global Wordnet Conference*, page 272.

Im, S. and Pustejovsky, J. (2009). Annotating event implicatures for textual inference tasks. In *The 5th Conference on Generative Approaches to the Lexicon*.

Im, S. and Pustejovsky, J. (2010). Annotating lexically entailed subevents for textual inference tasks. In *Twenty-third international flairs conference*.

Izko, H. S., Uriarte, O. A., and Sierra, B. Z. (2017). *Rol semantikoen etiketatzeak testuetako espaziodenbora informazioaren prozesamenduan daukan eraginaz*. PhD thesis, PhD thesis, Universidad del País Vasco/Euskal Herriko Unibertsitatea . . . .

Izquierdo, R., Suárez, A., and Rigau, G. (2007). Exploring the automatic selection of basic level concepts. In *Proceedings of RANLP*, volume 7. Citeseer.

Jezek, E., Magnini, B., Feltracco, A., Bianchini, A., and Popescu, O. (2014). T-pas: A resource of corpus-derived typed predicate-argument structures for linguistic analysis and semantic processing. In *Proceedings of LREC*, pages 890–895.

Jiang, Z. and Ng, H. (2006). Semantic role labeling of nombank: A maximum entropy approach. pages 138–145.

Johansson, R. and Nugues, P. (2007a). Lth: semantic structure extraction using nonprojective dependency trees. In *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)*, pages 227–230.

Johansson, R. and Nugues, P. (2007b). Using WordNet to extend FrameNet coverage. In *Proceedings of the Workshop on Building Frame-semantic Resources for Scandinavian and Baltic Languages, at NODALIDA*, Tartu, Estonia.

Juan Aparicio, M. T. and Martí, M. A. (2008). Ancora-verb: A lexical resource for the semantic annotation of corpora. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Kamtl, H. and Reyle, U. (1993). From discourse to logic: Introduction to model theoretic semantics of natural language, formal logic and drt.

Kipper, K. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon.* PhD thesis, University of Pennsylvania.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942.*

Laparra, E. (2015). Implicit semantic roles in discourse. Master's thesis, Euskal Herriko Unibertsitatea - Universidad del País Vasco (EHU-UPV).

Laparra, E., Agerri, R., Aldabe, I., and Rigau, G. (2017). Multi-lingual and cross-lingual timeline extraction. *Knowledge-Based Systems*, 133:77–89.

Laparra, E. and Rigau, G. (2009a). Integrating wordnet and framenet using a knowledge-based word sense disambiguation algorithm. In *Proceedings of the International Conference RANLP-2009*, pages 208–213.

Laparra, E. and Rigau, G. (2009b). Integrating wordnet and framenet using a knowledge-based word sense disambiguation algorithm. In *Proceedings of RANLP*, Borovets, Bulgaria.

Laparra, E. and Rigau, G. (2013). Sources of evidence for implicit argument resolution. In *Proceedings of the 10th International Conference on Computational Semantics*, IWCS '13, pages 155–166, Potsdam, Germany.

Laparra, E., Rigau, G., and Cuadros, M. (2010a). Exploring the integration of wordnet and framenet. In *Proceedings of the 5th Global WordNet Conference (GWC 2010), Mumbai, India.*

Laparra, E., Rigau, G., and Cuadros, M. (2010b). Exploring the integration of wordnet and framenet. In *Proceedings of the 5th Global WordNet Conference (GWC 2010), Mumbai, India.*

Leseva, S. and Stoyanova, I. (2019). Structural approach to enhancing wordnet with conceptual frame semantics. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 629–637.

Leseva, S., Stoyanova, I., Todorova, M., and Kukova, H. (2020). Putting pieces together: Predicate-argument relations and selectional preferences. *Towards a Semantic Network Enriched with a Variety of Semantic Relations*, page 49.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*, volume 348. University of Chicago press Chicago.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461.*

Li, Z., Zhao, H., Zhou, J., Parnow, K., and He, S. (2019). Dependency and span, cross-style semantic role labeling on propbank and nombank. *Transactions on Asian and Low-Resource Language Information Processing.*

Litkowski, K. (2004). Senseval-3 task: Automatic labeling of semantic roles. In *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 9–12, Barcelona, Spain.

Liu, C. and Ng, H. (2007). Learning predictive structures for semantic role labeling of nombank.

Liu, M. (2020). The construction and annotation of a semantically enriched database: the mandarin verbnet and its nlp applications. In *From Minimal Contrast to Meaning Construct*, pages 257–272. Springer.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Y., Scheuermann, P., Li, X., and Zhu, X. (2007). Using wordnet to disambiguate word senses for text classification. In *international conference on computational science*, pages 781–789. Springer.

Loper, E., Yi, S.-T., and Palmer, M. (2007). Combining lexical resources: mapping between propbank and verbnet. In *Proceedings of the 7th International Workshop on Computational Linguistics, Tilburg, the Netherlands*.

López de Lacalle, M., Laparra, E., Aldabe, I., and Rigau, G. (2016a). A multilingual predicate matrix. In Chair), N. C. C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

López de Lacalle, M., Laparra, E., Aldabe, I., and Rigau, G. (2016b). Predicate matrix: Automatically extending the interoperability between predicate resources. *Language Resources and Evaluation*.

López de Lacalle, M., Laparra, E., and Rigau, G. (2014a). First steps towards a predicate matrix. In *Proceedings of the 7th Global WordNet Conference (GWC2014)*, Tartu, Estonia.

López de Lacalle, M., Laparra, E., and Rigau, G. (2014b). Predicate matrix: extending semlink through wordnet mappings. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

Majewska, O., Vulić, I., McCarthy, D., Huang, Y., Murakami, A., Laippala, V., and Korhonen, A. (2018). Investigating the cross-lingual translatability of verbnet-style classification. *Language resources and evaluation*, 52(3):771–799.

Marcus, M. P., Marcin-kiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of English: the penn treebank. *Computational Linguistics*, 19:313–330.

Marcus, R., Palmer, M., Ramshaw, R. B. S. P. L., and Xue, N. (2011). Ontonotes: A large training corpus for enhanced processing. *Joseph Olive, Caitlin Christianson, andJohn McCary, editors, Handbook of Natural LanguageProcessing and Machine Translation: DARPA GlobalAutonomous Language Exploitation.*

Màrquez, L., Villarejo, L., Martí, M. A., and Taulé, M. (2007). Semeval-2007 task 09: Multilevel semantic annotation of catalan and spanish. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 42–47.

Maru, M., Scozzafava, F., Martelli, F., and Navigli, R. (2019). Syntagnet: Challenging supervised word sense disambiguation with lexical-semantic combinations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3534–3540.

Matuschek, M. and Gurevych, I. (2013). Dijkstra-wsa: A graph-based approach to word sense alignment. *Transactions of the Association for Computational Linguistics*, 1:151–164.

McDonald, R., Crammer, K., and Pereira, F. (2005). Online large-margin training of dependency parsers. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 91–98, Stroudsburg, PA, USA. Association for Computational Linguistics.

Meštrović, A. and Calì, A. (2016). An ontology-based approach to information retrieval. In *Semanitic Keyword-based Search on Structured Data Sources*, pages 150–156. Springer.

Meyer, C. M. and Gurevych, I. (2011). What psycholinguists know about chemistry: Aligning wiktionary and wordnet for increased domain coverage. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 883–892.

Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, B., and Grishman, R. (2004). The nombank project: An interim report. In *In Proceedings of the NAACL/HLT Workshop on Frontiers in Corpus Annotation.*

Mihalcea, R. and Moldovan, D. I. (2001). extended wordnet: Progress report. In *in Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*. Citeseer.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Miller, G. A., Leacock, C., Tengi, R., and Bunker, R. T. (1993). A semantic concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Miller, T. and Gurevych, I. (2014). Wordnet-wikipedia-wiktionary: Construction of a three-way alignment. In *LREC*, pages 2094–2100.

Moeller, S., Wagner, I., Palmer, M., Conger, K., and Myers, S. (2020). The russian propbank. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5995–6002.

Mohit, B. and Narayanan, S. (2003). Semantic extraction with wide-coverage lexical resources. In *Companion Volume of the Proceedings of HLT-NAACL 2003-Short Papers*, pages 64–66.

Monachesi, P., Stevens, G., and Trapman, J. (2007). Adding semantic role annotation to a corpus of written dutch. In *Proceedings of the Linguistic Annotation Workshop*, pages 77–84.

Moon, L., Christodoulopoulos, C., Fisher, C., Franco, S., and Roth, D. (2018). Gold standard annotations for preposition and verb sense with semantic role labels in adult-child interactions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3004–3014.

Mousselly-Sergieh, H. and Gurevych, I. (2016). Enriching wikidata with frame semantics. In *Proceedings of the 5th Workshop on Automated Knowledge Base Construction*, pages 29–34.

Mousser, J. (2011). Classifying arabic verbs using sibling classes. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*.

Mújdricza-Maydt, E., Hartmann, S., Gurevych, I., and Frank, A. (2016). Combining semantic annotation of word sense & semantic roles: A novel annotation scheme for verbnet roles on german language data. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3031–3038.

Navigli, R. (2018). Natural language understanding: Instructions for (present and future) use. In *IJCAI*, volume 18, pages 5697–5702.

Navigli, R. and Ponzetto, S. P. (2010). Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225.

Navigli, R. and Velardi, P. (2005). Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. *IEEE transactions on pattern analysis and machine intelligence*, 27(7):1075–1086.

Niemann, E. and Gurevych, I. (2011). The peoples ́ web meets linguistic knowledge: Automatic sense alignment of wikipedia and wordnet. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011).*

Niles, I. and Pease, A. (2001). Towards a standard upper ontology. In *Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001*, pages 2–9.

Ohara, K. H., Fujii, S., Ohori, T., Suzuki, R., Saito, H., and Ishizaki, S. (2004). The japanese framenet project: An introduction. In *Proceedings of LREC-04 Satellite Workshop "Building Lexical Resources from Semantically Annotated Corpora"(LREC 2004)*, pages 9–11.

Padó, S. and Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Padró, L., Agic, Z., Carreras, X., Fortuna, B., García Cuesta, E., Li, Z., Stajner, T., and Tadic, M. (2014). Language processing infrastructure in the xlike project. In *LREC 2014: Ninth International Conference on Language Resources and Evaluation: Reykjavik, Islàndia: May, 26-31, 2014: proceedings*, pages 3811–3816. European Language Resources Association (ELRA).

Paek, H., Kogan, Y., Thomas, P., Codish, S., and Krauthammer, M. (2006). Shallow semantic parsing of randomized controlled trial reports. In *AMIA*

Annual Symposium Proceedings, volume 2006, page 604. American Medical Informatics Association.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Palmer, M. (2009). Semlink: Linking propbank, verbnet and framenet. In *Proceedings of the Generative Lexicon Conference*, pages 9–15.

Palmer, M., Babko-Malaya, O., and Dang, H. T. (2004). Different sense granularities for different applications. In *Proceedings of the 2nd International Workshop on Scalable Natural Language Understanding (ScaNaLU 2004) at HLT-NAACL 2004*, pages 49–56.

Palmer, M., Bonial, C., and McCarthy, D. (2014). Semlink+: Framenet, verbnet and event ontologies. In *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929-2014)*, pages 13–17.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Pease, A., Niles, I., and Li, J. (2002). The suggested upper merged ontology: A large ontology for the semantic web and its applications. In *Working notes of the AAAI-2002 workshop on ontologies and the semantic web*, volume 28, pages 7–10.

Pennacchiotti, M., De Cao, D., Basili, R., Croce, D., and Roth, M. (2008). Automatic induction of framenet lexical units. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 457–465.

Philpot, A., Hovy, E., and Pantel, P. (2005). The omega ontology. In *Proceedings of OntoLex 2005-Ontologies and Lexical Resources.*

Pianta, E., Bentivogli, L., and Girardi, C. (2002). Multiwordnet: developing an aligned multilingual database. In *First international conference on global WordNet*, pages 293–302.

Pilehvar, M. T. and Navigli, R. (2014). A robust approach to aligning heterogeneous lexical resources. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 468–478.

Popov, A. and Sikos, J. (2019). Graph embeddings for frame identification. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 939–948.

Popov, A., Simov, K., and Osenova, P. (2019). Know your graph. state-of-the-art knowledge-based wsd. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 949–958.

Pradhan, S. S., Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2007). Ontonotes: A unified relational semantic representation. In *International Conference on Semantic Computing (ICSC 2007)*, pages 517–526. IEEE.

Pyatkin, V., Roit, P., Michael, J., Goldberg, Y., Tsarfaty, R., and Dagan, I. (2021). Asking it all: Generating contextualized questions for any semantic role. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1429–1441, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

QasemiZadeh, B., Petruck, M. R., Stodden, R., Kallmeyer, L., and Candito, M. (2019). Semeval-2019 task 2: Unsupervised lexical frame induction. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 16–30.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training. OpenAI.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020a). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J., et al. (2020b). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Rospocher, M., van Erp, M., Vossen, P., Fokkens, A., Aldabe, I., Rigau, G., Soroa, A., Ploeger, T., and Bogaard, T. (2016). Building event-centric knowledge graphs from news. *Journal of Web Semantics*, 37:132–151.

Ruiz-Casado, M., Alfonseca, E., and Castells, P. (2005). Automatic assignment of wikipedia encyclopedic entries to wordnet synsets. In *International Atlantic Web Intelligence Conference*, pages 380–386. Springer.

Ruppenhofer, J., Sporleder, C., Morante, R., Baker, C., and Palmer, M. (2010). Semeval-2010 task 10: Linking events and their participants in discourse. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 45–50, Los Angeles, California, USA.

Rutkowski, S., Rychlik, P., and Mykowiecka, A. (2019). Estimating senses with sets of lexically related words for polish word sense disambiguation. In *Wordnet Conference*, page 118.

Şahin, G. G. and Adalı, E. (2018). Annotation of semantic roles for the turkish proposition bank. *Language Resources and Evaluation*, 52(3):673–706.

Salaberri, H., Arregi, O., and Zapirain, B. (2015). brol: The parser of syntactic and semantic dependencies for basque. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 555–562.

Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Scao, T. L., Raja, A., Dey, M., Bari, M. S., Xu, C., Thakker, U., Sharma, S. S., Szczechla, E., Kim, T., Chhablani, G., Nayak, N., Datta, D., Chang, J., Jiang, M. T.-J., Wang, H., Manica, M., Shen, S., Yong, Z. X., Pandey, H., Bawden, R., Wang, T., Neeraj, T., Rozen, J., Sharma, A., Santilli, A., Fevry, T., Fries, J. A., Teehan, R., Biderman, S., Gao, L., Bers, T., Wolf, T., and Rush, A. M. (2021). Multitask prompted training enables zero-shot task generalization.

Scarton, C., Duran, M. S., and Aluísio, S. M. (2014). Using cross-linguistic knowledge to build verbnet-style lexicons: Results for a (brazilian) portuguese verbnet. In *International Conference on Computational Processing of the Portuguese Language*, pages 149–160. Springer.

Scott, S. and Matwin, S. (1998). Text classification using wordnet hypernyms. In *Usage of WordNet in Natural Language Processing Systems*.

Segers, R., Laparra, E., Rospocher, M., Vossen, P., Rigau, G., and Ilievski, F. (2016a). The predicate matrix and the event and implied situation ontology: Making more of events. *Proceedings of GWC2016*.

Segers, R., Laparra, E., Rospocher, M., Vossen, P., Rigau, G., and Ilievski, F. (2016b). The predicate matrix and the event and implied situation ontology: Making more of events. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 364–372.

Segers, R., Rospocher, M., Vossen, P., Laparra, E., Rigau, G., and Minard, A.-L. (2016c). The event and implied situation ontology (eso): Application and evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1463–1470.

Segers, R., Vossen, P., Rospocher, M., Serafini, L., Laparra, E., and Rigau, G. (2015a). Eso: A frame based ontology for events and implied situations. *Proceedings of MAPLEX*, 2015.

Segers, R., Vossen, P., Rospocher, M., Serafini, L., Laparra, E., and Rigau, G. (2015b). Eso: A frame based ontology for events and implied situations. In *Proceedings of MAPLEX 2015*, Yamagata, Japan.

Segond, F., Schiller, A., Grefenstette, G., and Chanod, J.-P. (1997). An experiment in semantic tagging using hidden markov model tagging. In *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*.

Shen, D. and Lapata, M. (2007). Using semantic roles to improve question answering. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 12–21.

Shi, L. and Mihalcea, R. (2005). Putting pieces together: Combining framenet, verbnet and wordnet for robust semantic parsing. In *International conference on intelligent text processing and computational linguistics*, pages 100–111. Springer.

Sinha, S. K. (2008). Answering questions about complex events. Technical report, CALIFORNIA UNIV BERKELEY DEPT OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE.

Sio, J. U.-S. and da Costa, L. M. (2019). Building the cantonese wordnet. In *Wordnet Conference*, page 206.

Subirats, C. and Sato, H. (2003). Surprise! spanish framenet. In *In Proceedings of the Workshop on Frame Semantics at the XVII. International Congress of Linguists*. Citeseer.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706.

Surdeanu, M., Johansson, R., Meyers, A., Màrquez, L., and Nivre, J. (2008). The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of the Twelfth Conference on Natural Language Learning*, CoNLL '08, pages 159–177, Manchester, United Kingdom.

Swier, R. S. and Stevenson, S. (2004). Unsupervised semantic role labellin. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 95–102.

Swift, M. (2005). Towards automatic verb acquisition from verbnet for spoken dialog processing. In *Proceedings of Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, pages 115–120.

Taulé, M., Martí, M. A., and Recasens, M. (2008a). Ancora: Multilevel annotated corpora for catalan and spanish. In *LREC*.

Taulé, M., Martí, M. A., and Recasens, M. (2008b). Ancora: Multilevel annotated corpora for catalan and spanish. In *LREC*.

Taulé, M., Peris, A., and Rodríguez, H. (2016). Iarg-ancora: Spanish corpus annotated with implicit arguments. *Language Resources and Evaluation*, 50(3):549–584.

Tonelli, S., Bryl, V., Giuliano, C., and Serafini, L. (2012). Investigating the semantics of frame elements. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 130–143. Springer.

Tonelli, S. and Giuliano, C. (2009). Wikipedia as frame information repository. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 276–285.

Tonelli, S., Giuliano, C., and Tymoshenko, K. (2013). Wikipedia-based wsd for multilingual frame annotation. *Artificial Intelligence*, 194:203–221.

Tonelli, S. and Pianta, E. (2009). A novel approach to mapping framenet lexical units to wordnet synsets (short paper). In *Proceedings of the Eight International Conference on Computational Semantics*, pages 342–345.

Tripodi, R., Conia, S., and Navigli, R. (2021). United-srl: A unified dataset for span-and dependency-based multilingual and cross-lingual semantic role labeling. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2293–2305.

Tufis, D., Cristea, D., and Stamou, S. (2004). Balkanet: Aims, methods, results and perspectives. a general overview. *Romanian Journal of Information science and technology*, 7(1-2):9–43.

Vaidya, A., Palmer, M., and Narasimhan, B. (2013). Semantic roles for nominal predicates: Building a lexical resource. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 126–131.

Varelas, G., Voutsakis, E., Raftopoulou, P., Petrakis, E. G., and Milios, E. E. (2005). Semantic similarity methods in wordnet and their application to information retrieval on the web. In *Proceedings of the 7th annual ACM international workshop on Web information and data management*, pages 10–16.

Vàzquez, G., Alonso, L., Capilla, J. A., Castellón, I., and Fernández, A. (2006). Sensem: sentidos verbales, semántica oracional y anotación de corpus. *Procesamiento del Lenguaje Natural*, pages 113–119.

Vial, L., Lecouteux, B., and Schwab, D. (2019). Sense vocabulary compression through the semantic knowledge of wordnet for neural word sense disambiguation. *arXiv preprint arXiv:1905.05677*.

Vossen, P. (1998a). Introduction to eurowordnet. In *EuroWordNet: A multilingual database with lexical semantic networks*, pages 1–17. Springer.

Vossen, P. (1998b). A multilingual database with lexical semantic networks. *Dordrecht: Kluwer Academic Publishers. doi*, 10:978–94.

Vossen, P., Agerri, R., Aldabe, I., Cybulska, A., van Erp, M., Fokkens, A., Laparra, E., Minard, A.-L., Aprosio, A. P., Rigau, G., et al. (2016). Newsreader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news. *Knowledge-Based Systems*, 110:60–85.

Vossen, P., Fokkens, A., Maks, I., and van Son, C. (2018). Towards an open dutch framenet lexicon and corpus. *policy*, 12:352.

Wattarujeekrit, T., Shah, P. K., and Collier, N. (2004). Pasbio: predicate-argument structures for event extraction in molecular biology. *BMC bioinformatics*, 5(1):155.

Weischedel, R., Pradhan, S., Ramshaw, L., Palmer, M., Xue, N., Marcus, M., Taylor, A., Greenberg, C., Hovy, E., Belvin, R., et al. (2011). Ontonotes release 4.0. *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium*.

Wen, D., Jiang, S., and He, Y. (2008). A question answering system based on verbnet frames. In *2008 International Conference on Natural Language Processing and Knowledge Engineering*, pages 1–8. IEEE.

Xia, Q., Li, Z., Zhang, M., Zhang, M., Fu, G., Wang, R., and Si, L. (2019). Syntax-aware neural semantic role labeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7305–7313.

Xue, N. (2006). A chinese semantic lexicon of senses and roles. *Language resources and evaluation*, 40(3-4):395–403.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Zaghouani, W., Diab, M., Mansouri, A., Pradhan, S., and Palmer, M. (2010). The revised arabic propbank. In *Proceedings of the fourth linguistic annotation workshop*, pages 222–226.

Zapirain, B., Agirre, E., and Màrquez, L. (2008). Robustness and generalization of role sets: Propbank vs. verbnet. In *Proceedings of ACL-08: HLT*, pages 550–558.