Cross-Lingual Frame Comparability: Computational and Linguistic Perspectives

Von der Fakultät Informatik, Elektrotechnik und Informationstechnik der Universität Stuttgart zur Erlangung der Würde eines Doktors der Philosophie (Dr. phil.) genehmigte Abhandlung.

Vorgelegt von

Jennifer Sikos

aus Santa Clara, California, USA

Hauptberichter	Prof. Dr. Sebastian Padó
Mitberichterin	Prof. Dr. Katrin Erk

Tag der mündlichen Prüfung: 24 Oktober 2022

Institut für Maschinelle Sprachverarbeitung
der Universität Stuttgart

2023

Erklärung (Statement of Authorship)

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig verfasst habe und dabei keine andere als die angegebene Literatur verwendet habe. Alle Zitate und sinngemäßen Entlehnungen sind als solche unter genauer Angabe der Quelle gekennzeichnet. Die eingereichte Arbeit ist weder vollständig noch in wesentlichen Teilen Gegenstand eines anderen Prüfungsverfahrens gewesen. Sie ist weder vollständig noch in Teilen bereits veröffentlicht. Die beigefügte elektronische Version stimmt mit dem Druckexemplar überein.

I hereby declare that this text is the result of my own work and that I have not used sources without declaration in the text. Any thoughts from others or literal quotations are clearly marked. The thesis was not used in the same or in a similar version to achieve an academic grading or is being published elsewhere. The enclosed electronic version is identical to the printed versions.

Contents

Abstract

Frames are descriptions of commonplace scenarios or events. Because they describe everyday scenes, such as *buying* or *eating*, it seems reasonable to assume that many frames in one language would carry over directly to other languages. However, the specifics of how that scene is realized can be highly specific to a culture; it is still an open research question as to how well (and how many) frames actually apply across languages. This thesis concerns cross-lingual frame comparability - the degree to which a frame can be transferred from one language to another. It addresses several aspects of frame comparability: what is frame comparability; how a computational system can measure cross-lingual frame comparability; and how frame comparability affects cross-lingual models of frames.

The work we describe incorporates both linguistic and computational perspectives of frames across languages. Within the linguistic perspective, we draw inspiration from research in existing linguistics literature and conduct our own analyses of frames within and across languages. From the computational perspective, we build systems to a) measure frame comparability and b) predict frames across languages using different cross-lingual and monolingual metrics.

Overall, we contribute to a deeper understanding of cross-lingual frames by defining two dimensions along which frames can be compared: a *lexicographic*

dimension, where the definition of a frame in the lexicon might vary, and a *usage-based* dimension, where uses of the frame in text might vary depending on the sentential or surrounding context. We adopt this lens throughout the rest of the dissertation, culminating in a summary of the ways these dimensions interact and how they could be brought together for a more comprehensive and coherent story of cross-lingual frame comparability.

Zusammenfassung

Frames sind Beschreibungen alltäglicher Szenarien oder Ereignisse. Da sie Alltagsszenen wie *Kaufen* oder *Essen* beschreiben, liegt die Annahme nahe, dass sich viele Frames in einer Sprache direkt auf andere Sprachen übertragen lassen. Die Besonderheiten, wie eine Szene realisiert wird, können jedoch für eine Kultur sehr spezifisch sein; Es ist noch eine offene Forschungsfrage, wie gut (und wie viele) Frames tatsächlich sprachübergreifend gelten. Diese Dissertation beschäftigt sich mit der sprachübergreifenden Frame Vergleichbarkeit - dem Grad, in dem ein Frame von einer Sprache in eine andere übertragen werden kann. Sie befasst sich mit mehreren Aspekten der Frame Vergleichbarkeit, zum Beispiel: Was ist Frame Vergleichbarkeit wie ein Computersystem die sprach-übergreifende Frame Vergleichbarkeit messen kann; und wie sich Frame Vergleichbarkeit auf sprachübergreifende Frame-Modelle auswirkt.

Die Arbeit beinhaltet sowohl linguistische als auch rechnerische Perspektiven von Frames in verschiedenen Sprachen. Innerhalb der linguistischen Perspektive lassen wir uns von der Forschung in der bestehenden linguistischen Literatur inspirieren und führen unsere eigenen Analysen von Frames innerhalb und zwischen Sprachen durch. Aus rechnerischer Sicht bauen wir Systeme auf, um a) die Frame-Vergleichbarkeit zu messen und b) Frames über Sprachen hinweg mithilfe verschiedener sprachübergreifender und ein-

sprachiger Metriken vorherzusagen.

Insgesamt tragen wir zu einem tieferen Verständnis sprachübergreifender Frames bei, indem wir zwei Dimensionen definieren, entlang derer Frames verglichen werden können: eine *lexicographische* Dimension, bei der die Definition eines Frames im Lexikon variieren kann, und eine *auf Verwendung basierende* Dimension, wobei die Verwendung des Rahmens im Text je nach Satz oder Umgebungskontext variieren kann. Wir übernehmen diese Sichtweise für den Rest der Dissertation und enden in einer Zusammenfassung der Art und Weise, wie diese Dimensionen interagieren und wie sie für eine umfassendere und kohärentere Darstellung der sprachübergreifenden Frame-Vergleichbarkeit zusammengebracht werden könnten.

Acknowledgements

A dissertation is a significant undertaking, one which I expected to tackle during my Mutterzeit with twin newborns. Then, a few months into the endeavor, Covid-19 came and the world went into lock down. The childcare plans we'd made fell through, and I had no idea how I'd ever be able to finish this work. In all honesty, it would not have been at all possible for me to finish this document if it weren't for the confidence and encouragement of these folks I need to thank here.

First is my advisor Prof. Dr. Sebastian Padó, who has been a source of encouragement and support, and who set a high standard of research from which I have benefited from immeasurably. Working with him has deepened my competency in research and broadened my knowledge of the field, but it also taught me the importance of mentors caring personally for employees - a model of true leadership by which I hope to follow. Next, I need to thank Dr. Michael Roth for his contributions and eye for detail that have always brought marked improvements to this work and his kind words of support in seeing me through this process.

I also need to thank my family - Les, who worked hard to give me the space and time to finish this. Thank you for encouraging me to pursue my dreams. To Harrison and Lena, who are constant reminders that true learning starts with curiosity.

List of Figures

List of Tables

List of Publications

1. Jennifer Sikos and Sebastian Padó (2018a). "FrameNet's Using relation as a source of concept-based paraphrases". In: *Constructions and Frames* 10.1, pp. 38–60
2. Jennifer Sikos and Sebastian Padó (2018b). "Using embeddings to compare framenet frames across languages". In: *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, pp. 91–101
3. Jennifer Sikos and Sebastian Padó (2019). "Frame identification as categorization: Exemplars vs prototypes in embeddingland". In: *Proceedings of the 13th International Conference on Computational Semantics-Long Papers*, pp. 295–306
4. Jennifer Sikos, Michael Roth, and Sebastian Padó (2022). "Improving Multilingual Frame Identification by Estimating Frame Transferability". In: *Linguistic Issues in Language Technology* 19. DOI: `10.33011/lilt.v19i.939`. URL: `https://journals.colorado.edu/index.php/lilt/article/view/939`

Part I.

Introduction and Background

1. Introduction

Speakers can use creative means to express the same concept in different ways. Although a human would recognize that *X sold a laptop to Y* also means that *Y bought a laptop from X*, in many cases, computational systems still struggle to make the inferences necessary to understand the relationship between these statements. In linguistics, many theoretical frameworks have been proposed to explain the fact that this type of inference comes so naturally to humans. The goal of Frame Semantics (Fillmore et al., 1976) is to account for this surface-level variation by explicitly defining the conceptual knowledge that is being expressed (i.e., *something was bought by someone*). According to frame semantics, humans are able to understand variation in language because, despite different surface forms, each of the variations is referencing the same commonplace scene or scenario called a **frame**. In this thesis, we look at frames, and the variation in language that they capture, as they are applied across languages. We ask whether a computational system can capture similarities and divergences in frames across languages, and what implications these similarities have on multilingual models of frames.

There are many components of a frame that make them a challenge to compare cross-lingually. First is a set of participants that are expected to appear as part of a scene - for instance, in a frame about a commercial transaction, the participants would be a Buyer, a Seller, and Goods. There are also specific words which evoke the frame, and the participants appear as part of the argument structure of those words. In the paraphrase above (*X sold a laptop to Y / Y bought a laptop from X*), *sold* and *bought* are the words that evoke a commercial transaction frame, where *X* is the SELLER, *Y* is the BUYER, and *a laptop* is the GOODS. While many of these roles and terms can translate readily to different languages, there are also attested cases of mismatches in both (see Section 2.3).

The universal status of frames has long been discussed by researchers in the frame semantics community (Boas, 2009; Baker et al., 2018; Boas, 2005b). Several aspects of the theory make it compelling for cross-lingual research; regardless of his or her native language, a speaker is likely to have a way to express an event about a commercial transaction. The words that evoke the frame would very likely be those that are able to express the participants BUYER, SELLER, and GOODS. It is under this assumption that many frames are thought to transfer quite well across languages. In fact, it has been proposed that they are a reasonable place to look for interlingual representations (Boas, 2005b). At the conceptual level, many commonplace scenarios can be shared across cultures and languages; however, typological divergences lead to important differences in the concrete ways that frames are expressed from language to language (Petruck and Boas, 2003). Over the last decade, linguists have increasingly asked how applicable frames are from one language to another: can they be adopted across languages without any alteration, do they fail to apply at all, or somewhere in between? The experience seems to be somewhere in between, where certain frames transfer more readily across languages than others (Baker and Lorenzi, 2020). From a theoretical standpoint, understanding how frames vary across languages contributes to having a more complete picture of how 'universal' frames are (Boas, 2020a).

This thesis concerns **frame comparability**, where comparability is the degree to which a frame can transfer from one language to another. Broadly speaking, there are two sides to understanding a frame: the first side is generalizations of the frame's structure (words that evoke the frame, participants, etc.) that are part of its lexicographic definition, and the second side is the frame's actual use in text (Fillmore and Baker, 2001). With these two components of a frame in mind, our definition of frame comparability includes two aspects of the comparability of frames: the *lexicographic* aspect and the *usage-based* aspect.

On the lexicographic side, the driving question is about the conceptual structures that belong to a specific frame (such as the words which evoke the frame and its participants,) and how those structures are represented differently in a dictionary of frames for any given language pair (Boas and Dux, 2013). Because languages differ in their conceptualization of events, these typological differences (discussed in further detail in Chapter 2) should be reflected in the frame lexicon of those target languages. The Taking frame, discussed in further length in Section 2.3, is an example of lexicographic divergences in a frame across English and German, where the Possessor role exists in the German lexicon but not the English:

English

Core Role	Agent Theme Source
LUs	take.v grab.v

German

Core Role	Agent Theme Source Possessor
LUs	nehmen.v reissen.v

Figure 1.1.: Lexicon entry for the Taking frame in English and German

The principal concern in lexicographic frame comparability is centered around the frame inventory and how a frame's internal structure can accommodate the linguistic properties of a target language.

The second aspect is from the usage side, where the main concern is how well frames align across data from different languages. This aspect includes analyses over translated texts where the goal is to assess how well frame structures align across the language pair, as well as analyzing frames over naturalistic texts in different languages and deciding how similar they appear across those languages. The distinction between the lexicographic and usage-based aspect of frames arises in specific cases where a frame doesn't transfer well under a certain translation or context but is still comparable. For instance, translation shifts can cause frames to transfer poorly across supposedly parallel texts, where part of speech shifts or certain types of paraphrasing (such as Example 1), can cause frames to be mismatched (Padó, 2007; Ohara, 2020):

(1) a. [We $_{\text{AGENT}}$] propose to [*increase* $_{\text{CAUSE\_CHANGE\_POSITION\_ON\_A\_SCALE}}$] [tea $_{\text{ITEM}}$] [prices $_{\text{ATTRIBUTE}}$].

b. Wir schlagen [*höhere* $_{\text{CHANGE\_POSITION\_ON\_A\_SCALE}}$] [Teepreise $_{\text{ITEM}}$] vor.
(We propose higher tea prices.)

Despite the problems that can occur when aligning frame structures across language data, those same frames might still exist in the target language, in the sense that there are other words that can trigger the frame and the participants are the same. Therefore, certain frames that transfer poorly in multilingual corpora can still be comparable from the lexicographic side.

In general, the lexicographic side of a frame represents the frame's potential - what roles, predicates, and other world knowledge should be expected in the frame based on analyses and linguistic introspection. The usage side of a frame is how it is actually realized in text; namely, what roles or predicates frequently appear in the text, and allows for certain lexicographic structures to appear or to be missing in certain contexts. Clearly, it is often the case that the usage and lexicographic aspects overlap, where cases of poor frame alignment in parallel text also indicate that a modification to the definition of the frame is warranted. In fact, much of the work in constructing frame lexicons for different languages involves analyzing frames over translated data (Boas, 2005b). In Section 2.3 describes in more detail the relationship between lexicographic and usage comparability of frames.

Investigating Frame Comparability

The second piece of our work is how we investigate frame comparability, where we adopt methods from both linguistic and computational perspectives. The linguistic perspective is concerned with qualitative assessments of frame behavior across languages, where the linguistic properties of a frame in different languages are analyzed. Linguistic analyses provide a more thorough explanation and insight into the conditions under which frames do not transfer well cross-lingually and have the potential to inform how to interpret or build computational models of frames across languages. The computational perspective centers around cross-lingual computational models of frames, including computational representations of frames and computational measures of frame similarity. Computational models can operate over data too large for a linguist to manually assess and are also the means by which frame semantics is productionized in industry. While both methods provide added value to understanding frame comparability, there can be various challenges in both linguistic and computational investigations.

Challenges of Frame Comparability from the Linguistic Perspective To build a frame lexicon for a new target language, linguists must either begin without assuming knowledge of any existing frames, therefore doubling the work in cases where a frame definition from another language would readily apply to theirs, or analyze each existing frame and its components for suitability to the target language (Boas, Dux, and Ziem, 2016; Atzler, 2011). Empirically analyzing frames across parallel, translated text is also a painstaking process for the linguist because of the variation brought about by translation shifts and paraphrasing.

Challenges of Frame Comparability from the Computational Perspective Frame comparability also affects researchers in Natural Language Processing (NLP) who seek to build computational systems that can detect frame structures in different languages. Many of these systems rely on extensive, annotated data for training machine learning models. For each new language, data needs to be manually annotated. Those annotations require either a preexisting frame lexicon or an investment into its construction, introducing the lexicographic aspect of frame comparability. Creating machine-readable frame semantic resources is demanding in terms of the resources required (Torrent et al., 2014; Ohara et al., 2003; Borin et al., 2010). An alternative to building language-specific frame annotations is to build a system that projects frame structures from one language to another across parallel (translated) texts. However, this introduces many of the issues in usage-based frame comparability where, as described above, this approach encounters issues of frames that don't transfer well as a consequence of translation shifts.

Research Questions

This thesis is concerned with both the linguistic and computational perspectives of frame comparability. The computational systems that we construct incorporate aspects of lexicographic and usage-based perspectives of frame comparability, either in the motivation, design, or analysis of the system. In subsequent chapters, this work will further illustrate in detail the problem of frame comparability in linguistics and its impact on NLP systems. We propose a path towards addressing the issues brought about by frame comparability by answering three basic research questions.

Research question 1: How can we measure frame comparability? First, we address how modern computational systems can provide a quantitative measure of usage-based frame comparability. Much of the prior work in measuring frame comparability has been solely qualitative, where linguists looking at a specific language pair will analyze how well the frame and its components appear in both languages. We describe an explicit, quantitative measure of frame comparability across a language pair based on **word embeddings**. In a case study of cross-lingual frames, we demonstrate how a graded measure of frame comparability can capture relationships between frames across languages automatically. Computational measures of frame comparability can help to inform and accelerate linguistic research, where greater attention can be directed towards frames that are predicted to have low comparability and therefore would reduce the overall workload on the part of the linguist.

Research question 2: How much is frame comparability an issue for models of frame identification across languages? We address the difficulty of building NLP applications for frames across languages; in particular, we discuss the consequences that arise when frames have either a high or low comparability in terms of their use in text. Frame-semantic resources have been constructed for several languages, and due to the high cost of annotation, we consider it worthwhile to determine whether frame annotations from one language will successfully transfer for training systems in another language. Work in this direction can provide additional insight into the frames that are highly comparable (in an usage setting) across languages, as they should be more useful in multilingual training. The requirement for language-specific frame annotations can also be reduced; results of a multilingual system would reveal which frames transfer poorly, requiring more language-specific data to learn, and which frames do not need a high number of annotations as they can be learned from annotated data from other languages – findings that additionally contribute to the lexicographic aspect of comparability. We build a system to identify frames in different languages and present results which demonstrate exactly how much comparability impacts training a multilingual system.

Research Question 3: How can we cope with low frame comparability? There are potentially several causes of low frame comparability, and there are different ways that researchers can cope with these cases. Mismatches in frames that clearly reflect a difference in the frame lexicon, such as role mismatch or language-specific concepts, can be reconciled by either modifying, removing, or creating new frame definitions for the language of interest. In this case, the frame definition itself warrants alterations made by the linguist.

In this thesis, we address low frame comparability from the usage side, where we are interested in frame mismatches across parallel text. However, the methods we propose for coping with frame mismatches in text contribute to the lexicographic side. We focus on cases of poor frame alignment where, although frames are mismatched in text, they still warrant their own entry in the lexicon. We do a linguistic analysis of frame paraphrases, where the frames in the individual sentences differ but the sentences are actually paraphrases of one another – an issue that arises frequently in translated corpora. Although we conduct our analysis over monolingual text, we discuss how the work can be extended to account for cross-lingual cases of frame paraphrases. We describe the linguistic constraints that elicit frame parallelism in these cases and further detail how a computational system might use these constraints for detection of frame equivalence in the way a frame is used in text.

1.1. Thesis Contributions

The majority of the work in this thesis has been presented in corresponding publications, and we list the corresponding publication for each contribution listed. The main contributions include work in computational models of frames that can be used to evaluate and cope with frame comparability across languages. Major contributions, with corresponding publications, include the following:

- This is was first work to explicitly use embeddings to evaluate frame similarity across different language pairs. We discuss the construction of frame embeddings and graded measurements of frame similarity in (Sikos and Padó, 2018b).
- The work presented here is the first to do multilingual frame identification as a standalone task, which is a critical sub-task in a full frame semantic parsing system. We begin this work by experimenting with the best model design for an automatic, multilingual frame identification system, which we first defined in (Sikos and Padó, 2019).

- Reduction of training data for multilingual frame identification. Building on our multilingual frame identification system, we describe a technique to automatically determine which frames will benefit a system for an unseen language, thereby reducing the data one would need for training a frame identification system for a new language ("Improving multilingual frame identification by estimating frame transferability"). The prediction of which frames transfer over languages incorporates features from both the annotated data (usage-based), as well as features from the available frame lexicon (lexicographic).

- Definition of the linguistic constraints that account for conceptual paraphrases which evoke different frames. Although the issue of paraphrase relations has been identified in the past, this work is the first to concretely define the conditions which trigger different frames to be evoked across paraphrastic sentences. We use existing structures from the FrameNet lexicon as sources of conceptual paraphrases. Linguistic constraints and frame paraphrases are discussed in (Sikos and Padó, 2018a).

1.1.1. Contributions to Dimensions of Frame Comparability

We have described in the introductory paragraphs above two dimensions along which we study frames across languages: 1) a lexicographic/usage-based view of frames, and 2) a computational/linguistic approach to researching frames. This thesis touches on all of these dimensions at different points. An overview of the contributions from each section is given in Figure 1.2, where:

- Lexicographic/Linguistic: we describe in Part IV a set of side conditions that can be incorporated into a frame lexicon. We also analyze frames and their lexicographic structures Part IV, Section 7.3.1

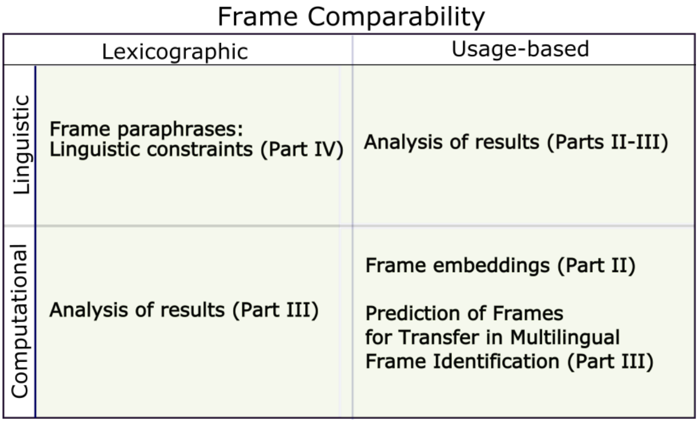


Figure 1.2.: Contributions of this thesis include computational and linguistic work, where both lexicographic and usage-based aspects of frame comparability are discussed.

- Lexicographic/Computational: Part III we describe our experiments with cross-lingual frame selection. We compare performance of frames that were selected based on their lexicographic divergences as part of the results analysis.

- Usage-based/Linguistic: Parts II-III we compare the output of our embedding models (based on frames in context) with the current literature on frames across languages from linguistics. Part IV also describes our analysis of cases where different frames appear in paraphrased text.

- Usage-based/Computational: Parts II-III describe our experiments in building embedding-based models of frames.

1.2. Contributions to Published Work

The above contributions were published in corresponding peer-reviewed papers with multiple authors. In work on evaluating frame similarity with frame embeddings, described in Chapter 4, I contributed by implementing the model and conducting the analysis of frames with high and low comparability, as well as compiling prior linguistics research in cross-lingual frames as motivation for the study. The semantic neighborhood check (Section 4.3.2) and rules for combining German multi-word predicates, as well as input to the study's design was carried out by my co-authors. I implemented and evaluated frame identification system defined in (Sikos and Padó, 2019), as well as background literature review for exemplar and prototype theory. Motivation to the experiment design was conducted by my co-authors, who contributed to the analysis as well. Work on multilingual frame identification and frame selection ("Improving multilingual frame identification by estimating frame transferability") was collaboratively designed by all co-authors, as well as the evaluation metrics and the breakdown of the frame selection procedure. I constructed the model and experiments, as well as any plots or results. Finally, I analyzed and did the classification of frame pairs in (Sikos and Padó, 2018a), and created frame paraphrases in the *Using* relation for my analyses and classification. Side conditions and motivation was conducted by my co-author as well as much of the background and substantial parts of the discussion in the study.

1.3. Thesis Overview

This thesis is structured as follows. Chapter 1 describes the motivation and background relevant to the contributions of the thesis. Chapter 2 outlines the theoretical background, including a brief discussion of the history of frame semantics. Chapter 3 gives background to the computational methods described in the rest of the thesis, specifically distributional semantic models (embeddings) and frame semantic parsing. Chapter 4 describes work in building and aligning embeddings of frames across languages, and chapter 5 gives an overview of our model architecture for frame identification. Chapter 6 adopts this architecture for predicting frames across languages, and describes our work in selecting frames from different languages to be used in training a target language. Chapter 7 gives a linguistic analysis of frame paraphrases - utterances which appear similar but evoke different frames. Chapter 8 concludes.

2. Theoretical Background

Frame semantics is a theoretical framework that has developed over decades of research in linguistics, and is based on an even longer history of research in semantics. Its theoretical foundations contributed to a different approach towards a semantic analysis of language. It does this in two important ways: first, it explicitly introduces world knowledge (vis-á-vis everyday scenarios) into a model of language use, and second, it ties together predicates whose semantic behaviors are similar under a unified model of meaning. Section 2.1 begins with a brief discussion of earlier theories of lexical semantics that contrast with the frame-based view of language. To situate frame semantics with the broader program of cognitive theories of grammar, we briefly discuss the history of psychologically explanatory models of language. Section 2.2 describes further details of the history and theory of frame semantics and its framework for language understanding. Section 2.3 gives background to frame comparability in linguistics and describes many of the typological differences that cause poor frame alignment across languages.

2.1. Lexical Semantics Before Frames

Behaviorism had been the dominant paradigm for research in linguistics in the early 20th century. During this time, the field of linguistics was devoted to descriptive work, where linguistic research was restricted to surface-level, observable phenomena. (Harris, 1995). The Chomskyan revolution in the mid-20th century turned linguistics away from purely descriptive work and towards more formal representations of language. In Chomsky's view, as a study of language, linguistics ought to concern itself with the cognitive structures that mediate language use (Chomsky, 1964). By introducing formalism into linguistic analysis, linguists sought to describe the underlying representations that account for surface-level linguistic behaviors. During this time, Chomsky introduced his transformational grammar, where different surface forms of a shared concept (ex, the passive vs active voice) can be derived by systematic rules that apply to an underlying, deep representation (Chomsky, 1957).

One of the early theories in lexical semantics introduced the concept of feature attributes, where a word's meaning can be decomposed into its underlying primitive features (Katz and Fodor, 1963; Katz, 1964). In a feature-based view of word meaning, the term *bachelor* would have semantic primitives Male and Not Married (Resnik, 1996).

Alternatively, a formal semantic theory is one in which truth conditions are essential to composing meaningful expressions (Chomsky, 1955). Within this framework, sentences in natural language are written in predicate logic, where truth conditions dictate similarities in word meaning: if the same circumstances cause A and B to be true, then there is an entailment relation between A and B (Montague, 1970; Janssen, 2011). Central to these methodologies is the adherence to the Principle of Compositionality (Frege, 1884), where the meaning of the whole can be understood by the meaning of its individual components. Within these theoretical frameworks, lexical semantics is primarily interested in, and restricted to, the semantics of a single sentence.

Although the Chomskyan revolution brought formalism to the forefront linguistic theory, Chomsky makes clear that only specific, rule-abiding phenomena constitutes the internal system of language that belongs in a grammar (what he calls a speaker's "competence"). Anything outside of that purview belonged to the "performance" aspect of language, and should not be incorporated into a model of language.

By the 1970s a new wave of thought began to take hold: that, when it comes to representations of natural language, "context is of overwhelming importance in the interpretation of text. Implicit real-world knowledge is very often applied by the understander, and this knowledge can be very highly structured" (Schank and Abelson, 1988). During this time, theories of language that were based on a psychologically explainable account of language started to incorporate into their models exactly the type of phenomena that Chomsky considered "performance" and formal semantics was not able to account for. Counter to the representation of meaning as being purely truth-based or mathematical, linguists began to describe language in terms of the world knowledge required to understand it (Abelson, 1976; Croft and Cruse, 2004).

One of the first to achieve this was Charles Fillmore, who published his seminal work that introduced the concept of the frame – pushing the notion that knowledge of the real world was necessary to truly understand what a speaker means when using language (Fillmore et al., 1976). Frame semantics entered into the picture as an alternative to formal logic and mathematics-based representations of meaning where, according to Fillmore, work in lexical semantics should incorporate more rich descriptions of the human experience. Theories that do so are about a semantics of understanding (U-semantics) (Fillmore, 1985). A U-semantic theory contrasts with truth-conditional (T-semantic) theories, where a U-semantic theory does not require an expression to be assigned a truth value, but instead can incorporate context like cultural or experiential knowledge about the world as part of a grammar of language. With these ideas in mind, Fillmore explicitly formalized background knowledge required for language understanding as part of his theoretical framework.

2.2. Frame Semantics

2.2.1. Motivation

Frame Semantics begins with the assumption that a word's meaning is only able to be understood when the background setting, or knowledge, is available to the listener. This includes certain pragmatic knowledge like cultural facts or common experiences. In a classic example, to understand the word *breakfast*, one would have to understand that, in the culture in which the term is used, there are relatively fixed meal times where breakfast occurs in the morning hours, and that breakfast food generally consists of pancakes, eggs, toast, etc. (for Americans) which is typically only eaten in the morning (Petruck, 1996). The term can be used to refer to the type of food that is eaten or the morning meal time. All of this background knowledge is necessary if one were to make sense of the utterances, *they had breakfast for dinner*, or as a response to a question about the meeting time between two parties: *they met for breakfast* (Fillmore, 1982).

It is clear that frames can incorporate a broad range of world knowledge, but it is important for any grammar of language that representations in the model must be shared amongst all speakers of that language. This means that the knowledge that belongs to a frame represents standard, or prototypical experiences. From its beginnings, Frame Semantics drew from theories of natural categories, specifically prototype theory (Rosch, 1973b). In prototype theory, a category is an abstraction over its members where some members have more importance to defining the category than others (Posner and Keele, 1968). Frames represent abstract, prototypical categories of events or scenarios where certain uses of its lexical units are more prototypical examples of that frame than others (Fillmore, 1982; Tannen, 1993). In the *breakfast* example above, both sentences reflect non-prototypical uses of the term, where a prototypical use would simply access the scenario where an individual eats a meal in the morning: *he read the newspaper while eating breakfast.* Although this reading is more prototypical of the definition of the frame, all three uses require access to the same background knowledge for a listener to understand their meaning.

Fillmore's case grammar (Fillmore, 1968) had a deep influence on the types of background knowledge that belong to the definition of a frame. In case grammar, words have a deep case, that is, a specific number of arguments which can be filled by specific participants. While his original work included six participant types that were thought to have more universal applicability (Agent, Patient, Theme, etc.), the subsequent implementation in frame semantics has extended this work to a much larger set (see Section 2.4.1).

2.2.2. Theoretical Framework

To reiterate from above, frames provide a conceptual background for language understanding, where a word's meaning is understood by the frame it evokes. The conceptual background accompanies linguistic constraints which specify how the frame is expressed in language. Specifically, a frame is an everyday scene or scenario where specific words evoke the scene. In a frame such as a Commercial_transaction, several predicates (*buy, purchase, sell*) are capable of evoking the frame. Those same predicates could have several different senses, some of which are unrelated to the frame: it is clear that *He got an iPod from the Apple store* is not accessing the same knowledge structure as *He got her point*, despite the predicate *got* being present in both sentences. Frame semantics accounts for different word senses because each unique sense activates a different frame. In the first sentence, *got* activates a COMMERCIAL_TRANSACTION frame, while the other activates a GRASP frame, where a person is understanding some phenomenon. The specific sense of a word (and its part-of-speech tag) that activates a frame is referred to as a frame's **lexical unit (LU)**.

Each frame has a set of participants, called **semantic roles**, which belong to the conceptual structure of the frame. In the paraphrase below, both sentences discuss a COMMERCIAL_TRANSACTION frame, triggered by the lexical units *bought* and *sold*, which have the semantic roles BUYER, SELLER, and GOODS:

(2) [Aylin $_{\text{BUYER}}$] [*bought* $_{\text{COMMERCIAL\_TRANSACTION}}$] [a new book $_{\text{GOODS}}$] from [Derek $_{\text{SELLER}}$].

(3) [Derek $_{\text{SELLER}}$] [*sold* $_{\text{COMMERCIAL\_TRANSACTION}}$] [a new book $_{\text{GOODS}}$] to [Les $_{\text{BUYER}}$].

At the linguistic level, these semantic roles appear as part of the valence patterns (that is, the number of required arguments,) of the frame's LUs. In Examples 2 and 3, the LUs *buy* and *sell* are trivalent, meaning they have three core arguments: the semantic roles BUYER, SELLER, and GOODS. As the examples above demonstrate, frames can generalize not only over the lexical level, but syntactic variation is also allowable within a frame. In the paraphrase above, *buy* and *sell* both have the same semantic roles but the syntactic distribution of those roles are different across sentences. In Example 2, the BUYER is the subject, the GOODS the direct object, and the SELLER the indirect object. Example 3 shows the SELLER is the subject and the BUYER the indirect object.

Certain roles are expected as part of the argument structure of the predicating LU and/or are critical to understanding the meaning of the frame. However, additional, **non-core roles** might be described in text that are not thought to be necessary to the frame. In the example of a frame about commerce, additional roles might include MONEY, TIME, and LOCATION. The semantic roles that are part of the argument structure of the LU, and are necessary components in the frame, are **core roles**. It is not always the case that all the core semantic roles will appear in any given context. If one of the core roles is not realized, as in the SELLER in Example (4), its existence is nonetheless implied. This signals its status as a semantic role that is core to the meaning of the frame, and belongs as part of the conceptual structure of the frame's lexicon entry:

(4) [Aylin $_{\text{BUYER}}$] [*bought* $_{\text{COMMERCIAL\_TRANSACTION}}$] [a new book $_{\text{GOODS}}$]

It is clear that, although they are conceptual categories, frames are also highly structured in the systematic, syntactico-semantic properties of their LUs. A semantic role can appear only once for a single predicate; for instance, in a COMMERCIAL_TRANSACTION frame, the LUs *buy, purchase, sell*, or *get* would express at most one argument characterizing a BUYER, SELLER, or GOODS.

In sum, frames are defined by a set of lexical units and semantic roles. Word relationships are accounted for by the fact that they reference the same frame, and the lexical units in a frame have argument structures that support the expression of the frame's semantic roles. Because they describe everyday experiences, the question of how universal frames are has existed since the theory was proposed; for the last several decades, linguists have analyzed frames to

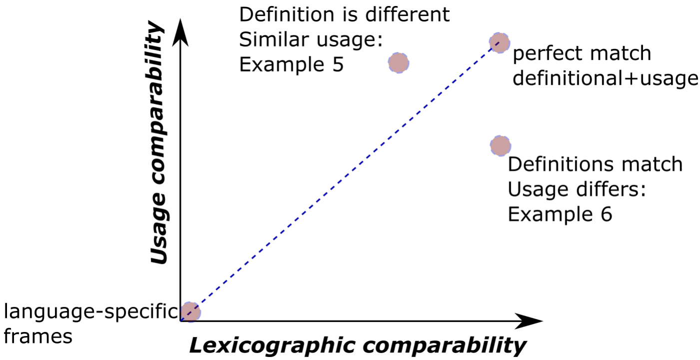


Figure 2.1.: Lexicographic and usage-based aspects of frame comparability.

determine their comparability across many different languages, shedding light on the issues that arise when applying frame structures from one language to another.

2.3. Frame Comparability in Linguistics

Comparing frames across languages can mean addressing the lexicographic (definitional) aspect of frames – specifically, what information about a frame is stored in the frame lexicon for different languages, and the usage-based aspect of the frame, that is, how frames are expressed across data from different languages. The distinction between linguistic observations that are incorporated into a lexicon and corpus data that describes attestations of language use has a long history of discussion in linguistics (Hanks, 2009; Harris, 1995; Hanks and Allan, 2013; Gries and Ellis, 2015). Fillmore discusses that there is a necessity for both the "armchair linguist," that is, someone who analyzes the native intuitions about the grammar of a language, and the corpus linguist, where attestations of the native language are studied over (sometimes large) collections of data (Fillmore, 1992). FrameNet (discussed below in Section 2.4.1) therefore incorporates elements of both – a lexicon, which represents realization potential of the frame, and examples drawn from corpora which demonstrate the frame's actual realization in text.

In chapter 1, we introduced a definition of cross-lingual frame comparability where comparability is the degree to which a frame can apply from one language to another. Given the different objectives of lexicographic and usage-based perspectives, the specific criteria of what constitutes high or low frame comparability needs to be modified for each purpose. Below, we offer concrete definitions of high and low frame comparability in both lexicographic and usage-based aspects, discussing how they relate and where they differ. The relationship between the two aspects is visualized in Figure 2.1, where instances of frames can be plotted in terms of their lexicographic and usage comparability across languages. Many frame instances would be hypothesized to fall within or near the dotted line, where there is a perfect correlation between the dictionary definitions of the frame across languages and its attestations in translated text. There are, however, cases where this correlation breaks down, which we discuss in the definitions below.

Lexicographic Frame Comparability For the lexicographic aspect of frame comparability, a frame is highly comparable when there exists lexical units in the target language that fit the meaning of the frame in the source language, and semantic roles are consistent. Focusing on the semantic roles and definition of the frame means that the lexicographer is less concerned with issues of certain translation shifts, such as shifts in POS, that might cause mismatches over parallel text, but is instead concentrated on the question of whether the definition of the frame in language A can be directly adopted for language

B. Lexicographically, several frames have been found to be highly comparable across languages, even those that are typologically unrelated. Japanese and Brazilian Portuguese, for instance, have been found to have frames which were originally defined for English (Ohara, 2020).

Frames that have low lexicographic comparability would include those in which there is a mismatch in the definition of the frame across languages. These include cases where there are critical typological divergences that cause frame structures to differ. This is especially the case for mismatches in semantic roles that can be attributed to differences in argument structures. For instance, German dative case is more flexible than English, allowing for an additional semantic role, POSSESSOR for the LU *nehmen.v* (*take*) (Burchardt et al., 2009):

(5)	[Er $_{\text{AGENT}}$]	[*nahm* $_{\text{TAKING}}$]	[ihm $_{\text{POSSESSOR}}$]	[das Bier $_{\text{THEME}}$]
	He	took	him	the beer
	[aus der Hand $_{\text{SOURCE}}$]			
	out of the hand			

“[He $_{\text{AGENT}}$] [*took* $_{\text{TAKING}}$] [the beer $_{\text{THEME}}$] [out of his hand $_{\text{SOURCE}}$]”

Example 5 (discussed in further detail in Section 2.3,) suggests that TAKING has a lower lexicographic comparability across English and German, where the POSSESSOR is a core role in German that is not defined for English, shown in Figure 2.2 below.

Low lexicographic comparability in the TAKING frame can still appear as high usage-based comparability in contexts where the POSSESSOR is not expressed in German, demonstrating the difference between the realization of the frame over parallel text and the lexicographic definition of the frame:

(6) [He $_{\text{AGENT}}$] [*took* $_{\text{TAKING}}$] [a beer $_{\text{THEME}}$] [from the fridge $_{\text{SOURCE}}$].

[Er $_{\text{AGENT}}$] [*nahm* $_{\text{TAKING}}$] [ein Bier $_{\text{THEME}}$] [aus dem Kühlschrank $_{\text{SOURCE}}$].

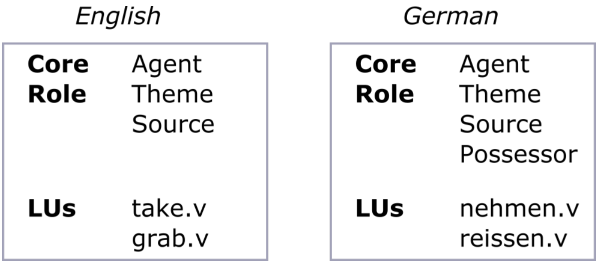


Figure 2.2.: Lexicon entry for the Taking frame in English and German

A predominate number of instances such as Example 6 would suggest the frame can be plotted along the lexicographic/usage-based scale of comparability in Figure 2.1 as having a lower lexicographic but higher usage-based comparability.

For both lexicographic and usage-based comparability, frames that have the lowest comparability are frames that cannot be compared across languages at all – those that mismatch at the conceptual level. These include highly culturally-specific experiences which are unlikely to appear in another culture and language. An example of a culture-specific frame is the German idea of *Kulanz*, which can be thought of as 'a graceful act of courtesy in regards to a commercial transaction', a concept that does not exist in even closely related languages such as English (VanNoy, 2017). For these frames, there is no translational equivalence and comparability is essentially impossible. These frames are plotted at the origin in Figure 2.1, as they need to be defined from scratch by the linguist for each new language.

Usage-based Frame Comparability When doing a usage-based comparison of frames across language data, the highest comparability between frames as they are used in corpus data is one in which, across languages, one should expect that "all translations of lexical units in one source language frame to fall into the same target language frame" (Baker and Lorenzi, 2020) and the semantic roles are the same in both languages. Concretely, this would mean that the ideal case of usage-based frame comparability is one in which parallel, translated text shows consistent use of frame structures that can be easily aligned. This would mean that, regardless of the translation of a source language LU, the target language LU would still evoke the same frame. Many frames that are suspected to have high usage-based and lexicographic comparability relate to events that describe common human experiences (MOTION or COMMERCIAL_TRANSACTION, for instance). This is generally the case because such frames are defined broadly, where many different lexical units with varying argument structures define the meaning of the frame (Padó, 2007). For this reason, these frames are more likely to be found consistently across even typologically unrelated language pairs.

There can be many types of low frame comparability from the usage perspective. This is in part due to the translation shifts that make alignment of frame structures across translations difficult, if not impossible (Samardžic et al., 2010). One example of low usage-based comparability is a semantic role that is dropped for translation purposes, such as voice or POS shifts, leading to so-called implicit semantic roles (Roth and Frank, 2013; Ruppenhofer et al., 2010; Gerber and Chai, 2012). Although they are not explicitly mentioned in the text, the role can be inferred by the context (Sikos, Versley, and Frank, 2016):

(7) The more [we $_{\text{SPEAKER}}$] [*refuse* $_{\text{AGREE\_OR\_REFUSE\_TO\_ACT}}$] [to democratize the institutions $_{\text{PROPOSED\_ACTIONS}}$] ...
Je mehr [die Demokratisierung der Institutionen $_{\text{PROPOSED\_ACTIONS}}$] [*verweigert* $_{\text{AGREE\_OR\_REFUSE\_TO\_ACT}}$] wird ...

"The more the democratization of the institutions are refused..."

Example 7 demonstrates a mismatch over parallel text, where the role Speaker is missing in the German translation due to a nominalization of the verb *democratize*. In cases of LUs that are often translated into different POS for a target language, it can be the case that the frame structures will align poorly for most translations; these cases demonstrate frames with low usage-based comparability, but do not pose a problem in terms of their lexicographic comparability. The frame Agree_or_refuse_to_act in Example 7 has the Speaker role in its definition for German, even though it is not realized in the given text. Frames with the lowest usage-based comparability are identical to the case of lexicographic comparability, where frames are language-specific and thus cannot be aligned.

Causes of Low Frame Comparability

Low frame comparability can be attributed to several different causes, many of which can be traced to typological differences in linguistic properties of the language pair. Typological differences, such as differences in lexicalization patterns or differences in argument structure, will not only affect the usage of frames across parallel data, but they will also require alterations to the definition of the frame in the frame lexicon; in other words, these issues will affect both the usage-based and lexicographic aspects of frame comparability.

Argument Structure Syntactic valence describes the number of arguments that a lexical unit is expected to have, and across languages, argument structure can differ for lexical units that evoke the same frame. When this argument structure differs, it affects the number and type of core semantic roles that occur with the frame. Since the core arguments are a central part of the definition of the frame, this leads to divergences in the way the frame can be defined across languages. In English, for instance, the dative case is much more restrictive than German; Burchardt et al., 2009 discuss the case of the TAKING frame, evoked in sentences such as:

(8) [He AGENT] [*took* TAKING] [the beer THEME] [from the cooler SOURCE]

In this frame, there are core semantic roles AGENT, THEME, and SOURCE, where the SOURCE is a location. In addition to a location as a SOURCE, the SOURCE in English can include a person as a dative argument in the SOURCE role:

(9) [He AGENT] [*took* TAKING] [the beer THEME] [from him SOURCE]

German, however, is more flexible in terms of the dative case, and both the location and person can be realized. Because they are both realized, there needs to be a distinction between the location, the SOURCE, and the person, which can be defined as a POSSESSOR. Both are realized with the lexical unit **nehmen** which evokes the TAKING frame:

(10) [Er AGENT] [*nahm* TAKING] [ihm POSSESSOR] [das Bier THEME]
He took him the beer
[aus der Hand SOURCE]
out of the hand
"He took the beer out of his hand"

In these cases, argument structure can determine core semantic roles in a frame, which impact the frame's definition and comparability of the frame across the language pair. The definition of the TAKING frame necessarily distinguishes the SOURCE role from the POSSESSOR role in German, whereas the English frame only defines a AGENT, THEME, and SOURCE core semantic roles. This mismatch not only reflects a low usage-based comparability, where the frames are not aligned across translations, but the core semantic roles are different as well, indicating that the frame has low lexicographic comparability.

Lexicalization Lexicalization concerns the semantic information that gets encoded in a word. This naturally impacts the comparability of frames across languages, as different languages necessarily express certain concepts in ways that can evoke different frames. This is especially the case for verbs of motion, where satellite languages such as English express the manner of motion as part of the verb (*She ran into the room*), while verb-framed languages such as Spanish often express this manner externally (*She entered the room running*) (Talmy, 1985). These typological differences in many cases lead to different frames being evoked when expressing the same concept (Ellsworth et al., 2006; Boas, 2013):

(11) a. ...[we $_{\text{SELF\_MOVER}}$] started to [*walk* $_{\text{SELF\_MOTION}}$] [to $_{\text{GOAL}}$] Merripit House.
...nos [*poníamos en camino* $_{\text{SETTING\_OUT}}$] [hacia $_{\text{DIRECTION}}$] la casa Merripit.
"we put in path towards the house Merripit."

In Example 11, the English lexical unit *walk* evokes the SELF_MOTION frame, which focuses on the movement of the SELF_MOVER. Many of the lexical units in the SELF_MOTION frame concern manner of movement (*crawl, hike, run, hobble*, etc.). The Spanish translation expresses the equivalent concept as the SETTING_OUT frame which focuses on the journey and destination rather than the manner of motion. Divergences in these lexicalization patterns can clearly impact frame comparability across languages; in the case of verbs of motion, the English sentence evokes a general frame relating to movement, while the Spanish frame is specific to the directionality of the path. While the SELF_MOTION and SETTING_OUT frames are conceptually distinct, in this case they are used to refer to the same event. These cases could be indicators that there is a likely relationship between frames, and occur often in translated texts.

Paraphrasing There are additional issues that only affect the usage-based aspect of frame comparability across languages, which include differences in translation choices. The typological differences above cause utterances in a source language to be paraphrased in a target language, and in many cases, the linguistic constraints of the target language require that a given sentence be translated as a paraphrase. However, there are additional instances in which a translator paraphrases from the source language by choice, leading to sentences which evoke different frames in the target language. These cases reflect low comparability in terms of the alignment of the frame across parallel data.

In many cases, translations which evoke different frames involve paraphrases which draw upon more nuanced world knowledge than is captured at the frame level: so-called **conceptual paraphrases** (Sikos and Padó, 2018b). Conceptual paraphrases can be found in monolingual paraphrases, which we discuss in later chapters, but they also appear cross-lingually where translators can employ different frames to express a similar concept (Padó, 2007):

(12) a. [We AGENT] propose to [*increase* CAUSE_CHANGE_POSITION_ON_A_SCALE] [tea ITEM] [prices ATTRIBUTE].

b. Wir schlagen [*höhere* CHANGE_POSITION_ON_A_SCALE] [Teepreise ITEM] vor. (We propose higher tea prices.)

Example 12a evokes the CAUSE_CHANGE_OF_POSITION_ON_A_SCALE frame, where an AGENT or CAUSE is affecting the increase of prices, and Example 12b evokes the Change_ position_on_a_scale frame, where no cause need to be present. While the frames are clearly related by a causative relationship, and there is a certain amount of overlap in semantic roles between the frames, there are many cases which demonstrate how these two frames are conceptually distinct from one another. In Cause_change_of_position_on_a_scale, the Agent role is a core role to the meaning of the frame, and even with the omission of the Agent, the existence of a causative force is clearly implied:

(13) [Oil $_{\text{ITEM}}$] prices were [*raised* $_{\text{CAUSE\_CHANGE\_OF\_POSITION\_ON\_A\_SCALE}}$] yesterday.

On the other hand, the Change_position_on_a_scale implies no clear cause or agency which produces the change in value:

(14) [The population of Smallville $_{\text{ATTRIBUTE}}$] [*increased* $_{\text{CHANGE\_POSITION\_ON\_A\_SCALE}}$] [fourfold $_{\text{DIFFERENCE}}$].

The paraphrase in Example 12 shows how both of these frames are evoked to describe the same event. Cross-lingually, these cases can emerge because of syntactic and semantic divergences. For English and German, for instance, the translator chose an adjective to express the increase in price, although **höhere** is a stative in German and does not take an agent. This decision meant that, although the translation expresses the same information, the agent role is absent from the German sentence and the Change_position_on _a_scale frame is evoked.

The above examples illustrate certain issues in frame comparability across languages - how, from a linguistic standpoint, assessments of usage-based and lexicographic frame comparability can come with significant challenges. Constraints in different languages make the transference of certain frames and frame structures untenable. Overall, low frame comparability increases the workload of the linguist, where low comparability requires either constructing a new frame lexicon and/or painstakingly evaluating each mismatched frame structure in translated texts. In chapter 3, we describe how these issues affect frame semantics in computational systems, where the computational linguist must grapple with issues relating to low comparability when creating a system for detecting frame structures across languages automatically.

2.4. Frame Semantic Resources

2.4.1. Berkeley FrameNet project

FrameNet(Baker, Fillmore, and Lowe, 1998) is the first and most prominent lexicographic resource where frames and their components are concretely defined. The project was initiated, and continues to be managed, at the International computer Science Institute (ICSI) in Berkeley where the frame lexicon has been established for the English language. The project continues to grow, and several versions of FrameNet have emerged over the years as the inventory of frames changes and develops over time. FrameNet has from its beginnings been designed as both a frame semantic resource for lexicographers, as well as a machine-readable resource for NLP scientists (Baker, 2014). In addition to a digital lexicon of frames, FrameNet provides annotated data that is used as a resource for training frame semantic parsers. Frames are connected to one another within the resource, making the entire lexicon a hierarchical graph structure. This graph structure has itself been used in models of frame semantics in NLP (Botschen, Mousselly-Sergieh, and Gurevych, 2017; Popov and Sikos, 2019).

Components of the Lexicon

FrameNet's frame lexicon features several core components that define a frame conceptually and linguistically. Each frame is first defined at the conceptual level, where a general description of the frame is provided. Frames can either be lexicalized, meaning they have surface expression through lexical units and semantic roles, or they can exist as purely conceptual devices that connect more specific, lexicalized frames. Frames that have a linguistic expression are called **lexical** frames, while frames that are abstract concepts that are not realized linguistically are called **non-lexical** frames. Non-lexical frames are not only generalized concepts, but they typically represent perspectives on lexical frames (Osswald and Van Valin, 2014). These include frames such as a COMMERCE_MONEY-TRANSFER, where money is exchanged between individuals; this abstract frame is lexicalized by two related frames concerning transactions where money is paid for a good or service: COMMERCE_COLLECT and COMMERCE_PAY. This abstract frame relates the two money-related transaction frames, in contrast to frames which focuses on the exchange of goods rather than money. Frames related to exchange of goods evoke a non-lexical COMMERCE_GOODS-TRANSFER, which includes lexical frames COMMERCE_BUY and COMMERCE_SELL. Figure 2.3 shows the diagram of these relationships, where the non-lexical frames are conceptually linking the lexical frames within the hierarchy. A vast majority of frames in FrameNet are lexical frames, which are the frames that have an explicit set of LUs.

Each frame defines the LUs which can evoke it in text; for example, *collect.v, collection.n,* and *charge.v* are LUs for the frame COMMERCE_COLLECT (shown in Figure 2.3). Within the resource, a single lexical unit is represented as a lemma and its part-of-speech tag. Multi-word expressions can also evoke frames, such as *chief executive officer.n* in the LEADERSHIP frame, or collo-

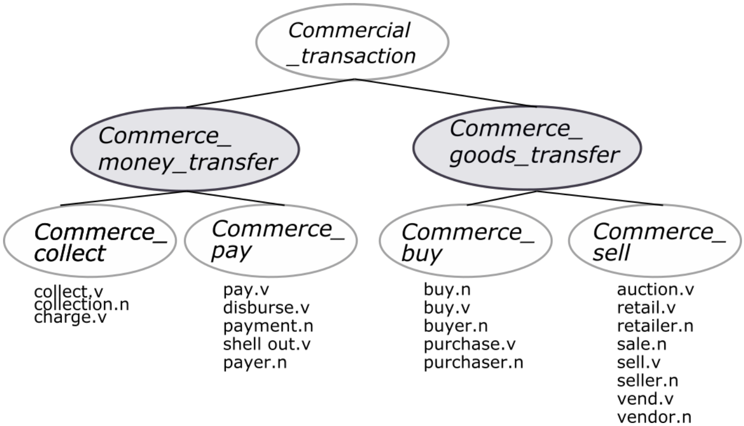


Figure 2.3.: Lexical frames COMMERCE_COLLECT, COMMERCE_PAY, COMMERCE_BUY, COMMERCE_SELL and their LUs and non-lexical frames COMMERCE_MONEY_TRANSFER, COMMERCE_GOODS_TRANSFER, which serve to conceptually link relationships between lexical frames but do not have LUs themselves.

quialisms such as *make a name for oneself.v* in the FAME frame. LUs can be verbal, nominal, adjectival, or adverbial, and many frames have a combination of different part-of-speech types in their set of LUs. Other frames, for instance, the CLOTHING frame, are predominately one category – in this case, nominals that describe items of clothing (*pullover.n*, *cape.n*, *vest.n*, etc.) FrameNet provides attestations of a frame's lexical unit in corpus data, where examples of the LU in corpora are annotated for their argument structure and realization of the frame's semantic roles in different syntactic configurations (Ruppenhofer et al., 2006).

Semantic roles are described as **frame elements (FEs)** in the resource, as the roles not only describe argument structure (where core and non-core roles of the predicating LUs are distinguished for each frame,) but they also include a highly specific description of the filler for that role (Boas, 2005a). These fillers are semantically transparent labels which describe the type of entities that fill a role. Certain fillers are found across several, sometimes unrelated frames, such as PLACE, which describes a location of the event, or TIME, which is a description of when the event occurs. However, many frames have fillers that are specific to the concept of the frame and its related frames, such as a BUYER for frames relating to commerce, or a DELIVERER in a frame about DELIVERY. Because the fillers for semantic roles are sometimes specific to a frame, the inventory of FEs in FrameNet is relatively large, with over 10k FEs currently in use within the resource. An example of core and non-core roles in FrameNet is shown in Figure 2.4. Comparison between the roleset in COMMERCE_SELL and COMMERCE_PAY shows that even closely related frames can have differences in their core role set. COMMERCE_PAY also has the core roles BUYER, SELLER, and GOODS, but it specifies additional core roles (MONEY, RATE) that are not a core part of the COMMERCE_SELL frame. The overlap in specific core roles (i.e., BUYER) demonstrates that there is a conceptual relationship between the two frames, but the difference in the core roles also depicts the conceptual difference between these two closely related frames.

In addition to fillers for specific semantic roles, FrameNet also provides information about the **semantic type** of its FEs (Petruck et al., 2004). Semantic types are a fixed set of attributes that can be broadly applicable, meaning they can occur in frame elements or frames that involve completely unrelated concepts. Common semantic types include "Sentient" for roles that involve a person such as the COGNIZER FE, or "State_of_affairs" for an EXPLANATION FE.

Finally, frames are connected to one another by so-called **frame-to-frame**

Core:

Buyer [Byr]	The Buyer has the Money and wants the Goods. Lee SOLD a textbook to Abby.
Goods [Gds]	The FE Goods is anything (including labor or time, for example) which is exchanged for Money in a transaction. Kim SOLD the sweater.
Seller [Slr]	The Seller has possession of the Goods and exchanges them for Money from a Buyer. So far, my company has SOLD more than three million copies.

Non-Core:

Back [Bac]	This FE indicates that the selling act reverses an earlier separate act in which the current Seller bought the Goods from the current Buyer. There is a plan for Mercer to buy the ground and SELL it back to the club for more money.
Explanation [Exp] **Semantic Type:** State_of_affairs	The Explanation for which an event occurs.
Imposed_purpose [Imp]	The Buyer's intended purpose for the Goods. He SOLD the filing robot to me for filing all my legal documents.

Figure 2.4.: FrameNet's core and non-core frame elements for the frame COMMERCE_SELL

relations, which is one of the additional conceptual devices within the FrameNet resource. Frame-to-frame relations specify that, not only are frames connected to one another conceptually, but that two frames can have linguistic, ontological, or other types of conceptual relationships that define *how* they relate to one another. FrameNet currently has 14 relations, some of the most common including ***Inheritance, Using, Perspective_on, Subframe, Precedes, Inchoative_of,*** and ***Causative_of***. Frame-to-frame relations such as **Inheritance** and **Subframe** are analogous to typical "is-a" in standard ontologies (Scheffczyk, Pease, and Ellsworth, 2006; Chow and Webster, 2007). Other frame-to-frame relations that are based on linguistic relationships include the ***Causative_of*** and ***Inchoative_of*** relations, where one frame can take a CAUSE or AGENT FE as a core role and its inchoative counterpart can't (Petruck et al., 2004). For example, the CAUSE_TO_FRAGMENT frame is

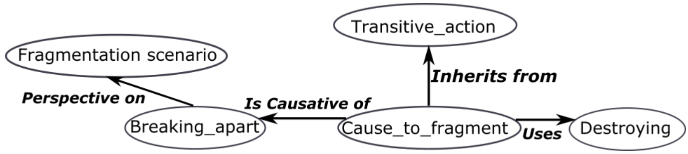


Figure 2.5.: Frame-to-frame relations

connected by the ***Causative_of*** relation to the inchoative BREAKING_APART frame, shown below in Example 15:

(15) a. [Mark $_{\text{AGENT}}$] [*broke* $_{\text{CAUSE\_TO\_FRAGMENT}}$] [the windscreen $_{\text{WHOLE\_PATIENT}}$] [into pieces $_{\text{PIECES}}$].

b. [The windscreen $_{\text{WHOLE}}$] [*broke* $_{\text{BREAKING\_APART}}$] [into pieces $_{\text{PIECES}}$] because of Mark.

Example 15a shows that the FE AGENT is the subject of the CAUSE_TO_FRAGMENT frame, but the inchoative form in Example 15b treats the AGENT as an oblique. These frame-to-frame relations not only connect closely related frames, they demonstrate how frames could conceptually relate to one another in context.

In sum, the FrameNet lexicon is composed of several key structures: the frame (lexical vs. non-lexical), its lexical units, frame elements, semantic types, and frame-to-frame relations. The current version of FrameNet defines over 1200 frames with a majority (1075) being lexical, and over 10k frame elements. On average, frames have approximately 12.7 LUs, and FrameNet has over 13k LUs in the entire database [1]. The work we will describe in later chapters uses these defined frames and their LUs to model frames across languages, as well as frame-to-frame relations for resolution of poor frame transfer.

[1]https://framenet.icsi.berkeley.edu/fndrupal/current_status

Annotations

FrameNet provides full text annotations for training supervised machine learning applications. Annotations were added to the 100-million word British National Corpus (BNC), which is a balanced corpus of text genres, from novels to newswire [2]. To date, several versions of the resource have been released, and the two releases used most prominently as training for standard frame semantic parsers are FrameNet 1.5 and 1.7 annotations.

In general, researchers developing a frame semantic resource can adopt two broad methods for frame annotation, 1) frame-by-frame annotations, where the goal is to create a complete coverage of frames in the lexicon, and 2) lemma-by-lemma annotation, which attempts to cover all senses of a specific set of lemmas that are encountered in running text (Djemaa et al., 2016). FrameNet provides annotations of each frame in corpora via exemplar sentences, following a strict frame-by-frame approach. The full text annotation provides frame annotations over continuous text where lemmas that are not already defined as part of an existing frame structure are disregarded. For completeness, all frames that appear in a sentence are annotated accordingly, meaning several sentences can have annotations of multiple frames (see Figure 2.6).

FrameNet's 1.5 full text annotations cover over 23k sets of frame annotations, with over 11k lexical units. All lexical units appear on average 20.7 times in the full text annotations.

2.4.2. FrameNets in Other Languages

Following the popularity of the Berkeley FrameNet project, several efforts emerged to construct frame semantic lexicons and, in certain cases, frame-

[2] http://www.natcorp.ox.ac.uk/

Figure 2.6.: Example sentence from FrameNet's full text annotations. A single sentence can evoke multiple frames, such as the LUs *Massive* and *smashing*, which evoke the Dimension and Cause_to_fragment frames, respectively.

annotated corpora for different languages. Thus far, these projects have ranged across typologically related and unrelated languages, including: German, French, Chinese, Dutch, Latvian, Finnish, Danish, Hebrew, Hindi, Japanese, Italian, and Korean (Baker and Lorenzi, 2020). While these projects have illustrated the challenges and successes in creating frame semantic lexicons for other languages, relatively few eventually produced annotated sets that approached the size of the English FrameNet corpus. Consequently, only a subset of these languages have demonstrated the applicability of their resources for frame semantic parsing in a different language (Chinese (Li, Wang, and Gao, 2010), Swedish (Johansson, Friberg Heppin, and Kokkinakis, 2012), German (Burchardt et al., 2006), and French (Michalon et al., 2016)). Other annotated resources that led to language-specific frame semantic parsers were constructed only for certain domains, such as the FrameNet Brasil project, in which annotated frames specifically related to tourism and sports (Costa et al., 2018). Most of these frames were not covered in the English version, meaning they could not be compared across the language pair.

Recently, multiple teams of researchers in FrameNets for different languages have begun frame semantic annotations over translated text. The Multilingual FrameNet project brings together different teams from across the globe to

	#	# of instances			similarity w/EN		
T Lang	frames	Train	Dev	Test	same	modified	unaligned
Berkeley FrameNet (EN)	1020	15044	4434	4458	-	-	-
SALSA Project (DE)	1001	26070	5530	5659	234	37	730
French FrameNet (FR)	105	16961	1732	2941	46	22	37

Table 2.1.: Size of frame-semantic resources across FrameNet projects for English (EN), German (DE), and French (FR). *# frames* is the number of frames defined in the corresponding frame lexicons, where the size of the English frame inventory is roughly similar in size to the number of German frames, but French has approximately one tenth of the number of frames compared to the English and German lexicons. Size of the annotated data (*# of instances*) is similar across languages. Finally, we can compare the similarity of the frame definitions in the German and French frames to definitions in the English lexicon (*similarity w/EN*), where definitions can be the same as English ("same"), modified for German and French ("modified"), or frames that are language-specific ("unaligned").

annotate translations of the TED talk, "Do schools kill creativity?". When completed, this work promises to lead to a more comprehensive understanding of how frames transfer across a parallel, translated text and the universal status of frames in general (Torrent et al., 2018).

The work described in this thesis features two of the languages, German and French, which have the following: pre-existing frame semantic parsers (for system comparison); frame semantic annotations; and a frame inventory that has frames which can be compared to the English FrameNet.

German FrameNet: The SALSA Project

One of the earliest attempts to adapt frames for another language was the SALSA project (Burchardt et al., 2006; Burchardt et al., 2009), a German frame semantic resource developed in Saarbrücken, Germany. Annotations for frames and their frame elements were added to an existing resource called the TIGER treebank (Brants et al., 2002), a syntactically parsed corpus of German news. Although the project initially began with the English FrameNet v1.2 and v1.3 frame inventory, several cross-lingual divergences between German and English quickly led to (sometimes significant) modification of frames for German data. Overall, there are 250 frames that come from FrameNet v1.3 and 5 taken from FrameNet v1.2, making the FrameNet-based frames roughly a quarter of the total subset of frames in the German SALSA lexicon. In contrast to the frame-by-frame annotation approach in the Berkeley FrameNet project, the SALSA project annotated its German corpus in a lemma-by-lemma strategy. This led to the creation of several so-called “proto-frames”: frames which covered senses of a lemma that were not available as frames in the existing frame inventory. These proto-frames are represented in the database as the lemma name and the sense number of the encountered instance; for example, the lexical unit “person” in German can refer specifically to the character portrayed in an art piece:

(16) Die [*Person* PERSON2-SALSA] des [Faust PERSON] gibt uns einige
The person of Faust gives us a few
Aufschlüsse über Goethes Verhältnis zur Religion und zu
clues about Goethe’s relationship to religion and to
ethischen Fragen.
ethical questions.

“The Faust character gives us insight into Goethe’s relationship with religion and questions of ethics.”

The *person* LU in Example 16 evokes the Person2-salsa proto-frame as the specific sense of the word is not available as an existing frame.

In terms of their annotation, the SALSA release predominately annotates these FrameNet-based frames: out of over 37k annotations, over 22k are from this group. Frames that have been altered from their original English definitions are a relatively small subset of the lexicon, where only 37 frames fall into this category. Similarly, this group represents the lowest number of annotations in the data. Currently, the resource has 37k annotated instances over 24k sentences with both verbal and nominal LUs (Rehbein et al., 2016).

Proto-frames (language-specific frames,) make up a majority of the frames in the SALSA lexicon, with 730 of the total frames belonging to this category. Over 12k LU annotations were for proto-frames, meaning they are the second largest category of annotated data in the SALSA v2.0 release. The distribution of annotations for each of these frame types suggests that a predominate number of the total annotations have frames that are on one end of the lexicographic comparability spectrum or the other: either they are frames whose definition is a complete match across English and German, or they are frames whose definition is language-specific and therefore cannot be compared. Table 2.1 shows the number of frames, annotations, and frame definitions in the SALSA project, where the proto-frames are in the "unaligned" category.

French FrameNet: the ASFALDA project

French FrameNet is one of the more recent frame semantic resources that has enabled the construction of a language-specific frame semantic parser (Michalon et al., 2016). Similar to the German SALSA project, the ASFALDA project (Djemaa et al., 2016) adapted frames from the English inventory and added frame semantic annotations to resources with existing syntactic annotation. For French, these included the Sequoia and the French treebanks (Marie and Djamé, 2012; Anne and Nicolas, 2004). Unlike SALSA, however, the French annotations proceeded in a frame-by-frame fashion, focusing effort on FrameNet frames that come from four hand-selected, specific domains: commercial transactions, verbal communication, cognitive positions, and causality. Once these domains were selected, all English frames that fit the domains of interest were compiled and analyzed for their suitability for the French language, leading several frames to be merged, split, or modified (Marie et al., 2014). As shown in Table 2.1, the French FrameNet project has a much smaller number of defined frames compared to the Berkeley FrameNet and SALSA frame lexicons, but the size of the annotated data is comparable to the English. French frames that are counted as "unaligned" to the English FrameNet were those whose definitions merged multiple frames from English, resulting in frames whose definitions were not readily alignable to a specific English frame. Authors of the resource focused on more comprehensive coverage of LUs, and therefore chose to restrict the number of annotations for any given lemma to 100.

3. Technical Background

The goal of computational semantics is to have a system that can automatically extract meaning from text. This meaning can be as wide-ranging as word sense disambiguation, anaphora, or inference and reasoning. Frame semantics has become well-known in computational semantics, where the goal is to produce a semantic representation of text.

This chapter describes work in computational semantics that is relevant to frame semantics and representations of frames, setting the stage for later work in frame representations that can be compared cross-lingually. We will cover a standard technique for automatically measuring word meaning in text – **embeddings** – in Section 3.1. This technique will be used later in modern systems that detect frames in text (**frame identification**,) which will be described in Section 3.2.

3.1. Embeddings

One of the ways to computationally represent the meaning of a word is by considering its context; that is, other words that it co-occurs with. The idea of representing word meaning as a measurement of its relationship to its neighbors is based on the *distributional hypothesis*, where words that occur in similar context have similar meaning (Harris, 1954). Based on this hypothesis, **word**

embeddings are vectorized representations of a word, where words that have a similarity in meaning are close in vector space. More specifically: a word embedding represents single word w in a vocabulary of words W as a vector of D dimensions: $\vec{w} = [x_1...x_D]$, where x is a single association feature – in the simplest case, a co-occurring word.

Word embeddings have been shown to capture semantic relationships, including ontological relationships such as *is-a* or basic linguistic relationships such as possessive/non-possessive or singular/plural (Turney and Pantel, 2010; Mikolov, Yih, and Zweig, 2013). Prior work has established that embeddings can be quite adept at capturing a general "semantic relatedness", such as *doctor – needle* or more specific "semantic similarity" relation, such as *doctor – physician* (Budanitsky and Hirst, 2006). Because they capture different types of semantic relationships, embeddings have been used as the basis for state-of-the-art performance in tasks such as word sense disambiguation, information retrieval, or automatic thesaurus generation (Manning, Schütze, and Raghavan, 2008; Yuret and Yatbaz, 2010; Iacobacci, Pilehvar, and Navigli, 2015; Curran and Moens, 2002).

Although embeddings in NLP tend to refer to word representations, an embedding can be computed over any pattern in a dataset. Phrases or entire sentences can be represented as vectors, as well as image data or even gene sequences (Arora, Liang, and Ma, 2016; Akata et al., 2015; Du et al., 2019). For this reason, it has been proposed that the distributional hypothesis be extended s.th. not only words, but *patterns* that occur in similar contexts will have similar meanings (Lin and Pantel, 2001). This thesis will demonstrate that frames can be seen as one of these patterns, where lexical units that occur in similar contexts will belong to a similar frame, and more broadly, frames that occur in similar contexts will be conceptually related. By learning

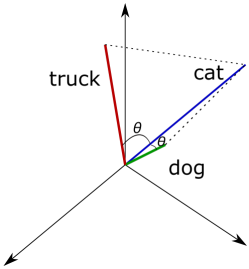


Figure 3.1.: Cosine similarity ($cos(\theta)$) and euclidean distance (dotted lines).

representations based on co-occurrence in the data, it is not necessary to explicitly define features; instead, features are learned automatically in the training process.

Much of the work in word embeddings involves generating embeddings in an unsupervised fashion over a very large corpus, and annotated data is not necessary in building a vectorized representation of a word.

Measuring Similarity in Vector Space

As vectors in high dimensional space, a semantic relationship between two words can be computed by measuring their geometric distance. The most standard similarity measure for word embeddings is the **cosine similarity** metric, which is a normalized inner product of the two word vectors, w_1 and w_2:

$$sim(w_1, w_2) = cos(\theta) = \frac{w_1 \cdot w_2}{\|w_1\| \times \|w_2\|}$$

Words that are near in vector space, that is, words that are thought to be more similar, will have a cosine similarity value close to 1 while words that are dissimilar will have a value near -1. Cosine similarity is the standard metric to compute word similarity in vector space, although one could also measure vector similarity as the euclidean distance between two word embeddings. With euclidean distance, the lower the distance between the vectors, the higher their similarity will be; for example, similarity would be the inverse of the distance ($sim(w_1, w_2) = 1/dist(w_1, w_2)$) where

$$dist(w_1, w_2) = \sqrt{\sum_{i=1}^{D} (w_{1_i} - w_{2_i})^2}$$

One shortcoming to euclidean distance is shown in In Figure 3.1, where frequency can affect the length of a vector, and vector length can impact measurements of similarity. For this reason, cosine similarity is the standard metric for determining similarity in vector space.

Visualization Once an embedding is learned, word vectors can be plotted in a low-dimensional space to visualize word relationships. A commonly used technique for visualization of word embeddings is t-SNE, which measures the similarity of vectors as probability distributions and then minimizes the KL-divergence between similar probability distributions (Van Der Maaten, 2014). The results show semantically similar words as close together and dissimilar words far away in a low-dimensional plot:

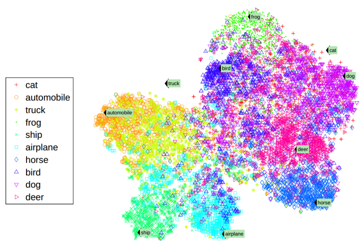


Figure 3.2.: tSNE visualization of word relations ((Socher et al., 2013))

While visualization of word embeddings shows clearly the types of semantic relationships that are learned, a more precise measurement of similarity between two words can be achieved using measurements of geometric distance described above.

Learning Embeddings

Vector space models have been used in NLP for several decades, and there are many non-neural network based approaches to constructing a vectorized word representations. These approaches involve decomposition over large matrices with dense matrix multiplication operations (Salton, 1971; Blei, Ng, and Jordan, 2003; Deerwester et al., 1990). In recent years, the predominant approach to learning word embeddings is via training in a neural network, as this method has been shown to be more effective in learning representations. There are many possible neural embedding architectures that can be used for learning embeddings, where embeddings are shaped by the objective function that is defined over the network. As we will describe below, it is less common in recent years for embeddings to be trained as an end product in and of themselves. Instead, embeddings are often trained as part of end-to-end systems, or used in a pre-training and fine-tuning setup where embeddings are first learned over large, unstructured corpora and then tuned with supervision over an annotated corpus for a variety of downstream tasks.

The work in this thesis uses two common approaches to learning embeddings: representations learned from **Word2Vec** (Tomas et al., 2013; Mikolov et al., 2013) and those learned from **BERT** (Devlin et al., 2019).

Word2Vec Word2Vec is a simple, 2-layer neural network in which a word representation is learned by either a) making predictions for target words given information about their context (CBOW architecture), or b) making predictions for context words given information about the target word (Skipgram architecture). Over a large corpus of unstructured text, every unique word in the vocabulary (W) is associated with a single dimension, and each input word vector has W dimensions. Input words are therefore represented as one-hot vectors where the vector is the size of W. In the CBOW variant, which we use in later chapters for our models of frames, the network then learns to predict a target word from words in the surrounding context. More formally, when adopting the CBOW architecture, the training objective (J_θ) is to maximize the log probability of a target word, w_t given its surrounding context, w_{t+j}, over a sequence of words T (Tomas et al., 2013):

$$J_\theta = \frac{1}{T} \sum_{t=1}^{T} \sum_{-c \leq j \leq c} \log p(w_t | w_{t+j})$$

where c is the size of the context window. One way to estimate $p(w_t|w_{t+j})$ is via the softmax function, where the probability can be computed over a single pair (w_t, w_j) of a target word w_t and a context word w_j:

$$p(w_t|w_j) = \frac{e^{(\vec{w}_t \cdot \vec{w}_j)}}{\sum_{w=1}^{W} e^{(\vec{w}_t \cdot \vec{w}_j)}}$$

However, computation of this softmax for all target words quickly becomes computationally intractable. For this reason, alternative approaches to learning $p(w_t|w_{t+j})$ include a hierarchical softmax, which is a computationally efficient way to define the softmax function, or via negative sampling.

There are many implementations of the Word2Vec model available, and at the time of its debut, word embeddings from the model achieved impressive results in inference and entailment problems. However, one of the primary shortcomings of the approach is the bag-of-words nature of a word's contextual features. In learning, there is no indication of the position of a contextual word in relation to the target. Another shortcoming of the approach is that, when training a vanilla Word2Vec model, word sense is conflated when word forms are identical. The model learns one single representation for any given token, so *she called him on the phone* and *she called him a thief* would conflate both senses of *call* into a single vector.

BERT Recent approaches to learning word representations address both of the shortcomings described above. Instead of learning word embeddings over a shallow network, recent models learn embeddings in a uni- or bi-directional language model (Peters et al., 2018; Alec et al., 2018), thus accounting for the position of the surrounding words in training a target word representation. The model additionally generates representations for each target word instance in the corpus, meaning the word representations are contextualized; *they bought a car* and *they bought his excuse* would have two different vector representations for the lemma *buy*.

The most popular of these approaches is the Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2019). BERT uses an encoder-decoder architecture, where the encoder converts and input sequence into a continuous representation, and the decoder generates the output sequence from the representation produced by the encoder (Cho et al., 2014; Sutskever, Vinyals, and Le, 2014). The system is composed of stacked, self-attention mechanisms called a Transformer (Vaswani et al., 2017).

Unlike the Word2Vec model, the BERT system also enables researchers to fine-tune embeddings for downstream tasks. BERT learns embeddings in two phases: the pre-training phase, where embeddings are built from large, unstructured corpora, and the fine-tuning phase, where the pre-trained embeddings are later adapted for specific NLP tasks with a small set of annotated corpora. To learn their pre-trained embeddings, the BERT language model masks the target word w_t in a sentence, and the objective of the model is to predict the original word given the seen context. BERT additionally uses a 'next sentence prediction' training, which is a binary task where the goal is to predict whether a sentence follows a target sentence. In the masking learning objective, input is a sentence with a percentage of tokens masked, and the hidden embedding for the masked token is fed to a final softmax layer. For sentence prediction training, authors randomly select negative sentence pairs that do not occur sequentially versus positive pairs. The loss function is then the sum of the mean token masking likelihood with the mean sentence prediction likelihood.

Fine-tuning a model is then straightforward, where parameters from the pre-trained embeddings are input and tuned for the defined task. At the time of its release, the BERT model demonstrated state-of-the-art results when fine-tuned for wide-ranging NLP tasks such as question answering, textual similarity, paraphrase detection, and sentiment analysis.

3.2. Computational Models of Frames

In NLP, **frame semantic parsing** is a task in which systems automatically detect frames and their semantic roles in text. Over the last several years, frame semantic parsing has become a well-known task, as it is shown to improve several downstream applications such as question answering and information retrieval (Shen and Lapata, 2007; Surdeanu et al., 2003; Chen, Wang, and Rudnicky, 2013; Søgaard, Plank, and Alonso, 2015).

A full semantic parsing system consists of three subtasks – target identification, frame identification, and semantic role labeling, each of which can be studied as a standalone task as they all pose their own specific challenges. First, **target identification** is the automatic detection of the predicates that evoke a frame in text. FrameNet has predicates of all parts-of-speech (POS), so a simple POS tagger is often not sufficient to detect frame-evoking predicates. Therefore, the standard approach to target identification involves using a small set of rules over syntactic parses that indicate a word is frame-evoking (Johansson and Nugues, 2007), with some researchers using training data to further refine these rules based on training data targets and lexical units in the FrameNet lexicon (Das et al., 2010a). Once the targets have been identified, their sense must be disambiguated to the appropriate FrameNet frame. A predicate can potentially evoke a single frame or several (sometimes related) frames, making the task substantial for certain predicates and straightforward for others. For instance, the verb *capitulate* evokes only one frame (Giving_in), while the verb *call* can evoke nine frames, several of which have close relationships to one another (i.e., Referring_by_name, Labeling, Simple_naming, Communication_means, and Name_conferral). In text, the disambiguation of an predicate to a frame is the subtask of **frame identification**, also called **frame assignment**. While target identification can rely heavily on syntactic parses, frame identification is much more semantic in nature, as the task is analogous to other word sense disambiguation efforts where the senses are the frames (Erk, 2005). The final step in a full semantic parse is the identification of the frame's semantic roles in text, a well-known task referred to as **semantic role labeling**. While prior systems used lexical syntactic features to train classifiers to label semantic roles, recent neural net-based models have included word representations learned from a network to achieve state-of-the-art results in frame-based semantic role labeling (Swayamdipta et al., 2017; Roth and Lapata, 2015).

3.2.1. Frame Identification

This thesis is particularly interested in the semantics of frames and how frame meaning is represented cross-lingually. We restrict our focus to the frame identification task, as this sub-task entails the construction of a (implicit or explicit) frame representation in order to successfully disambiguate a predicate to the appropriate frame. Focusing on this sub-task is a more straightforward way to construct and compare frame representations across languages, which we will describe in future chapters.

Frame identification systems face several critical design decisions. First, some approaches to frame identification assume the availability of a full frame lexicon; that is, for each target predicate, there is a resource that has a clearly defined subset of frames that the predicate can evoke. In these cases, frame identification is essentially a word sense disambiguation (WSD) task. However, that does not necessarily have to be the case. Some approaches to frame identification have treated the extraction of frame candidates for a target predicate as an additional parameter to learn from a subset of training data (Das et al., 2010b), or alternatively, many systems report on a so-called **no-lexicon** condition where the subset of frame candidates is not assumed to be available for a target predicate. This choice adds complexity to the disambiguation task, but it is a more realistic setting – it can potentially reveal knowledge gaps in the existing resource, and it can also apply to languages that lack a well-developed frame lexicon.

Second, one of the fundamental assumptions in frame semantics is that frames are open class categories; that is, even well-established frame lexicons are in no sense "complete". Therefore, systems could be constructed or evaluated in such a way that one can generalize about the existence of new frames. While attempts to automatically expand the frame lexicon have been addressed (Pavlick et al., 2015; Ferrández et al., 2010), some systems of frame identification report over the so-called **unseen** frame condition where a frame in the test data is not seen during training time.

In general, there are two factors in play within the frame identification task. On the one hand, there is the problem of *ambiguity*, where a predicate is ambiguous in terms of the frame it evokes, and on the other hand, there is the problem of *generalization*, where a system must cope with frames as open classes and frame lexicons as flexible and ever-expanding entities.

3.2.2. Prior Work

Frame Identification

There have been many approaches to frame identification over the years. Prior to neural network-based models, Das et al., 2010b used an extensive list of lexical and syntactic features in predicting a frame for a target predicate. Authors of this model construct frame prototypes from seen instances of frames in the training data, represented in their model as a latent variable, l, which is a sum over the frame instances. Over 600k binary features were used, including lemmas that appear in training, WordNet relations, POS tags, and syntactic dependencies. These features and their weights are shared across target predicates, frame instances, and prototypes in a conditional log-linear model. During prediction, an instance of a predicate is assigned the frame with the most probable prototype from a list of frame candidates. In earlier work by Johansson and Nugues, 2007, authors used a much shorter list of syntactic and semantic features to classify a predicate as evoking a frame in a SVM classifier.

Current approaches to frame identification involve learning frame embeddings to predict a predicate's frame given its context. The first of such approaches was described in Hermann et al., 2014, where authors used embeddings of predicates and their syntactic context to predict frames. Pretrained embeddings of target predicates and words in their syntactic context are mapped to a low dimensional space using the WSABIE algorithm, which simultaneously produces embeddings in the same latent space for frames (i.e., the labels used for classification). The objective of the model minimizes distance between the embedding of the predicate and the embedding of the frame, where at test time, the frame closest to the test predicate is chosen.

Instead of using matrix factorization for frame embeddings, later models constructed frame embeddings for classification using neural network. The most straightforward approach uses Word2Vec to learn embeddings, where frame labels replace the predicates which evoke them in text (Botschen, Mousselly-Sergieh, and Gurevych, 2017). The frame embeddings from this model were then used in predicting relationships between frames, but authors demonstrate that they can be used in disambiguating a predicate to its frame as well. Botschen et al., 2018 showed further improvement in frame identification for English and German when predicate embeddings are concatenated with multimodal data as hidden layers in a Multilayer Perceptron. Frames were predicted in the final output layer of the classifier.

Finally, while these systems addressed frame identification as a standalone task, Swayamdipta et al., 2017 used joint decoding with semantic role labeling where parameters were shared across the two sub-tasks. Authors use a segmental RNN architecture with a softmax margin criterion which favors recall of arguments in text. Altogether, this architecture allows them to forgo syntactic parsing as a prior step to extracting arguments. Although they achieve improved performance in argument labeling, frame identification results were still outperformed by the Hermann et al., 2014 model.

Finally, language model pre-training is the latest development in frame identification that has surpassed prior results. Tan and Na, 2019 fine-tune the BERT model with position-based attention mechanisms for frame-evoking predicates. Authors of this model use the English pre-trained model, which was trained over the BooksCorpus and English Wikipedia. For the frame identification task, they compute an additional hidden state with vectors for the target predicate and its context. The context vector uses a local attention mechanism to weigh input from earlier hidden states, where a fixed window of context, the hidden states before and after the target, is used to compute alignment scores (i.e., how much each hidden state contributes to prediction). Results surpassed the Hermann et al., 2014 model, demonstrating the efficacy of the BERT model for the frame identification task.

Frame Semantic Parsing for Different Languages

In cross-lingual frame semantic parsing and related work, much of the focus is on the semantic role identification and labeling. In earlier approaches, annotations were projected across syntactic parses to generate cross-lingual training data (Plas, Merlo, and Henderson, 2011; Padó, 2007). This work was followed by similar projection-based methods which used a filtered projection and bootstrapping for automatically generating semantic role annotations for different languages (Akbik et al., 2015; Akbik and Li, 2016). Alternatively, cross-lingual model transfer used shared, multilingual feature representations to train a model over source language annotations and apply it directly to another without the need for parallel data (Kozhevnikov and Titov, 2013). Most recently, machine translation has been proposed as a means to generate cross-lingual semantic role labels, where cross-lingual word representations can be used to train an end-to-end semantic role labeler for different languages

(McDonald et al., 2013; Daza and Frank, 2019a; Fei, Zhang, and Ji, 2020).

The above models were designed and tested across different language pairs, where typically annotations are only assumed in one source language and then projected or "translated" onto a target. However, there have also been several efforts to train a single model for simultaneously labeling multiple languages. (Johannsen, Alonso, and Søgaard, 2015) perform weakly supervised frame semantic parsers for nine languages, where authors first use bilingual dictionaries to construct small, frame-annotated datasets for nine languages. Authors then perform frame identification with multilingual word embeddings and universal dependencies to train a single model for all languages, demonstrating improved performance when using multilingual training. Similar gains were shown for experiments in semantic role labeling, where a single multilingual system outperformed individual, monolingual models (Mulcaire, Swayamdipta, and Smith, 2018).

3.3. Frame Comparability in NLP

Understanding how frames compare across languages is important from both a theoretical standpoint – namely, having a more complete picture of how 'universal' certain frames might be, as well as a computational standpoint. As described in Section 3.2, frame semantic parsing clearly offers benefits to a range of NLP applications. These applications in different languages would likewise benefit from frame semantic parsers, although English is the language that is predominately studied since it is one of the few with significant annotations and an established frame lexicon.

To build a frame semantic parser for a new language, researchers have conventionally used the techniques described above (Section 3.2.2). In the projection approach, frames and their semantic roles are projected across parallel data, and the cases where parallelism breaks down can cause errors in the transfer of the frame semantic annotations Padó, 2007; Marzinotto, 2020). This especially the case in translation shifts and paraphrasing, both of which have low usage-based frame comparability and cause errors in the projection. Figure 3.3 shows the case of English and French sentences where the English COGNIZER role is missing in French due to translation shift (example modified from (Akbik et al., 2015)):

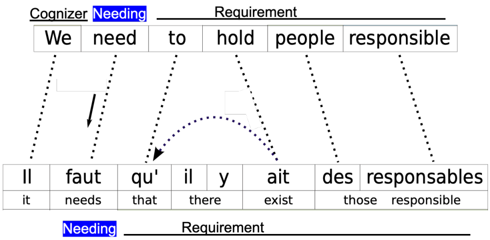


Figure 3.3.: Semantic role alignment across English-French translations.

Additional cases of poor usage-based comparability are described in Section 2.3, where most cases can cause problems for the projection approach. Alternative approaches involve model transfer for cross-lingual SRL, where no parallel data is assumed, but rather cross-lingual word representations, bilingual dictionaries, and/or parsers for the target language are used to train a model in one language and apply it to another (Kozhevnikov and Titov, 2013). However, if language-specific constraints are integrated into the cross-lingual transfer, such as constraints regarding the valency of predicates (Chang, Ratinov, and Roth, 2007), this can introduce errors from poor lexicographic comparability, where the potential arguments that are expected for a frame do not match across languages.

The most resource-intensive approach, (but generally ideal for the target language,) requires an investment on the part of the linguist; a new frame lexicon and frame semantic annotations can be created for a target language, which can be subsequently used for training a machine learning model on the curated data (Johansson, Friberg Heppin, and Kokkinakis, 2012; Michalon et al., 2016). The creation of frame annotations requires significant time and financial commitment on the part of the lexicographer. Most teams interested in obtaining a frame semantic resource for a new language will face the challenge of funding, where often there is not enough resources to build datasets with sufficient size and frame coverage for training a frame semantic parser. Because many frames have been attested across different language pairs, using frame definitions or annotations from another language could help to drastically reduce this requirement. It is for this reason that teams looking to create new frame semantic resources for their target language often use frame definitions from other languages as a starting point for creating their own. Efforts by the lexicographer looking to create a new frame semantic resource can be mitigated if there is existing evidence of frames that can be applied directly from a previously studied language.

In sum, assessments of cross-lingual frame comparability are crucial in the development of NLP techniques for multilingual frame semantic parsing. Most modern techniques for cross-lingual frame semantic parsing either a) rely on the assumption of frame equivalence across language pairs, or b) they require significant, potentially difficult to obtain, resources on behalf of the lexicographer. Computational approaches that can mitigate the workload of b) will help to reduce the effort of the lexicographer, as well as facilitate the development of systems that can adapt to non-parallel frame structures in translated text. In this thesis, we will later describe a new technique for frame semantic models in different languages, specifically, models based on Word2Vec and BERT embeddings to construct and compare representations of frames across languages.

Part II.

Computational Measures of Frame Comparability

4. Aligning Computational Representations of Frames across Languages

Linguists tend to study a frame across languages by evaluating a handful of frames or specific instance of frames at a time. While this practice leads to many important insights in the field (which we will discuss in Section 4.1), it also reinforces the need to complement these studies with computational systems, where a computational system can quantitatively assess frame comparability by computing how similar all instances of a frame in one language are against all instances of that frame in another. While this is feasible in a computational approach, such an endeavor would be too labor-intensive to accomplish by hand.

This chapter describes our approach in constructing and comparing computational representations of frames across languages, addressing the question of how frame comparability can be measured computationally. There are several ways that one might assess frame comparability; our approach is the first to directly compare frames across languages based on data originally written in those languages – distinguishing our work from comparisons based on translated data (Padó and Lapata, 2005; Tonelli and Pianta, 2008; Spreyer and

Frank, 2008), which is prone to effects of translationese (Gellerstam, 1986; Koppel and Ordan, 2011). We use embeddings (described in Chapter 3) to learn a computational representation of a frame based on its usage in corpora. We then assess the comparability of that frame via the similarity of its representation to its counterpart in the other language.

We measure frame comparability across two languages, English and German, which have frame-annotated corpora of similar sizes and a relatively robust body of literature in lexicographic comparison of frames across the language pair (Boas, 2009; Boas and Dux, 2013; Boas, Dux, and Ziem, 2016). There are several linguistics studies which discuss the comparability of frames in English and German, further motivating our selection of these languages; these linguistic studies enable us to compare our quantitative measure with qualitative assessments of frame comparability. Our goal in this chapter is to demonstrate how quantitative measures of frame similarity, which are arrived at computationally, can complement or support the qualitative measures of frame comparability that can be found by linguistic analysis. Much of the work described in this chapter was presented in a publication "Using Embeddings to Compare FrameNet Frames across Languages" at the 2018 COLING workshop on Linguistic Resources for Natural Language Processing (Sikos and Padó, 2018b). We have extended many of the analyses and experiments for this chapter.

This chapter is structured as follows. In Section 4.2.1, we first describe a background of linguistics studies which compare frames across the languages of interest, English and German. Next, we describe a background on frame embeddings (Section 4.2.1) and alignment of frames across languages (Section 4.2.2). We then proceed to our experiments. In Section 4.3.1, we sketch the annotated data we use for both experiments, along with the pre-processing steps for both English and German full-text FrameNet corpora. In Experiment 1 (Section 4.3.2), we describe methods we use to construct frame embeddings and evaluate the quality those embeddings. Experiment 2 is the alignment of frame embeddings across languages and the analysis of frame comparability using embeddings. This analysis involves a comparison of the linguistic, qualitative frame comparability discussed in linguistics literature with our quantitative, embedding-based measure of frame comparability. We then summarize our findings and conclude in Section 4.5.

4.1. Linguistic Studies in Frame Comparability for Frames in English and German

Some of the first work in cross-lingual frame semantics introduced linguistic analyses of frames across English and German (Boas, 2001; Burchardt et al., 2006). Much of these analyses adopted a strategy of translating, or analyzing translations of, sentences which are known to evoke an English frame and discuss their equivalent in German (Boas, 2020a). Below, we describe several studies about frames across English and German, which yielded a number of hypotheses we can test in later quantitative analysis.

Motion and Communication Frames In Chapter 2, we discussed linguistic analyses of frames across languages where there are divergences in lexicalization, semantic roles, valency, or cultural differences that appear in the definition of the frame. There are, however, specific frames that have a high cross-lingual comparability, where these frames are often speculated to have a more "universal" status (Baker and Lorenzi, 2020; Padó, 2007). These are generally frames that are more coarse-grained and relate to everyday, shared human experiences that are less defined by cultural or country-specific conventions (such as judicial processes, for instance). The MOTION and COMMUNICATION frames have been previously discussed (Boas, 2020b), as they are coarse-grained and more likely to be found across cultures; that is, they are thought to have a decent "cultural equivalence" (Boas, 2020b). We would therefore hypothesize that the MOTION and COMMUNICATION frames have a higher cross-lingual correspondence than other, more specific frames in the shared embedding space, indicating a high cross-lingual applicability.

Sub-frames Relating to Motion We described in Chapter 2 the divergences in frames that can occur due to differences in lexicalization patterns, and it has been previously observed that causation can be lexicalized differently for certain verbs in English and German (Boas, 2013). An example of this is below, where the English sentence evokes the SELF_MOTION frame, and the German translation evokes the SETTING_OUT frame due to differences in how the manner of motion is lexicalized (Boas, 2013):

(17) ...we started to [*walk* SELF_MOTION] [to Merripit House GOAL].

(18) ...wir uns [*zu Fuss* MEANS_OF_MOTION] in [Richtung DIRECTION] auf
...we us [*by foot*] in [direction] of
Merripit House [*aufmachen* SETTING_OUT].
Merripit House [*go on*].

In Example 17, the English verb *walk* incorporates the manner of motion, evoking the SELF_MOTION frame, whereas the German translation in Example 18 shows *aufmachen* (*go on*) describing the motion, evoking the SETTING_OUT frame, and *zu Fuss* (*by foot*) describing manner and evoking a separate frame, MEANS_OF_MOTION. In this case, there is a mismatch in the frames evoked across the translations; however, the SETTING_OUT frame is related to the SELF_MOTION frame in the FrameNet hierarchy as they are both frames which relate to the abstract MOTION frame.

Additional examples of English and German translations show how the same motion frame can be expressed differently across the language pair (Boas, 2001):

(19) Tina [*ran* $_{\text{CAUSE\_MOTION}}$] Enno off the street.

(20) Tina [*drängte* $_{\text{CAUSE\_MOTION}}$] Enno (beim Rennen) von der Strasse.
Tina [*pushed aside*] Enno (by running) from the street.

The German translation in Example 19 expresses the sense of English *run* as the predicate *abdrängen* (*push aside*), while the manner of motion is expressed externally (*beim Rennen/by running*).

If it is the case that such lexicalization differences occur frequently in English and German text, we would expect that the English and German sub-frames for motion events will diverge. In a computational model, we would hypothesize that MOTION sub-frames will have a lower similarity in the embedding space to their parent MOTION frame, indicating a lower comparability.

Statement frame Although the COMMUNICATION frame is thought to be more universal, there are noted differences in the lexical units of its sub-frame, STATEMENT, across English and German (Boas, 2002). In English, the STATEMENT frame is about a SPEAKER communicating a MESSAGE and has the semantic roles SPEAKER, ADDRESSEE, MESSAGE, and MEDIUM. Its frame-evoking predicates are capable of expressing either a SPEAKER, which is a salient entity, or MEDIUM, the channel through which the message can be communicated. (Boas, 2002) discusses the case where both the SPEAKER and MEDIUM can fill the subject position in English or they can be present simultaneously in the same sentence, as in Examples 21-23:

(21) [The CEO $_{\text{SPEAKER}}$] [*announced* $_{\text{STATEMENT}}$] [that the company would be acquired $_{\text{MESSAGE}}$].

(22) [The press report $_{\text{MEDIUM}}$] [*announced* $_{\text{STATEMENT}}$] [that the company would be acquired $_{\text{MESSAGE}}$].

(23) [The CEO $_{\text{SPEAKER}}$] [*announced* $_{\text{STATEMENT}}$] [that the company would be acquired $_{\text{MESSAGE}}$] [by email $_{\text{MEDIUM}}$].

In German, there are specific predicates that can be translated for each of these alternations from the STATEMENT frame, where certain predicates can appear in some of the alternations in Examples 21-23, but not others. Example 21 can be translated with the predicates *bekanntgeben, bekanntmachen, ankündigen*, and *anzeigen*; Example 22 is translated with predicates *bekanntgeben, ankündigen*, and *anzeigen*; and Example 23 is translated with *ankündigen, ansagen*, and *durchsagen* (Boas, 2002). The only lexical unit in German that can express all constructions in Examples 21-23 is *ankündigen.* If it is true that our embeddings are capturing some semantic intuitions about frames and lexical units, we can hypothesize that the *ankündigen.v* lexical unit in German should fit well with the STATEMENT frame as it is defined in English.

4.2. Background

4.2.1. Frames and LU Embeddings

Before continuous representations were popularized, work in computing representations of frames involved a set of hand-engineered features including the POS tag of the predicate's head word, the syntactic dependency label of its parent, and WordNet relations (Das et al., 2014). Hermann et al., 2014 was the first to propose an end-to-end representation learning approach to the frame identification task (described in Chapter 3) where frame embeddings were learned as part of the model. These continuous representations of a frame - **frame embeddings**, are a compact representation of a frame's meaning, which are built over instances of that frame in a corpus. Properties of the frame, such as relationships between lexical units, semantic roles, semantic types, and frame relationships, can be thought as implicitly learned in the model, where the only explicit phenomena used in constructing the embeddings are the annotations of the target lexical units. Because these target LUs are available in the annotated data, it is also possible to construct **LU embeddings** - representations of the target predicate in context. The authors in Hermann et al., 2014 computed frame embeddings using matrix factorization over syntactic parse features and continuous word representations. To predict frame labels, they learn two matrices, one with predicates and the other with frames in their context; frame and predicate representations are then learned s.th. the distance between a predicate and its frame is minimized. At prediction time, an instance embedding is assigned its nearest frame embedding. The frame embeddings produced in the (Hermann et al., 2014) model were not themselves an end product, but rather were used to assign frames to predicates. Therefore, the assessment of the frame embeddings was an extrinsic evaluation relevant to the frame prediction task.

An alternative approach to learning frame embeddings is to build continuous representations of a frame using Word2Vec (Botschen, Mousselly-Sergieh, and Gurevych, 2017). In this work, authors used frame-annotated corpora in English and, similar to our approach in Section 4.3.2, authors constructed frame embeddings by replacing the frame-evoking predicate with the frame it evokes and then using this pre-processed corpus as input to the Word2Vec model. These representations were qualitatively assessed by nearest neighbor comparison (similar to our nearest neighbor approach described in Section 4.3.2) and used to predict frame-to-frame relations. Our work is the first to use frame embeddings to compare frames across languages (Section 4.3.2).

4.2.2. Aligning Monolingual Embeddings Across Languages

The frame embeddings described above were learned over monolingual texts, but for our purposes, we would like to have embeddings of frames from different languages in the same, shared space. This would allow us to directly compare the same frame across different languages. To this end, there are a few ways that embeddings from different languages can be aligned into a shared, cross-lingual vector space.

Prior research has demonstrated that methods involving cross-lingual supervision over parallel or comparable corpora produce effective cross-lingual embeddings (Luong, Pham, and Manning, 2015; Hermann and Blunsom, 2014; Upadhyay et al., 2016). We seek to measure distance of frame embeddings in a joint, cross-lingual vector space where the embeddings are based on monolingual (not parallel or comparable), annotated corpora. There are many possible ways this can be accomplished.

Recent work used adversarial training and refinement with matrix factorization to learn embeddings for multiple languages in the same, shared space (Lample et al., 2017). In adversarial training, the goal of the discriminator is to identify the language of origin for an embedding so that the model learns to make embeddings in different languages as similar as possible (Conneau et al., 2017). Another approach to learning multilingual word embeddings is to use a bilingual dictionary to learn a transformation from one embedding space onto the other, resulting in a multilingual vector space (Gaddy et al., 2016; Xing et al., 2015; Tomas et al., 2013; Artetxe, Labaka, and Agirre, 2017). This is a straightforward approach that can be readily adopted thanks to the availability of machine-readable, bilingual dictionaries. The approach we apply, described in more technical detail in Section 4.3.3, is a simple implementation of this type, where we learn a linear mapping to project embeddings from one language onto the same space as embeddings in another.

Finally, the last few years have seen an increase in pre-trained language models such as the multilingual BERT model (mBERT) (Devlin et al., 2019). In mBERT, embeddings from different languages are learned jointly; we describe the mBERT model in more detail in Chapter 6 below. Subsequent work showed how multilingual embeddings from joint training can be improved by either using a cross-lingual objective function (Conneau and Lample, 2019), or by adding auxiliary tasks such as paraphrase detection or cross-lingual masked language model training (Huang et al., 2019). Our experiments in this chapter preceded the release of these pre-trained language models; however, recent studies suggest that BERT embeddings still do not perform as well as other distributional semantic models on certain semantic tasks (Lenci et al., 2021), indicating that our embeddings might not be surpassed in quality compared to the basic mBERT model.

4.3. Experiments

We now proceed to our experiments in constructing and aligning frame and lexical unit embeddings across English and German. In Experiment 1, we construct monolingual embeddings in English and German and conduct different evaluations over the quality of those embeddings. Experiment 2 then proceeds to our work in cross-lingual frame embeddings where we map embeddings from one language onto the vector space of another. We then compare the results of our embeddings with the earlier linguistic analyses from existing literature, described in Section 4.1. An overview of the two experiments is shown in Figure 4.1, where the similarity of frames is compared mono-lingually in Experiment 1 and cross-lingually in Experiment 2. This similarity is our computational measurement of cross-lingual frame comparability.

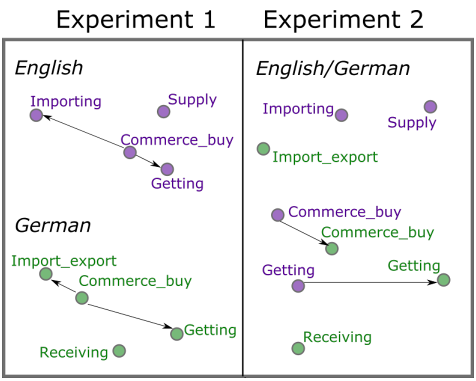


Figure 4.1.: Frame similarity is compared monolingually in Experiment 1, where English and German frames differ slightly in terms of the most similar frames in vector space, and cross-lingual frame similarity in Experiment 2 were frames that are more similar have a higher cross-lingual comparability.

The starting point of our work is the pre-processing of frame annotated data in English and German, with the goal of constructing embeddings for 1) frames, and 2) lexical units, in both languages. To this end, the pre-processing of the corpora should result in four datasets; one corpus for frames – a **Frame Corpus**, and the other for lexical units – a **LU Corpus**, in English and German.

4.3.1. Data

The data we use in these experiments are the full-text Berkeley FrameNet corpus (v1.5) for English, and the SALSA corpus for German. The construction of annotations for both resources are described in further detail in Chapter 2 (Section 2.4).

Pre-processing the Berkeley FrameNet corpus

For English, annotations provide tokenized, part-of-speech tagged, and lemmatized sentences, where the frame-evoking LUs are represented with both their unlemmatized surface forms as well as their standard lemma+pos tag form. We pre-process the corpus in three steps.

First, there are several LUs that are multi-word expressions. Since the Word2Vec model handles each individual token as a standalone unit, these multi-word expressions need to be concatenated together to be learned as a single frame-evoking predicate. This is especially critical for multi-word expressions where the tokens that compose the predicate are themselves predicates for other frames. For example, *ride* evokes the RIDE_VEHICLE frame (*Clara rode the train to Stuttgart*), but the multi-word expression LU *ride out* evokes the SURVIVING frame (*Clara is going to ride out the storm in her cellar*). To keep these predicates as single units in the model, we concatenate them in the corpus with an underscore (i.e., "ride_out").

Second, we detect named entities in the corpus, many of which also span across several tokens. These named entities are later used for learning a linear mapping across languages (see Section 4.3.3). Similar to the multi-word LUs, entities that span multiple tokens need to be concatenated together to have a single representation in vector space (ex, *San Francisco* is then written as "San_Francisco" in the corpus). We detect these entities over the corpus using the spaCy [1] named entity recognition (NER) tool, which is a convolutional neural network (CNN) model trained to predict entities and their labels.

Finally, to build the Frame Corpus, we replace each occurrence of a LU with the name of the frame it evokes. This corpus allows us to build a vectorized representation of a frame that is based on its sentential context. An example of the final English Frame Corpus and LU Corpus is shown in Table 4.1 below.

Corpus	*Example Text*
Original	"The Washington Post reported on the country's biological weapons labs"
LU	[The_Washington_Post, *report*, on, the, country, 's, *biological_weapon*, lab]
Frame	[The_Washington_Post, STATEMENT, on, the, country, 's, WEAPON, lab]

Table 4.1.: Example of original FrameNet sentence and the same input sentence in the Lexical Unit (LU) Corpus and Frame Corpus

FrameNet provides "exemplar" sentence annotations, which are annotations for the lexical units in a frame that are selected as representative examples of the LU and its semantic roles. However, these exemplar annotations have a single frame per sentence, meaning that, in cases where there are multiple frames in a sentence, the annotations only contain representations of one single frame. Instead we chose to construct the frame embeddings over the full-text annotations, as it has annotations for each frame encountered in a sentence. This ensures that every instance of a frame is actually annotated, which is a more useful setting for constructing the frame embeddings.

Pre-processing the SALSA corpus

While the pre-processing of the German annotations in SALSA proceed similar to the English annotations, there are a few additional phenomena, such

[1]https://spacy.io/api/entityrecognizer

as separable prefixes and multi-word predicates, that need to be accounted for when processing data in the German language. Similar to the English annotations, the SALSA corpus provides sentences that have been tokenized and lemmatized.

We detect entities in German with the spaCy pretrained German model, and again we convert entities that span multiple tokens into a single word in both the LU Corpus and Frame Corpus. The multi-word entities are concatenated with the underscore so that the surface form of the entity would be consistent with the same entity in the English corpus (thus allowing us to later align the entities across the English and German corpora).

Multi-word LUs and Separable Prefixes in German In German, there is a highly frequent class of multi-word predicates – that is, German predicates with separable prefixes. These need to be handled in pre-processing, as these predicates have particles that are an integral part of the meaning of the LU. Like English, this is especially the case where verbs that have separable prefixes with stems that can themselves be frame-evoking predicates for different frames. For instance, the predicate *zurückdenken* ("to think back") evokes one of SALSA's prototype frames (ZURÜCKDENKEN1-SALSA), where the prefix *zurück* can be separated from its stem *denken* in text. The LU *denken* ("to think") has its own entry in the SALSA lexicon, as it can evoke several different frames (AWARENESS, CATEGORIZATION, COGITATION, EXPECTATION, to name a few). Unlike English, however, the separation of a prefix and its stem for verbs like *zurückdenken* can be more drastic, where the stem can at times appear in the beginning of the sentence and its separable prefix coming at the end (see the *ablehnen* example in the "Original Corpus" example in Table 4.3).

VVFIN+PTKVZ	→	PTKVZ_VV
Example: gehört an	→	angehören
VM* VV*	→	no concatenation
Example: müssen rechnen		
PTKZU+VV*	→	VV*
Example: zu sagen	→	sagen

Table 4.2.: Rules for combining multi-word predicates in German SALSA corpus where certain particles are concatenated to the main verb to form a single predicate (first example), while other particles remain separated from the main verb s.th. only the main verb is annotated.

Prefixes should not be simply concatenated to every stem that is observed in the data, however, as there are also particles in German that should not be treated as part of certain LUs. These include particles for modals and the infinitive *zu* form (*to*). We use the POS tags provided by the corpus, originally taken from the TIGER Treebank annotations (Brants et al., 2002; Schiller et al., 1999), to discriminate between cases of particles that are separable prefixes (*PTKVZ*) from infinitive markers (*PTKZU*) and modals (*VM*). The rules for reconstructing LUs from their annotated tokens in German data is shown above in Table 4.2.

After pre-processing the SALSA corpus for entities and multi-word LUs, we again replace the frame-evoking LU with the name of the frame it evokes in German. This results in a LU Corpus and a Frame Corpus for the German annotations, shown in Table 4.3.

Corpus	*Text*
Original	"Konzernchefs lehnen den Milliardär als US-Präsident ab" (CEOs reject the billionaire as US President)
LU	[Konzernchef, *ablehnen*, der, Milliardär, als, US-Präsident]
Frame	[Konzernchef, JUDGMENT_COMMUNICATION, der, Milliardär, als, US-Präsident]

Table 4.3.: Example of original SALSA sentence and the same input sentence in the LU Corpus and Frame Corpus

4.3.2. Experiment 1: Building and Evaluating Frame Embeddings

In Experiment 1, we first construct frame and lexical unit embeddings for English and German. We then experiment with different tuning of these parameters and the quality of the monolingual embeddings in the evaluation phase. In this experiment, we assess the quality of monolingual embeddings before we move on to the tests of cross-lingual alignment of these embeddings in Experiment 2.

Building Monolingual Frame Embeddings

We described in Section 4.2.1 the different approaches to learning frame embeddings. For our experiments, we adopt the Word2Vec approach (described in technical detail in Section 3.1) where frame and lexical unit embeddings are constructed from the Frame Corpus and LU Corpus for English and German. In the Word2Vec model, the hyperparameters that can be tuned for model performance, so selection of these hyperparameters can have a large effect on the quality of the subsequent embeddings. We tune the following hyperparameters for the Word2Vec model: the learning rate (α), size of the context window (*win*), number of iterations (*iter*) and number of negative samples (*neg*) selected for learning. All embeddings are 300-dimensional, which we chose by convention. In other hyperparameters in the model, including the option to ignore words with frequency below a certain count (*min_count*), the negative sampling distribution shape (*ns_exponent*), and the threshold for downsampling frequent words (*sample*), we took the default option as we did not choose to affect the weighting of specific words in the vocabulary.

Evaluating Monolingual Frame Embeddings

Before moving on to Experiment 2, we do an internal evaluation of the monolingual frame and LU embeddings. For our purposes, we consider frame representations as clusters of individual predicates, where a frame embedding can be thought of as successfully capturing the semantics of a frame when it fulfills the following criteria: first, in a **LU centroid check**, the frame embedding should be most similar to the centroid of embeddings from all the LUs which evoke the frame. Second, in a **semantic neighborhood check**, the frame or lu embedding should be closest to frames or lus that are semantically similar to it and farther away from those that are semantically dissimilar.

LU centroid check The first evaluation compares the embedding of a frame learned over the Frame Corpus with the centroid of its LU embeddings, learned from the LU Corpus [2]. For example, if the embedding is a good representation, we should expect that the embedding for the frame COMMERCE_BUY is most similar to the centroid of the embeddings for its frame-evoking LUs (*buy, purchase, get*), and is likewise dissimilar from embeddings of LUs that don't evoke the frame. Concretely, if we define a centroid c_{LU} for the frame-evoking

[2]Recall that both the Frame Corpus and LU corpus are identical except for the frame and LU tokens. The embeddings therefore end up in a similar latent space which allows us to compare the two

neg	alpha	win	iter	R@1	R@5	R@10
5	.025	2	10	0.481	0.701	0.766
10	.025	5	20	0.832	0.951	0.970
10	.05	2	20	0.733	0.923	0.957
20	.025	2	30	**0.931**	0.987	0.993
20	.025	2	35	0.929	0.990	0.994
30	.025	2	30	0.912	0.988	0.993

Table 4.4.: LU centroid check for English

LUs in a frame f, we can compute c_{LU} as a simple, unweighted average:

$$c_{LU}(f) = \tfrac{1}{||\{\mathrm{lu} \mid \mathrm{f} \in F\}||} \sum_{\mathrm{lu} \in f} \overrightarrow{\mathrm{lu}}.$$

For each frame f, we can then take all the c_{LU} embeddings and produce a ranked list of similarity to those LU centroid embeddings using cosine similarity (defined in Section 3.1). The quality of the frame embedding in the LU centroid check is determined by the percentage of frame embeddings where the frame embedding is most similar/highest ranked to its LU centroid embedding. Using the LU centroid check, we compare the embeddings produced by one set of hyperparameters in the Word2Vec model against the embeddings tuned on a different set of hyperparameters.

We report model performance on ranked retrieval, where recall at 1 (R@1) is the percentage of cases where the frame embedding is most similar to its LU centroid, recall at 5 (R@5) are percentage of frames whose LU centroids are among the 5 most similar, and recall at 10 (R@10) for the top ten of the most similar frames. A sample of the hyperparameter tuning results for the English frame embeddings are shown in Table 4.4 and German frame embeddings are

neg	alpha	win	iter	R@1	R@5	R@10
5	.025	2	10	0.408	0.591	0.653
10	.025	2	20	0.843	0.925	0.938
10	.05	2	20	0.938	0.986	0.993
20	.025	5	20	0.904	0.972	0.979
20	.025	2	35	**0.965**	0.979	0.986
30	.025	2	30	0.931	0.993	0.993

Table 4.5.: LU centroid check for German

in Table 4.5. These results demonstrate the importance of hyperparameters in the quality of frame embeddings, where too few negative samples and low iterations results in noisy embeddings for both English and German. In general, conditions where the negative sampling was > 5, the R@10 values were quite high, suggesting that already the model is learning semantics of a frame. The optimal set of parameters for learning frame embeddings is similar in English to German, where the best results are achieved when the model has a number of negative samples > 10, α at .025, context window at 2, and number of iterations > 20.

The hyperparameters which yielded the best results from the LU centroid check differ slightly between English and German, and we take the embeddings from the model that produced the highest results for English and the highest results for German for Experiment 2.

Semantic Neighborhood check For the semantic neighborhood check, we perform separate analyses over the frame and LU embeddings, but the method of evaluation is the same. For the frame check, we find the top-N most similar frames to the target frame based on their cosine similarity scores (see Section 3.1) and we evaluate qualitatively whether they are semantically related to the target. The LU check similarly takes the top-N most similar LUs to the target LU. We chose to evaluate over the top 10 nearest neighbors, as 10 is typically sufficient to assess whether the model has learned the semantic similarities that one would expect.

First we qualitatively analyze the top ten nearest neighbors for frames, where we expect that a good frame embedding will be nearest to semantically related frames. In this evaluation, we would expect that frames form sensible semantic neighborhoods, where a frame like COMMERCE_BUY has COMMERCE_SELL and GETTING frames in its nearest neighbor list, while frames that are more distant on the FrameNet hierarchy, such as TELLING, will not be in the same semantic neighborhood.

Table 4.6 shows results for the COMMERCE_BUY frame, which verifies that the frame embeddings do indeed form semantic neighborhoods with frames that one would imagine are highly related (for example, GETTING, COMMERCE_SELL, IMPORTING). Comparison of the nearest neighbors with different hyperparameters in the training phase shows that the hyperparameters also affect the qualitative output of the embeddings, where the optimal parameters have nearest neighbors more conceptually related to the concept of COMMERCE_BUY. For instance, the English model with lower iterations and fewer negative samples includes dissimilar frames such as LEGALITY, RATIFICATION, and COLLABORATION, while the German model with the same hyperparameters includes frames such as SUITABILITY, RANKED_EXPECTATION, and also COLLABORATION. The models with higher iterations and more negative samples are all much more closely related to the concept of COMMERCE_BUY. Table 4.6 shows similar results in German, where the nearest neighbors for the COMMERCE_BUY frame are similar to those we see in the English frame

win=2, *neg*=5, *iter*=10	
English	German
Source_of_getting, Process_completed_state, Collaboration,Withdraw_from_participation, Undergoing,Suitability, Cause_change_of_position_on_scale, Legality, Collaboration_Make_agreement_on_action, Ratification	Source_of_getting, Suitability, Process_completed_state, Collaboration, Cause_change_of_position_on_scale, Collaboration_Make_agreement_on_action, Withdraw_from_participation, Legality, Exchange, Ranked_expectation
win=2, *neg*=10, *iter*=15	
English	German
Amassing, Activity_pause, Delivery, Willingness, Transfer, Receiving, Legality, Process_completed_state, Source_of_getting, Cause_to_Resume	Amassing, Delivery, Activity_pause, Process_completed_state, Cause_change_of_position_on_a_scale, Cause_expansion, Attributed_information, Import_export, Receiving, Seeking_to_achieve
win=2, *neg*=20, *iter*=35	
English	German
Commerce_sell, Getting, Supply, Importing, Degree_of_processing, Transfer, Import_export, Receiving, Amassing, Cause_motion	Getting, Import_export, Receiving, Cause_motion, Removing, Manufacturing, Activity_start, Using, Commitment, Bringing

Table 4.6.: Semantic neighborhood check for top 10 most similar frames to Commerce_buy

embedding space. In Figure 4.2, we visualize these same frames within the FrameNet hierarchy, where many of the top neighboring frames are close to the target frame – meaning they are separated by only one or two frame relations. The frame-to-frame relations are not included in the FrameNet annotations, and the visualization in Figure 4.2 was manually drawn; therefore, the semantically related frames we see in the neighborhood check come only from the usage of the frame in text.

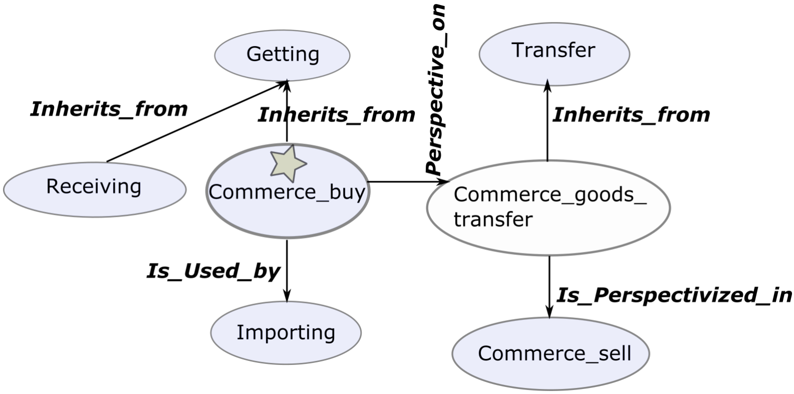


Figure 4.2.: Frames in the top 10 nearest neighbor list from Table 4.6 for the English COMMERCE_BUY frame (starred) represented in the FrameNet hierarchy. Most are one or two relations away from the target frame, demonstrating the embeddings are capturing semantic neighborhoods of frames

We additionally perform an analysis of the top 10 nearest LUs to a target LU to determine if they also form reasonable semantic neighborhoods to the target LU. In Section 4.1, we discussed the German *ankündigen* predicate and *announce*, which evokes the STATEMENT frame in English. Because we will proceed with our quantitative analysis of this case in Experiment 2, we will first do a semantic neighborhood check to determine if the predicates in the embedding space are nearest to semantically related predicates.

Results of the semantic neighborhood check for the predicates *announce* and *ankündigen* in Table 4.7 shows that there is an effect of the hyperparameter settings and the quality of the LU embeddings. The case is more clear with the English predicate, where many unrelated words appear in the space with smaller iterations and negative samples (*English, until, sometimes*), but the last group of nearest neighbors, with higher iterations and negative samples, are all clearly related to the *announce* predicate. Interestingly, the nearest neighbors in the English models with the lower number of iterations and negative samples include nominals, adverbs, and prepositions, while the model with a larger number of negative samples and iterations has all verbs as the top nearest neighbors. This suggests that with the ideal hyperparameters, the model has learned more of the syntactic properties of the predicates. The German *ankündigen* predicate shows that there is an effect in hyperparameter settings, but the effect is not as marked as the English. One potential reason is the size of the German data, which is larger than the English annotations (see Table 2.1). Surprisingly, the nearest neighbor results for *ankündigen* show many verbs relating to cognition (*know, believe, think, guess*) instead of verbs relating to communication, revealing some underlying differences in the SALSA and FrameNet corpora.

4.3.3. Experiment 2: Quantitative and Qualitative Assessments of Frame Comparability

The goal of this chapter is to describe a computational metric for assessing a frame's cross-lingual comparability. Experiment 2 combines the monolingual embeddings from Experiment 1 into a cross-lingual vector space where frames can be assessed in terms of their usage-based, cross-lingual comparability.

win=2, *neg*=5, *iter*=10	
English	German
report, say, sometimes, visit, estimate, clearly, possible, English, consider, until	*glauben, wissen, erklären, wünschen, denken, nein, fragen, vorstellen, verstehen, erinnern* (*believe, know, explain, wish, think, no, ask, imagine, understand, remember*)
win=2, *neg*=10, *iter*=15	
English	German
say, report, however, clear, declare, consider, fact, far, indicate, possible	*kennen, erleben, passieren, erklären, fragen, brechen, wissen, glauben, denken, vermuten* (*know, experience, happen, explain, ask, break, know, believe, think, guess*)
win=2, *neg*=20, *iter*=35	
English	German
reveal, say, tell, argue, promise, demand, claim, declare, confirm, respond	*entstehen, entnehmen, gelegen, überraschen, behandeln, erklären, verstecken, behaupten, thematisieren, kennen* (*arise, extract, located, surprise, treat, explain, hide, assert, address, know*)

Table 4.7.: Top 10 most similar predicates to the *announce* and *ankündigen* predicates with different hyperparameter settings

Aligning and Comparing Frame Embeddings across Languages

Thus far, the embeddings we have described are monolingual embeddings for English and German. We adopt the approach described in Section 4.2.2 where we learn a linear map to project the monolingual embeddings onto a shared, cross-lingual vector space. We adopt a straightforward method that maps embeddings from one vector space onto another via a linear transformation (Mikolov, Le, and Sutskever, 2013).

One simple strategy for building this bilingual seed dictionary automatically is to use named entities (Zhong et al., 2015; Chang et al., 2016) since they tend to have reliable correspondence across languages as they are considered rigid designators (Kripke, 1972) – meaning they are not prone to shifts in word sense as they reference the same object. For languages such as English and German, they are easier to align due to transliteration, where the characters appear similar across languages (for example, US president/US-präsident in English/German). As described in Section 4.3.1 use the spaCy tool to detect entities in text, which also provides a large set of labels for different entity types. Because the entity types include cardinal numbers, dates, and percentages, we select only the entities which are part of standard classes (organizations, persons and person groups, locations, and geopolitical entities). On first match, we select entities whose name is identical in English and German (i.e., *San Francisco – San Francisco*). However, there are clearly cases of entity matches where there is a slight variation in the naming, such as *President Clinton* in English translated as *Präsident Clinton* in German. To increase the size of the bilingual seed dictionary, we use a bilingual dictionary for English-German to match any named entities where the surface form varies but the individual or location are the same (for example, *Rome – Rom*). In total, over 40.2k entities were detected in the English FrameNet sentences, and over 15.3k entities were detected in the German SALSA corpus. The intersection of the entities that appeared in both corpora was ∼3k.

The linear mapping takes a seed dictionary D of word pairs $\langle x^s, x^t \rangle$, where x_i^s is an embedding for a word in the source language, and x_i^t is the embedding for its target language translation. We use the bilingual dictionary of named

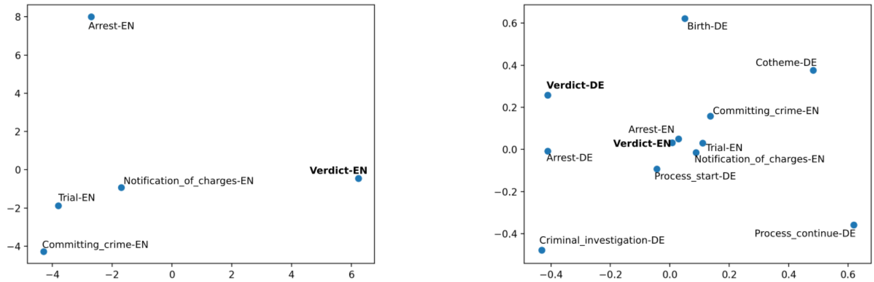


Figure 4.3.: English space around VERDICT (left), and joint, cross-lingual space (right). Many semantic relationships in English (-EN) are preserved in the mapping with German (-DE) frames.

entities to find a transformation matrix $\mathbf{W}$ where $\mathbf{W}x_i^s \approx x_i^t$. The matrix $\mathbf{W}$ is learned by minimizing the mean squared error (MSE) between the source and target embeddings for all pairs in the dictionary:

$$\mathrm{MSE} = \sum_{i=1}^{D} \parallel \mathbf{W}x_i^s - x_i^t \parallel^2$$

Once the transformation matrix is learned, we can apply it to embeddings from one source language vector space to a target language vector space by computing $z = \mathbf{W}x$. The result of this process is a shared, cross-lingual vector space where embeddings from both languages – despite being learned from different corpora – can be compared directly.

The result of this mapping is shown in Figure 4.3, where 300-dimensional frame embeddings have been reduced into 2-dimensional space with tSNE (see Section 3.1). Visualization of these frame embeddings verifies that the semantic neighborhoods of a frame are relatively preserved when mapped to a joint, cross-lingual vector space. Specifically, the English VERDICT frame, whose nearest neighbors in the 2-dimensional space are the English frames NOTIFICATION_OF_CHARGES, TRIAL, ARREST, and COMMITTING_CRIME, are still near one another in the joint space. We see in the joint space that the German frame vectors CRIMINAL_INVESTIGATION and PROCESS_START are also near the English VERDICT.

In Experiment 1, we used cosine similarity to evaluate monolingual embeddings that were learned in the same vector space. The alignment of monolingual frame embeddings onto a shared vector space means that we can now measure a frame's cross-lingual similarity with cosine similarity as well. A frame that has a high cosine similarity score with its cross-lingual counterpart indicates that its distributional features are similar across languages, and therefore can be considered as 'highly comparable', while a low cosine similarity between a frame's embedding across languages indicates a low comparability between frames.

Frames with Highest and Lowest Comparability

To get a sense of the cross-lingual frame embeddings, we start by analyzing the frames with the highest and lowest comparability in the shared, cross-lingual vector space in Table 4.8.

Frames with the highest measures of cross-lingual similarity are shown below, where GRASP is at the top of the list. In the GRASP frame, there is a COGNIZER that is understanding or comprehending some PHENOMENON, and it is one of the several frames that belong to an abstract domain of cognitive events (GRASP, JUDGMENT, MEMORY) that have high cross-lingual similarity in the shared vector space. Frames that relate to abstract event properties, such as ACTIVITY_RESUME and LIKELIHOOD are also frames that are high on the list. Many of the frames on the list of high comparability have a high number of annotations in English and a decent number in German, and are also frames that are more general (as discussed above).

Frame	Comparability	Freq (EN/DE)
Grasp	0.66	287/21
Communication	0.54	84/25
Verdict	0.51	215/77
Arrest	0.51	189/103
Building	0.49	393/52
Activity_resume	0.49	37/8
Judgment	0.48	1212/33
Memory	0.48	209/41
Cotheme	0.48	665/7
Likelihood	0.48	577/6

Table 4.8.: Frames with the highest comparability, where *comparability* is the cosine similarity metric between the frame's English and German vectors. *Freq* gives the number of annotations for the English (EN) and German (DE) frame.

Taking the frames with the lowest comparability scores, many of these low comparability frames appear to result from large differences in the LUs that either compose the frame or that are chosen for annotation in the corpus – indicating that some of these frames have a different meaning across the languages. The Origin frame in the English FrameNet contains nominals for many nationalities and ethnicities, such as *American.n, Assyrian.n, Byzantine.n*, while the German SALSA frame defined LUs as kinship terms such as *Kind.n, Sohn.n, Enkel.n* (*child, son, grandson*). The low cosine similarity score is therefore illustrating the difference in how the two frames are defined across the languages, where the interpretation of the frame's meaning is not overlapping.

Two frames on the list have a dramatic difference in the diversity and size of LUs in their frame. In English, the PEOPLE_BY_VOCATION frame has an extensive list of LUs relating to different professions, such as *farmer.n, actor.n, waitress.n, painter.n, judge.n, journalist.n*, etc., while the German SALSA frame only lists one LU, *Kumpel.n* (*miner*). Here, the low similarity is likely due to the sparsity on the German side, where a greater number of LUs and more diverse annotations could improve the computational measurement of comparability. Table 4.9 shows that several of the frames with the lowest comparability have low annotations counts in the German SALSA corpus, which could potentially explain the lower scores.

The COMMITTING_CRIME frame tells a similar story to PEOPLE_BY_VOCATION, although in this case there is arguably a difference in the interpretation of the frame's meaning as well; in English, LUs include *commit.v, crime.n, perpetrate.v*, while the SALSA frame again only lists one LU, *Täter.n* (*offender/culprit*). Although these LUs are conceptually in the same semantic space, their distribution in text is quite different. To demonstrate, an example of the COMMITTING_CRIME frame in English (Example 27) has the semantic roles PERPETRATOR and CRIME, while the German example in 24 has the predicating LU as the PERPETRATOR:

(24) jetzt sucht man der flüchtig [*Täter.* $_{\text{COMMITTING\_CRIME/PERPETRATOR}}$]
now searching we the fleeing culprit
"now, we are searching for the fleeing culprit."

(25) ...[politically motivated violence $_{\text{CRIME}}$] [*perpetrated* $_{\text{COMMITTING\_CRIME}}$] against

noncombatant targets by [sub - national groups or clandestine agents PERPETRATOR]

COMMITTING_CRIME in English does not have any predicating LUs that could also express the PERPETRATOR of the crime; however, the only nominal predicate (*crime.n*) does express the CRIME semantic role. These differences clearly emerge as different distributional patterns in the corpus, which affect the assessment of frame comparability. In this case, it seems accurate to suggest that these are two frames that, while named identical, are interpreted and annotated quite differently across English and German.

Frame	Comparability	Freq (EN/DE)
ORIGIN	-0.14	194/8
PEOPLE_BY_VOCATION	-0.11	586/12
UNDERGOING	-0.07	38/136
INGEST_SUBSTANCE	-0.02	187/7
EMPLOYING	0.00	151/428
TEXT	0.03	1080/4
SENSATION	0.03	471/1
TAKING_SIDES	0.04	189/18
COMMITTING_CRIME	0.07	48/55
SENTENCING	0.07	77/39

Table 4.9.: Frames with the lowest comparability, where *comparability* is the cosine similarity metric between the frame's English and German vectors. *Freq* gives the number of annotations for the English (EN) and German (DE) frame.

Comparison of Computational Measure of Frame Comparability with Qualitative Studies

We assume that the qualitative measures of frame similarity that come from linguistics, such as the presence of certain semantic roles, will be latent features of our embeddings, thus influencing the representations of the frames. In order to determine whether this assumption holds, we compare our embeddings with the qualitative analyses that have been conducted over those same frames across the English-German language pair. In this section, we will revisit the hypotheses discussed earlier in Section 4.1, and we end with an additional examination into the frames with the highest and lowest comparability overall for a broader analysis of our embeddings.

Motion and Communication frames Earlier, we presented the hypothesis that the MOTION and COMMUNICATION frames would have a higher cross-lingual comparability than more specific frames, as they are frames that are abstract and have a higher "cultural equivalence" (Section 4.1).

To test these hypotheses, we compare the COMMUNICATION and MOTION frames in our cross-lingual frame embedding space. The COMMUNICATION frame is inherited by more specific child frames COMMUNICATION_RESPONSE and COMMUNICATION_MANNER, while MOTION is inherited by three child frames: MOTION_DIRECTIONAL, SELF_MOTION, and CAUSE_MOTION. Figure 4.4 visualizes the relationships between the English and German versions of these frames; in general, the more abstract frames are closer in the low-dimensional space than their child frames, although the tendency is actually more pronounced for the COMMUNICATION frame. The visualization is consistent with the similarity scores, where the COMMUNICATION and MOTION frames are closer together in the cross-lingual vector space than their more

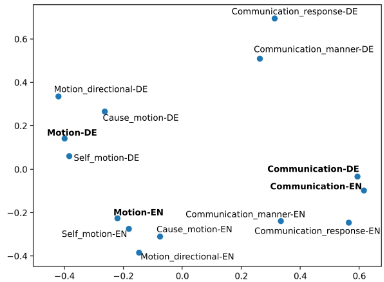


Figure 4.4.: Cross-lingual similarity scores for COMMUNICATION and MOTION.

Frame	Similarity DE-EN
COMMUNICATION	0.54
COMMUNICATION_RESPONSE	0.17
COMMUNICATION_MANNER	0.14
MOTION	0.39
CAUSE_MOTION	0.15
MOTION_DIRECTIONAL	0.27
SELF_MOTION	0.46

Figure 4.5.: Cross-lingual similarity between the less-specific COMMUNICATION and MOTION frames and more specific child frames.

specific sub-frames. Additionally, Table 4.8 confirms that the COMMUNICATION frame has one of the highest overall comparability scores for German and English.

The question we then ask is, could we extrapolate these results with all frames that are thought to be more abstract? To answer this question, we define abstract/general frames more precisely by counting FrameNet's frame-to-frame relations that specify parent/child relationships between frames. Recall from Chapter 2 that FrameNet frames are organized as a hierarchy of concepts where abstract, general frames are towards the top of the hierarchy and concrete frames are towards the bottom. The lower the frame sits on the hierarchy, the more specific its lexical units are, and the more specific its set of core roles and valence patterns. For this reason, frames that tend to exist towards the bottom of the hierarchy are thought to be those that are more language-specific, as they are more closely tied to linguistic constraints. Specifically, we take the relations "X is Inherited by Y", "X is Used by Y", and "X has Subframe Y" where X is a more abstract parent frame to a specific Y frame. For any given frame, if the frame has a high number of these relations, then it should be a more abstract frame and therefore more similar across the English-German language pair. We take the top 10 frames with the highest comparability in Table 4.8 and find that they have a much higher average number of relations in position X (4.89), while the 10 least comparable frames in Table 4.9 have a relatively low position (1.7). This demonstrates that, overall, the highest and lowest frames tended to be more and less abstract, respectively.

Sub-frames Relating to Motion in the Embedding Space In Section 4.1, we hypothesized that the Motion sub-frames, such as Cause_Motion and Self_Motion would have a lower comparability across German and English due to differences in lexicalization patterns. Table 4.5 indeed shows that the Cause_Motion frame has quite a low similarity score in the shared space, and a lower cross-lingual similarity than its parent Motion frame. However, the Self_Motion frame actually contradicts our hypothesis: the cross-lingual similarity score is quite high in this case; in fact, it shows a higher cross-lingual similarity than the Motion frame. One reason for this might be the predicates that appear for the Self_Motion frame in SALSA (*marchieren(march)*, *tanzen(dance)* , *hasten(hasten)*) do in fact express the manner of motion, similar to English. In these cases, the lexicalization differences demonstrated in the translated examples in the linguistic analyses don't appear to correspond with the quantitative cross-lingual similarity score. One possibility is that the cases of lexicalization differences are not predominate enough to affect the usage-based similarity score that comes from the annotated data; while specific translations may demonstrate divergences in manner of motion verbs, on the whole, many German predicates can express manner of motion.

Statement frame We described in Section 4.1 the STATEMENT frame in German, where we hypothesized that the *ankündigen* LU embedding would fit well with the STATEMENT frame as it is defined in English, and second, that the other German predicates would have a lower compatibility with the English STATEMENT frame. Recall that we have two embedding models for German and English: a model built from the LU Corpus that has an embedding for each LU, and a model built from the Frame Corpus that has embeddings for each frame. Thus far, we have been comparing embeddings of frames across languages, but we could compare LUs across languages. To test the hypothesis about the *ankündigen* predicate and STATEMENT frame, we compare the embedding of the LU *ankündigen* with the frame embedding of the English STATEMENT. This tells us how well the predicate from German fits into the frame's 'meaning' in English, where a high similarity score would indicate that the predicate is a good match for the frame.

Interestingly, when we look at the data in the SALSA annotations, we find that despite the fact that STATEMENT exists as a frame in SALSA, the *ankündigen* predicate is actually annotated for the frame HERALDING for an overwhelming majority of its instances (85/87). HERALDING has similar roles to the STATEMENT frame – a COMMUNICATOR, who informs the public about an EVENT. The EVENT is a verbal commitment the speaker makes to carry out a specific action:

(26) Regierung in Rom kündigt Preisstopp und Sparprogramm
Government in Rome announces price stop and savings program.
an.

"The Government in Rome announces a price freeze and savings program."

When we compare the cross-lingual embeddings of the German LU *ankündigen.v* with the set of available English frames, we find that the most similar English frame embedding is actually the STATEMENT frame. This indicates that the actual usage of *ankündigen* in SALSA is semantically very similar to the English STATEMENT, and it could potentially draw an implicit link between the HERALDING and STATEMENT frames which do not currently share any frame-to-frame relation. This coincides with the findings from the qualitative studies where *ankündigen* is one of the only German predicates which fits the constructions of the English STATEMENT frame.

4.4. Follow-up Studies

Following our study, Baker and Lorenzi, 2020 built frame embeddings using the FastText tool (Bojanowski et al., 2017). Instead of constructing frame embeddings directly from the available annotated corpora, authors compute frame comparability scores based on embeddings trained over Wikipedia data and aligned with an adversarial approach (described in Section 4.2.2). Authors took cross-lingual embeddings of lemmas and calculated a semantic neighborhood n of size k around a target language lemma embedding $\vec{t}$ using the cosine similarity metric (similar to the semantic neighborhood check described in Section 4.3.2). Given the set of lemmas for a frame in two different languages, were L_s is the set of lemmas for a source language and L_t is the set of lemmas for a target language, authors score the frame comparability by calculating the number of lemmas from the first language's lemma list that are in the semantic neighborhood of each $\vec{t}$:

$$sim(L_s, L_t) = \frac{\mid \{a \in L_s, b \in L_t : \vec{b} \in n(a)\} \mid}{\mid L_s \mid} \tag{4.1}$$

The authors remarked that one shortcoming of this approach is that the senses of the lemmas are not accounted for in the embeddings they use, meaning there is no guarantee that the lemma embeddings they take are true lemmas of the frame of interest. Conflation of different word senses is a recurring problem to many approaches to word embeddings in general. The problem would similarly affect our LU embeddings: a surface form in the annotated data that is identical to the surface form of a different word sense would be conflated into a single LU representation within the Word2Vec model. Importantly, because our method is based on annotated data, the frame embeddings do capture word sense distinctions; this is one critical distinction between our method and Baker and Lorenzi, 2020 approach. We further address the question of word sense and frame embeddings in the following chapters when we turn to contextualized vector representations of frames.

4.5. Summary

Part II of this thesis has addressed the first question described in the Introduction: *how can we measure frame comparability*? We focused on computational representations of frames and methods to automatically compare these representations across languages. The metric we chose overwhelmingly speaks to the usage-based comparability of frames across languages, where the computational model is capturing (dis)similarities in frames based on an aggregated representation of their usage in context.

The computational representation we chose for this work, a frame embedding, was constructed automatically using the Word2Vec model. As described in the Technical Background (Section 3.1), the quality of embeddings is dictated by the tuning of hyperparameters in the model. To ensure our results and analyses are not merely artifacts of the way the embeddings were built, we established two tests, the LU centroid check and the semantic neighborhood check, to validate the quality of the frame and LU embeddings. These metrics assess how well the embedding space is aligned to the linguistic definition of the frames in the frame lexicon, where a 'good' frame embedding should have sensible nearest neighbors that are close relatives in the FrameNet hierarchy, and the embeddings of lexical units should be most similar to the embedding of the frame they evoke. We use the LU centroid check for tuning hyperparameters of the model, where the goal is to have the frame embedding as close as possible to the centroid of its LU embeddings. The semantic neighborhood check confirmed that the top 10 nearest neighboring frame and LU embeddings were, indeed, semantically similar to the target frame or LU embedding. The result of the efforts described in this chapter are vectorized representations of frames in English and German, both of which are in their own vector space as they were trained on monolingual corpora.

Importantly, we demonstrated how embeddings based on frame annotations can be aligned across languages and compared using a geometric distance metric. Comparison of the same frame across different languages revealed that specific observations from the linguistics literature (Boas, 2001; Boas, 2020b; Boas, 2013) were consistent with the findings in the embedding space. We found that in the case of two abstract frames – MOTION and COMMUNICAtion, which were thought to have a more "universal" status (Boas, 2020b), do indeed have a higher similarity in the embedding space than more specific frames. We then generalized further to the top 10 most and least comparable frames in the cross-lingual embedding space, and found that the top 10 most comparable frames had a higher overall score for abstractness (which we based on counts of relations in the FrameNet hierarchy). We took a more specific analysis from the prior linguistics literature which showed that only one German predicate, *ankündigen*, could appear in the same alternations as the English predicate *announce* in the Statement frame. Our embeddings confirmed that, although the German annotations chose the Heralding frame for the *ankündigen* predicate, the English frame most similar to the predicate was Statement. These studies showed, from specific observations to broad generalizations, that our quantitative assessments of frame comparability correspond to many of the insights from the linguistics literature.

An additional finding in our study was that the frame and LU embeddings reflected (dis-)similarities in the annotation decisions of the corpora on which they were trained. It was not always the case that frames that are named similarly and have identical lexicographic definitions to the English frames (for example, Cause_Motion or Communication_Response) have been interpreted, and more specifically, annotated, the same across the English and German research teams. Again, this seems to be the case even for frames that were taken directly from the English inventory and used in the annotation of German predicates. Because these annotation choices were reflected in the embeddings, we can use our quantitative measure of frame similarity to get an automatic, and "high-level" assessment of how a frame's meaning is interpreted, and compares, across languages.

Finally, the work described in this chapter was conducted before the appearance of contextualized embeddings (Devlin et al., 2019; Peters et al., 2018; Alec et al., 2018); specifically, before the appearance of large, multilingual embeddings in which word representations are constructed simultaneously for multiple languages, such as mBERT. Similar analyses to those we have conducted in this chapter (Section 4.3.3) could presumably be made over these mBERT embeddings. However, in Part III, we will describe multilingual, contextualized embeddings for frame identification and our experiments in cross-lingual frame comparability with mBERT.

Part III.

Comparability and Cross-lingual Models of Frames

5. Model Selection for Frame Identification

Part III of this thesis addresses the question of how much frame comparability is an issue for cross-lingual models of frame identification. Before diving directly into the multilingual aspect, we experiment in this chapter with monolingual models of frame identification, with the ultimate goal of selecting a model design that will be used for later cross-lingual experimentation.

Chapter 3 introduced the task of frame identification and gave an overview of the existing models of frame identification in NLP. In this chapter, we take a well-known neural network-based architecture for learning contextualized embeddings, BERT (Section 3.1), and we design two different systems for training BERT to predict frames to predicates in context. The first design involves pre-trained representations from BERT, while the second fine-tunes the pre-trained embeddings (see Section 3.1 for details about BERT's pre-training and fine-tuning). We chose English for our experiments, as a predominant number of prior frame identification systems were trained and evaluated over English data, making it more straightforward to compare to the existing state-of-the-art models. The result of our efforts is a new state-of-the-art model in frame identification that has been trained to predict frames mono- or multilingually. A significant part of the work described in this chapter is from the paper "Frame Identification as Categorization: Examplars vs Prototypes in Embeddingland", which was published at the 2019 International Workshop on Computational Semantics (IWCS) (Sikos and Padó, 2019).

This chapter is structured as follows. We begin with a theoretical motivation for our experimentation where we describe the competing theories of categorization which form the basis of different model design choices. Next, we describe prior work in testing categorization theories with machine learning methods, followed by a description of how we can design the BERT model to categorize items according to one theoretical motivation or another. We then proceed to our experimental setup in Section 5.2 and results of our different model architectures in Section 5.3. Section 5.4.

5.1. Frame Identification as Categorization

Categorization is a well-studied and debated subject in cognitive science, and experiments on human subjects have been widely conducted in psychology where synthetic datasets are used to test how humans categorize (Malt, 1989; Goldstone, 1995; Goldstone, Lippa, and Shiffrin, 2001; Hampton, 1995). The theory of frame semantics developed during a time when theories on categorization began to play a more prominent role in studies of human cognition (Lakoff, 1999). In fact, frame semantics itself drew heavily from existing theories of categories (Fillmore, 1982) where frames are categories of events; each instance of a frame-evoking predicate is an item belonging to that category. In cognitive psychology, there are two prominent theories that explain how humans categorize over objects: **prototype** and **exemplar** theories. Both theories explain how humans generalize over categories, where a newly encountered object can be classified into a latent category.

Prototype Theory In prototype theory (Rosch, 1973a; Posner and Keele, 1968), a category is represented by a set of objects which, in the collective, represent the typical member of the category – the prototype. Newly encountered objects are classified into a category by their similarity to that category's prototype. In computational models, a representation of category would be computed as the centroid of the objects that instantiate it.

Exemplar Theory In an exemplar account of categorization (Nosofsky, 1986; Hintzman, 1986), a category is represented by all the individual objects which compose the category – the exemplars. New objects are classified as members of the category whose exemplar is nearest to the object. Computationally, an exemplar model of classification is a nearest-neighbor classification.

Most importantly, both theories are proposing different views on how humans abstract, leading us directly to our main question about model design for frame identification: is a representation of a frame category necessary/desirable when it comes to the automatic classification of its predicates? Does the construction of a frame 'prototype' help in frame assignment at all?

5.1.1. Machine Learning Models of Prototypes vs Exemplars

Analogies have been drawn between concepts in machine learning and categorization theories, as classification is a core objective of many machine learning efforts (Biehl, Hammer, and Villmann, 2016; Zubek and Kuncheva, 2018; Jäkel, Schölkopf, and Wichmann, 2008; Kibler and Aha, 1987; Voorspoels, Vanpaemel, and Storms, 2009). In terms of prototype-based machine learning models, vector space models often represent word meaning as a composite of the same word in different contexts, where a summary representation is generated and modified as more instances of the word are seen by the classifier. Additive and multiplicative functions are typically used to combine individual representations (Mitchell and Lapata, 2008), and the summary representation can then be used for classification of a new word by comparing its instance vector to its prototype summary representation using a distance metric (Schütze, 1998; Lowe, 2001). Prototype-based vector space models have been adopted for tasks beyond word sense disambiguation, such as paraphrase detection (Erk and Padó, 2009; Thater, Dinu, and Pinkal, 2009), image classification (Quattoni, Collins, and Darrell, 2008), and face recognition (Klare and Jain, 2013; Yi, Lei, and Li, 2015). An extension of this prototype-based word classification can be made with a multi-prototype model of classification, where multiple vectors can represent prototypes of different senses (Reisinger and Mooney, 2010; Huang et al., 2012).

Exemplar-based approaches include kernel methods where data points are compared via a similarity function (Jäkel, Schölkopf, and Wichmann, 2008); a popular exemplar algorithm is a k-nearest neighbors classification, where an instance vector is compared to its nearest neighbors in space (Li and Zhang, 2011). This is one of the variants taken in (Erk and Pado, 2010), in which authors predict word sense by taking all exemplars of a target word — that is, the prior seen instances of the word, and activating the relevant exemplars. Relevant exemplars were established by either the k-nearest neighbors, where the k-most similar neighbors are chosen based on an existing similarity metric, or a percentage-based metric where a certain top x percentage of most similar exemplars were chosen. The authors of this exemplar model used simple bag-of-words co-occurrence vectors for the target and exemplars, and find that the exemplar models outperformed the existing prototype models.

Computational models of each theory differ in terms of the decision boundary that the classification is based upon. In prototype theory, the decision boundary lies between the prototypes, making a linear boundary; in contrast, the boundary between exemplars is complex and non-linear, making the cost for training requirements higher. This is the traditional trade-off between exemplar and prototype models: exemplar models are more expressive and should work better than prototype models given the availability of more data, while they perform worse than prototype models given less data.

5.2. Models of Frame Identification with BERT

Recall from Chapter 3 that BERT consists of two primary stages: the pre-training stage, and the fine-tuning stage. With the BERT system, it is feasible to directly test how well the pre-trained embeddings have already implicitly learned frame categories without having seen any frame-labeled data. Results of this test would tell us how much generalized semantic knowledge is sufficient for classifying frames, and whether the semantic categories that have been learned over massive, unstructured corpora overlap with the linguist's intuition of frame categories. The division of learning categories with or without directed, specialized knowledge is analogous to the 'bottom-up' and 'top-down' input that shapes learning in human cognition (Smith and Sloman, 1994). The 'bottom-up' information is knowledge that is implicitly learned from the input – the pre-trained repesentations, while the 'top-down' information is knowledge that is explicitly given from an informed source – the fine-tuned representations.

We contrast four frame identification models over two dimensions in a 2×2 design: prototypes vs exemplars, and pre-trained representations vs fine-tuned representations. In the first dimension, given as columns in Figure 5.1, we compare prototype and exemplar model designs. The prototype model uses a frame prototype for classification, where the frame prototype is an aggregate of the instances of the frame. The softmax classification is itself a prototype approach, as the softmax function is the similarity of a data point to each of the classes; weights are learned for each class, and those class weights can essentially be thought of as 'prototypes'. Models that are designed with exemplars take all the instances of a frame and use those instances directly in classification. In both of these designs, the instances of a frame are LUs in context, where we take all the annotations of a frame from the Berkeley FrameNet full-text corpus. In the second dimension, shown as rows, we examine the degree to which fine-tuning is necessary to learn frame representations for classification. This dimension compares pre-trained representations with fine-tuned representations - analogous to the 'bottom-up' and 'top-down' distinctions in cognition. The pre-train only models use generic representations that were learned in an unsupervised fashion, whereas the fine-tuned representations use the frame annotations to tune the generic embeddings for frame assignment. Both the pre-train and fine-tune models are supervised methods as they rely on the frame annotations for learning, but the critical distinction is whether the embeddings themselves need to be tuned to learn frame categories.

5.2.1. Model Designs

Exemplar Models

The exemplar models compare a test LU with exemplars, where exemplars are the instances of other LUs that have been labeled for their frames. A test LU is assigned the same frame label as its most similar exemplar.

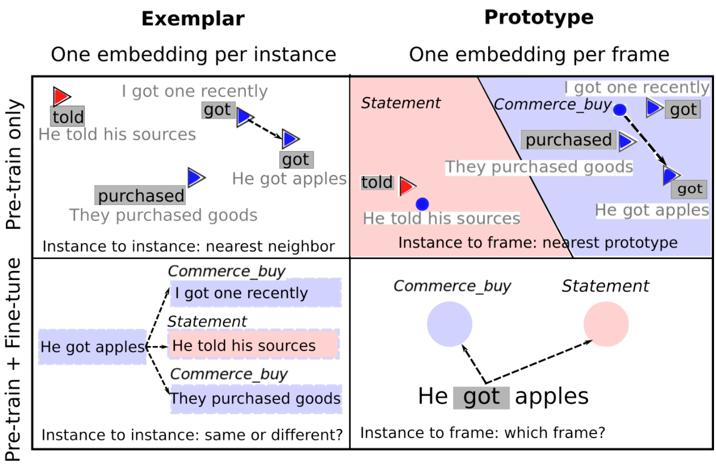


Figure 5.1.: Two dimensions of frame identification experiment: prototype vs exemplar, and pre-trained only vs fine-tuned

Pre-trained Exemplar The pre-trained exemplar model is the model with the simplest architecture, as there are no explicit representations of frames that need to be built. A pre-trained embedding for the test LU is compared with the pre-trained embeddings of other LUs in the training data ('exemplars'). The model then assigns the test LU the frame label of its nearest exemplar. We use the standard cosine similarity metric to determine the distance between the test LU and exemplars.

Fine-tuned Exemplar In the fine-tuned exemplar model, embeddings should be tuned s.th. predicate embeddings that evoke the same frame should be more similar than predicates that evoke different frames. The fine-tuned exemplar model poses the fine-tuning step as a binary classification problem, where the aim is for the model to decide, based on predicates in their sentential context, whether they are evoking the same frame or different frames. The input to the model consists of the concatentated embeddings of two predicates in context, where the BERT model adds a [SEP] token between the two sentences, shown in Table 5.2.1.

Input Sequence	The doctor treated the patient [SEP] He got apples
Label	different

Table 5.1.: Example of the fine-tuned exemplar model input where the input is a negative sample. The first sentence has the LU *treat*, which evokes a Medical_Intervention frame, while the LU in the second sentence, *got*, evokes the Commerce_Buy frame.

The loss function is a simple cross-entropy loss, where the prediction label is compared against the gold label, where the binary labels are "same/different". To create the negative input sentences, we use the frame lexicon to extract a list of frame candidates for each target predicate. We then sample a negative input predicate from the set of frame candidates, where the negative predicate evokes a different frame. If the target predicate has only a single frame, we randomly select a negative instance from the entire frame lexicon. For each predicate in the training data, we randomly sample two positive instances an two negative instances. At prediction time, we pair a test predicate with all the instances of frame candidates for that predicate. We classify the test predicate with the frame whose instances have the highest "same"-frame probability.

Prototype Models

Both prototype models below construct a representation of frames from all the known instances in the data. These instances are combined to form a single frame embedding, which is then used for the classification of new LUs for frame assignment.

Pre-trained Prototype In the pre-trained only prototype model, we take each frame and compute an unweighted centroid of all the pre-trained, contextualized representations of its instances. These instances are each of the LUs that evoke the frame and their context. The centroid is therefore the frame's 'prototype', constructed without any expert linguistic knowledge or preconceived notion of that frame, since it is composed of representations that come out of a model trained over large, unlabeled corpora.

Fine-tuned Prototype In the fine-tuned prototype model, we take the pre-trained embeddings of LUs and tune those embeddings with data that has been annotated for frames. Classification is set up as a token-level task where each token is assigned a frame at prediction time. The input to the model is therefore a sequence of tokens, and gold class labels are the frames that are evoked in the annotations, shown in Table 5.2.1. Tokens that do not evoke any frames are given a non-label, "O" class assignment. At test time, we take all LUs that are known to evoke a frame and evaluate the frame label over those tokens.

Input Sequence	The	doctor	treated	the	patient
Label	O	Medical_professionals	Medical_Intervention	O	Medical_Interaction_Scenario

Table 5.2.: Input sentence for the fine-tuned prototype model

The loss function of the model is a straightforward cross-entropy where each token is assigned a frame; as is the case with the pre-trained only model, no global optimization takes place.

5.2.2. Experiment Setup: Dataset and Hyperparameters

To compare with standard frame identification systems, we use the Das et al., 2014 train/test/dev split over the FrameNet 1.5 full-text annotations (explained in further detail in Chapter 3).

Since its release, there has been an exponential growth in pre-trained models for BERT; for our experiments, we take the standard English pre-trained model, which was trained on the BooksCorpus and Wikipedia. The BERT-large, cased model is trained with the highest number of layers (24), hidden units (1024), and self-attention heads (16). According to authors of the BERT model, performance over the contextualized, pre-trained representations is shown to improve when the final n layers are concatenated (Devlin et al., 2019), which we do for n=4 in our pre-trained (prototype and exemplar) model experiments.

The fine-tuned models all re-use hyperparameters from the pre-trained model, and the final layer is a standard softmax classification layer. The architecture of the BERT model has a high computational cost due to the attention mechanisms, so the fine-tuned models have a limitation on the sequence length. We set the length to 180 for the prototype model since the prototype model uses single sentences as input, and we increase the sequence length to 200 for the exemplar model, as it takes two text sequences as input. These lengths are chosen to keep as many tokens as possible in training while remaining computationally efficient. The fine-tuning model additionally requires several epochs to converge, and we found that the fine-tuned, prototype model requires 30 epochs. Although most of the BERT classification tasks only required 3-4 epochs to converge (Devlin et al., 2019), the frame identification model incorporates significantly more class labels – there are over 1k classes for frame identification, while the tasks described in other

BERT models had at most 4 classes. For the fine-tuned exemplar model, with only 2 class labels ("same/different"), we run the model for 5 training epochs before convergence.

5.3. Results

Performance of the models are evaluated over two different accuracy metrics. The first accuracy assumes the existence of the FrameNet frame lexicon, where the accuracy is evaluated over the frame that is the most probable given the set of frame candidates for the test LU. However, the lexicon accuracy includes cases of LUs that only evoke a single frame, making the task trivial; the second metric takes the lexicon accuracy only for predicates that are ambiguous, meaning they are capable of evoking more than one frame.

Table 5.3 shows results of the classifiers, where the best performing model overall is the fine-tuned prototype model. With an accuracy of 91.26%, the model outperformed the previous state of the art. In general, both prototype models seem to be better suited to the task of frame identification than the exemplar models, at least on the currently available frame annotations. It is no surprise that fine-tuning improves performance over pre-trained models, which we see in the results that frame identification indeed profits from task-based tuning. However, the pre-trained model alone works surprisingly well for frame classification; in fact, the simple pre-trained prototype model performs competitively to the prior SEMAFOR model (Das et al., 2014), which was trained with extensive linguistic and ontological features (see Chapter 3). Thus, the pre-trained vector space models do have a claim to robust performance, and the relationships learned over the pre-trained, unsupervised model seem to correspond well to the types of semantic relationships found in frames.

	Model	Full Lexicon	Ambiguous
Prior	(Das et al., 2014)	83.60	69.19
	(Hermann et al., 2014)	88.41	73.10
	(Hartmann et al., 2017)	87.63	73.80
	(Botschen et al., 2018)	88.82	75.28
	Model	**Full Lexicon**	**Ambiguous**
Ours	Pre-train only Exemplar	82.52	64.44
	Pre-train only Prototype	84.67	69.18
	Fine-tuned Exemplar	84.09	65.06
	Fine-tuned Prototype	**91.26**	**80.77**

Table 5.3.: Accuracy results for frame identification in both exemplar and prototype models, with and without fine-tuning.

Sentence Length We can further analyze the results of the models by comparing sentence length, as the BERT model incorporates long-range dependencies via its attention mechanisms. The model also includes bidirectionality, as context before and after the target predicate are used for prediction, introducing a highly contextualized representation. However, because of its architecture, longer sentences in the model should be prone to noise, which is exactly what we see in the pre-trained exemplar and prototype models in Figure 5.2.

While accuracy improves with higher sentence length in the fine-tuned, prototype model, sentence length seems to have a negative effect on the fine-tuned exemplar model. This is a surprising effect to come out of the fine-tuning process, but it could be an indication that the exemplar model struggles with

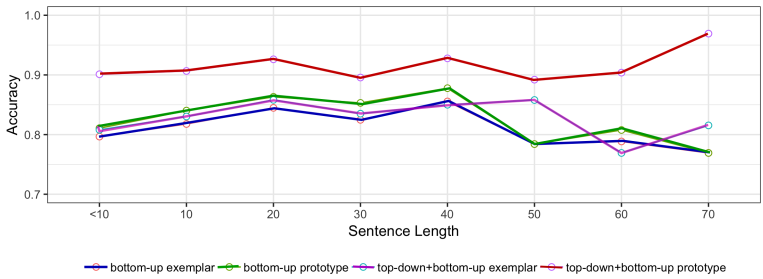


Figure 5.2.: Impact of sentence length on accuracy

determining which tokens in the input sequence are being compared – in many cases, multiple frames can be evoked in the same sentence so it is not explicit in the fine-tuned exemplar model which frames are under comparison. One modification to this architecture might be to add a positional index for the target predicate at input, following the design of (Tan and Na, 2019), so that the model limits the scope of the attention to the context around the target predicate.

Top Frames and Predicates Frame and predicate performance varies under the different model types, where certain frames benefit from the fine-tuning process more than others. We can speculate that frames that do not benefit much from fine-tuning are those that already form a compact topic cluster, where the lexical units are more likely to be specific and have a lower number of possible senses.

Table 5.4 shows the top 5 most accurate frames and predicates from the fine-tuned prototype model, which are compared against their accuracies in the other exemplar models and the pre-train only prototype model. For the top frames, the prototype model shows a clear performance gain for the CAPA-

		Fine-tuned		*Pre-train only*	
		Prototype	Exemplar	Prototype	Exemplar
Frame	CAPABILITY	1.00	0.73	0.48	0.73
	POSSESSION	1.00	0.94	0.92	0.81
	WEAPON	1.00	0.97	0.98	1.00
	LOCATIVE_RELATION	0.97	0.84	0.89	0.79
	TEMPORAL_COLLOCATION	0.89	0.76	0.76	0.71
Predicate	people.n	1.00	1.00	0.97	0.97
	know.v	0.96	0.89	0.90	0.87
	have.v	0.92	0.85	0.85	0.74
	in.prep	0.91	0.69	0.80	0.59
	can.v	0.91	0.59	0.29	0.62

Table 5.4.: Accuracies for top 5 frames and predicates from the fine-tuned prototype model and their accuracies in the other prototype and exemplar models.

BILITY frame when fine-tuning occurs, although the exemplar model shows no gain in the fine-tuning model. This presumably is because the frame consists of function words (*can.v, able.a*) which have several senses and require more fine-tuning. This is a similar case with the other frames LOCATIVE_RELATION, AND TEMPORAL_COLLOCATION, except for the WEAPON frame, which performs equally well across all model types. This is attributable to the fact that the predicates in the WEAPON frame form a coherent semantic class with low ambiguity (*bomb.n, missile.n, shotgun.n*). The best performing predicates show a similar trend, where function words (*can.v, in.prep*) benefit significantly with fine-tuning while certain content words such as *people.n* show a relatively stable performance across all models. This suggests that the highly frequent function words have contextualized embeddings that are more spread out over the high dimensional space and benefit more from top-down input during the fine-tuning process.

5.4. Summary

In this chapter, we experiment with new model designs for the frame identification task using the BERT architecture. The results lead to two important conclusions: first, prototypes and exemplars perform somewhat equivalently over the FrameNet corpus in models where there is no fine-tuning procedure. When fine-tuning is available, the prototype model becomes the more successful design – in other words, constructing a representation of a frame yields better performance in the classification task. This is likely the case because fine-tuning moves the representation of frame instances farther apart, thus potentially creating a boundary with a larger margin across frame classes.

Second, although fine-tuning is preferable, there are still decent frame identification predictions that come out of the pre-trained embeddings alone. This is strong indication that the types of semantic relationships that come directly from embeddings learned over large, unstructured corpora correspond well to the semantic relationships in the classes of frames that are defined by linguists. These two results taken together suggest that the prototype model, in the context of representation learning, can effectively learn categories that are based on non-linear representations - thus making them more successful in categorization tasks. Taking the results from this study, the next chapter demonstrates how frame representations in the prototype model design can be created for frame identification models in different languages.

6. Frame Selection for Multilingual Frame Identification

We described in Chapter 4 how embeddings can be used to measure frame comparability across languages. We can now ask whether the knowledge that certain frames are more comparable than others can benefit cross-lingual models of frames. This chapter revisits the hypothesis from Chapter 4, in which certain frames are more comparable than others, and that assessments of frame comparability can be made by aggregating instances of frames from data in different languages. Specifically, high usage-based, cross-lingual comparability should be reflected in the frame embeddings, where frames with a similar meaning will have similar embeddings in different languages. If they do, then it might be possible to re-use frame annotated data in one language for the benefit of a frame identification system in another.

Because frame identification is a task that requires annotated data, much of the initial efforts in frame semantic parsing were restricted to English, the first language with a large, frame-annotated resource. However, research in cross-lingual frame semantics has grown over the years, and several teams across the globe have created resources for frame semantic parsing in different languages (see Chapter 2 for a discussion of available FrameNets in different languages). Unfortunately, for languages that still lack frame-annotated data, the creation of these resources is both time and resource-intensive. For example, annotations for German, French, and English (the languages we consider in our experiments), took years to develop; as a consequence, there are only a select number of languages with which we can carry out experiments in cross-lingual frame identification. This is also a strong motivation for pursuing research which would reduce the need for frame annotations for a new, target language.

The experiments in this chapter investigate to what extent manual annotation efforts for a target language can be reduced by assuming that a) there are comparable frames in different languages, and that b) one can use the available annotations of frames from a different language to train a system in a target language. We hypothesize that frame identification in German, French, and English will benefit from cross-lingual training, and that we can predict which frames will maximally benefit a cross-lingual system. We predict frames for cross-lingual training based on their clusterability (McCarthy, Apidianaki, and Erk, 2016): how distinguishable a frame is from other frames in high dimensional space, and how spread out its instances are.

The frame prediction models described in this chapter have similar goals to active learning, where we are aiming to select data that will improve model training with less data, but the sample selection methods we use differ from those in traditional active learning. While the goal in active learning is to select training examples to maximize the performance of a classification system, our primary goal is to improve frame classification in order to learn something about frames and their transferability across languages. Therefore, traditional selection methods that would boost classifier performance (such as uncertainty sampling (Lewis, 1995) or expected error reduction (Roy and McCallum, 2001),) could presumably lead to greater classifier gains by targeting individual instances for training, but would not produce insights into frames across languages as a whole. Therefore, we formulate our experiments in this chapter around the selection of frames, not instances, to better understand the comparability of those frames across languages. More specifically, our goal is to build computational models that complement linguistic analyses of frames across languages.

6.0.1. Cross-lingual Frame Prediction with Clusterability Metrics

This chapter introduces three separate experiments in cross-lingual frame identification; contributions of each experiment are shown in Figure 6.1. We test frame clusterability with both cross-lingual (Experiment 2) and monolingual (Experiment 3) measures, where the former assumes the availability of data from two or more languages and the latter only assumes the availability of data from one language.

In Experiment 1, we train a frame identification system on data from other, **supplementary (S)** language(s) and apply it as-is to a **target (T)** language. This experiment requires no training data annotations for the target language, and establishes a baseline for how well a frame identification model can learn frames without any target language data (also called **zero-shot** learning). We call this zero-shot scenario the *S only* experiment, as we do not use any target language data in training. In addition to establishing a baseline for cross-lingual frame identification, in Experiment 1 we can make certain observations about which frames can be sufficiently learned with no target language data whatsoever, which can be illustrative of their usage-based cross-lingual similarity.

Experiment 2 essentially provides an upper bound in terms of frame se-

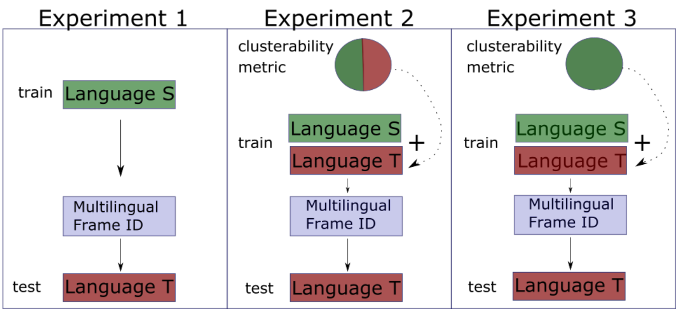


Figure 6.1.: Overview of three experiments we conduct in Chapter 6. Experiment 1 establishes a baseline with only supplementary language S used in training a system for a target language T. Experiment 2 uses features of cross-lingual frame clusterability from S and T to predict frames from T to select for training. Experiment 3 uses features of monolingual frame clusterability from S to predict frames from T to select for training.

lection for cross-lingual frame identification. In this experiment, we test the scenario where annotations for the target language are available; if annotations for both target and supplementary languages were available, we aim to find the frames that maximally improve performance in target language training. In this experiment, we add a small amount of annotations from the target language to the *S only* data. We will establish a process of frame selection where we choose the frames from the target language to add to the available, supplementary language data. Given the challenges in developing resources for frame identification, we would like to minimize the amount of training data in the target language while maximizing the performance of the cross-lingual frame identification model. This experiment adopts the scenario of an "annotation budget" for the target language, where we assume that only a select number of annotations can be created. The goal is to select target language frames that should be annotated and to test whether that selection process improves the performance of a frame identification system over a simple, random frame selection. In this design, we do not conduct actual annotations, but rather we imagine a scenario in which a certain number of annotations could be made available for the target language and we select this number from the existing target language annotations. The selection process in Experiment 2 is based on a *cross-lingual* clusterability metric. More specifically, the cluster properties of a frame within and across the target and supplementary languages are used to estimate how beneficial that frame would be in training a frame identification system for the target language. This experiment establishes our frame selection procedure and setup, and it illustrates how cross-lingual properties of a frame corresponds to cross-lingual models of frame identification for different languages.

Finally, Experiment 3 is the most realistic scenario, where annotations are available for the supplementary language but no target language annotations are available. Here, we adopt the standpoint of a researcher who wants to create a new frame semantic resource for a target language but who only has existing annotations from different languages as a starting point. Similar to Experiment 2, the goal of this experiment is to reduce the need for manual annotations in the target language by deliberately selecting target language frames based on certain criteria. However, unlike Experiment 2, our selection criteria only uses the annotations we assume to be available to the researcher - the supplementary language data. The clusterability metric in this case is based on *monolingual* frame clusterability, where the language we base our selection on is the other, supplementary language and not the target language. Experiment 3 is an alternative version of Experiment 2 but is refined for the production setting where no target language data is available.

In terms of cross-lingual frame comparability, we would hypothesize that frames that are selected for annotation in the target language in Experiments 2 and 3 should be those whose usage-based comparability is low, presumably because the frames that performed well from the supplementary languages (*S only* condition) have a higher comparability and therefore require less annotations from the target language. We additionally ask the question of how overlapping the frames' usage-based comparability is with its lexicographic comparability. That is, can we correlate results from our selection method with the frames' lexicographic comparability to determine if the frames that are selected for transfer based on their usage are also the frames that have a low overlap in their definitions? Much of the work in this chapter, predominately Experiment 3, was reported in our article "Improving Multilingual Frame Identification by Estimating Frame Transferability", which was published in the journal of Linguistic Issues of Language Technology (LiLT) 2022.

6.1. Related Work

Thus far, research in frame identification has focused on training models in a monolingual fashion, where models are trained for a language with resources curated specifically for that language (Gildea and Jurafsky, 2002; Das et al., 2010b; Erk and Pado, 2006; Johansson, Heppin, and Kokkinakis, 2012). Johannsen, Alonso, and Søgaard, 2015 were the first to tackle training complete frame semantic parsers in a multilingual setup. Authors of this model created small corpora of frame-semantic resources for nine languages over Wikipedia and Twitter texts. They use pre-trained multilingual embeddings as features for a linear classifier, and ultimately find that a single, multilingual model can perform better than multiple monolingual models. Similar findings have appeared in related NLP tasks; (Mulcaire, Swayamdipta, and Smith, 2018) combine existing data from 7 languages to perform semantic role labeling and find that, even with inconsistencies and divergences in the label set, concatenation of multilingual data leads to performance gains over monolingual models.

Other work has focused on cross-lingual labeling of semantic roles where no annotated data in the target language was assumed to exist. Recent models have incorporated multilingual embeddings in a neural network for predicting semantic role labels in a target language across parallel, multilingual corpora (Cai and Lapata, 2020). Instead of earlier work which aligned syntactic arguments for label projection across parallel text (Padó and Lapata, 2005), the authors predict labels over a LSTM with multilingual embeddings and a noise filter for predicted role labels in the target language. Other recent work performs semantic role labeling for a target language without target annotations or even parallel corpora; instead, (Daza and Frank, 2019b) use an encoder-decoder with multilingual embeddings to translate and jointly label sentences in the source and target languages. Following this work, (Daza and Frank, 2020) automatically translate and project Propbank role labels from English to three target languages. These automatic translations are then validated and corrected by human annotators, and several shifts in translation were observed to affect the target language role and predicate labels. These shifts included nominalization of verbal predicates, light verb constructions (Hwang et al., 2010), separable verb prefixes for German (discussed in more detail in Section 4.3.1), and shifts in the argument heads for named entities.

Prior work in cross-lingual frame identification has not directly addressed whether current multilingual embeddings capture frame similarity across languages; we evaluate how well embeddings learned in a joint, multilingual space can be used to predict frames in different languages without alignment or translations. Additionally, we seek to gain a high-level insight into frame (dis-)similarities across annotations, where the multilingual embedding space can also reveal how linguists annotated similar frames in different languages.

6.1.1. Active Learning and Frame Selection

Our annotation budget scenario is a pool-based sampling strategy (Lewis, 1995) in active learning, where a large pool of unlabeled data is assumed to be available (in this case, the target language data), and data is selected for annotation based on a pre-defined informative measure (in this case, our performance prediction model). This strategy produces a list of frames that are ranked for their comparability across the language pairs of interest.

6.2. Background

6.2.1. Multilingual BERT (mBERT)

BERT (see Chapter 3) provides multilingual, pre-trained embeddings that were learned over a large corpus of Wikipedia entries for over 100 languages. Because the size of each language's Wikipedia data varied, languages with the smallest datasets risked being overwhelmed in the model by languages with larger datasets. To mitigate this potential risk, the authors balanced the data by weighting the probability of the language (that is, the amount of the language's data over all the Wikipedia data) by a factor of 0.7. Using this weighting, well-resourced languages were under-sampled and low resource languages were over-sampled during training [1].

[1] https://github.com/google-research/bert/blob/master/multilingual.md

BERT's multilingual embeddings make it possible to fine-tune a single model over multiple languages and test that same model naively on other, unseen languages. This makes adaptation of the system straightforward for cross-lingual, zero-shot experiments - that is, experiments where a system is trained on one language and applied as-is to another. In fact, multilingual BERT (mBERT) has been shown to perform well on many different cross-lingual tasks (Wu and Dredze, 2019; Pires and Garrette, 2019; Karthikeyan et al., 2019). Additional setups such as multilingual corpus concatenation, where text from different languages can be concatenated and used to train a single, multilingual system, is also possible within this framework.

6.2.2. Lexicographic Frame Comparability in English, German, and French FrameNets

Frame-annotated resources have emerged across the globe, many with consistent frame inventories to the English Berkeley FrameNet (see Chapter 3). Most teams have taken the additional step of modifying frames originally developed in English for their own purposes, where modifications can include divergences in the core semantic roles, or the frame's LUs. The JUSTIFYING frame, shown in Figure 6.2, diverges lexicographically across English, German, and French, where the set of LUs and core semantic roles are different across each resource.

Other frames diverge lexicographically because of different approaches to the boundaries between certain frame classes, where some research groups chose to separate or merge frame classes for their own purposes. One example of this is the GRANT_PERMISSION and DENY_PERMISSION frames, where the boundary between the frames differs across all three resources. While English FrameNet 1.5 has several frames related to directives (GRANT_PERMISSION,

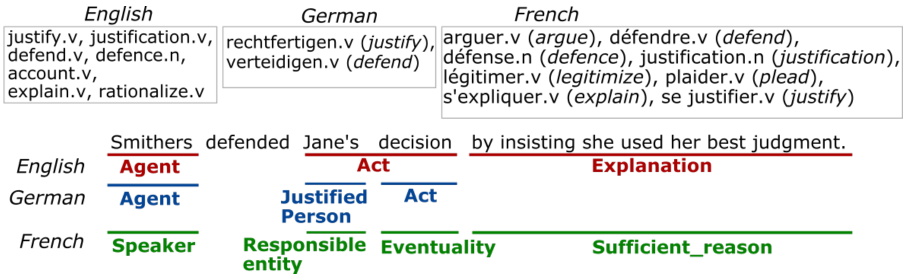


Figure 6.2.: Lexicographic entry for the JUSTIFYING frame in English, German (SALSA), and French FrameNets where each project has defined the frame differently.

DENY_PERMISSION, PERMITTING, and PROHIBITING), the German SALSA only adopts the definition for DENY_PERMISSION. Other German lexical units that would evoke these frames in English, such as *Erlaubnis.n (permission.n)*, are in SALSA-specific frames such as SCHLIESSEN2, a German-specific frame which describes an authorized individual giving permission to entry to a certain place, and *schliessen.v (close down)* which is used in a frame describing the act of an authority starting or ending an official function (the SCHLIESSEN1 frame). The French FrameNet conflates all four English frames to a large, single frame called GRANT_PERMISSION-PERMITTING where lexical units relating to giving and denying permission are all categorized under the same frame. A visual of these distinctions are given in Figure 6.3 below.

The reasons for these dissimilarities in frame structures can be typological (discussed in Section 2.3) as well as subjective, where the interpretation of the frame differs for each FrameNet team depending on their purpose (Baker and Lorenzi, 2020). In this section, we compare the frame lexicons from English to German and English to French. Because English frame definitions were

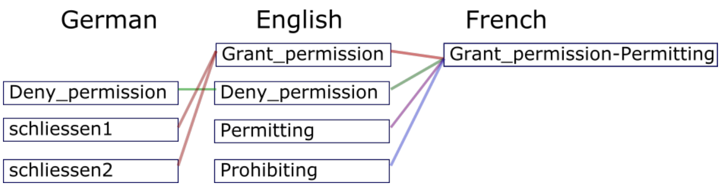


Figure 6.3.: Lexicographic entry for the GRANT_PERMISSION frame in English, German (SALSA), and French FrameNets where each project diverges in their frame class boundaries.

heavily used in both projects, we will describe how the frame lexicons diverged from the English definitions and which definitions were retained. This gives an overall sense of the frame overlap for the languages we consider.

Frames in German SALSA vs English FrameNet

There are three types of frames currently found in the SALSA lexicon: a) frames whose definitions are taken directly from the English FrameNet, b) frames whose definitions were altered from the original English FrameNet for a better fit with German predicates, and c) so-called proto-frames which are frames defined for a specific predicate sense in German that does not have a corresponding English frame (see Section 2.4.2). In terms of frames in a), FrameNet frames that were adopted for German as-is include a broad spectrum taken from the FrameNet hierarchy, including frame distinctions as specific as JUDGMENT_COMMUNICATION and JUDGMENT_DIRECT_ADDRESS.

Other FrameNet frames underwent alterations to their definition to better suit the German language. These include the addition of semantic roles, such as the PARTICIPATION frame, which SALSA added the CONTRIBUTION role for explaining what the PARTICIPANT does in the event. The CONTRIBUTION role does not exist in the English PARTICIPATION frame, but it can be found for German verbs like *beteiligen* and *teilnehmen* (*to participate, take part in*):

(27) [An der neugegründeten GmbH $_{\text{INSTITUTION}}$] [*beteiligten* $_{\text{PARTICIPATION}}$]
[The newly founded company] [*participate*]
sich [die Gesellschafter $_{\text{PARTICIPANT}}$] mit
themselves [the shareholders] with
[Beträgen zwischen 2000 und 30,000 Mark $_{\text{CONTRIBUTION}}$]
[amounts between 2000 and 30,000 marks]
"The shareholders contributed to the new company in amounts ranging between 2000 and 30,000 marks."

French FrameNet vs English FrameNet

Frames from the English FrameNet v1.5 were the basis for the definitions in the French ASFALDA FrameNet project. There are two frame types specified in the French lexicon, a) frames that are taken directly from the English definition, where only the core and certain non-core roles are used in the definition of the frame, and b) frames that have been modified from the original English to adapt to the French language. These include frames where the core roles are modified and/or the frame was merged with another FrameNet frame. There are additional instances of frames that have different LUs in French; one instance is the COMMERCE_BUY frame, where the definition is relatively consistent across English and German, but the French definition includes the *acquirer.v (acquire.v)* predicate, which evokes the GETTING frame in English:

(28) a. [Matra $_{\text{BUYER}}$] [*acquiert* $_{\text{COMMERCE\_BUY}}$] [la division $_{\text{GOODS}}$].
b. [Matra $_{\text{RECIPIENT}}$] [*acquired* $_{\text{GETTING}}$] [the division $_{\text{THEME}}$].

Differences show in Examples 28a and 28b are a result of the methods used in building the French FrameNet, where the GETTING frame is not part of the French FrameNet. This is largely due to the domains chosen for annotation in French (discussed in Section 2.4.2), where the COMMERCE_BUY frame is part of the commercial transactions domain. It is likely the case that the English GETTING frame is too broadly defined, including many LUs that are not fitting with commercial transactions (ex, *win.v*) and was therefore not incorporated in the French FrameNet. Therefore, LUs relating to commerce that are annotated for the GETTING frame in English are annotated under the COMMERCE_BUY frame in French.

6.3. Experiments

We now describe our three experiments in multilingual frame identification, where we aim to understand how well a frame identification system will work with minimal training data in the target language. Experiment 1 is the baseline condition where no target language annotations are available for training. Experiments 2 and 3 present our work on frame selection for the target language, where we use cross-lingual and monolingual selection metrics, respectively. In these experiments, the goal of the frame selection process is to define a procedure which will ultimately maximize the benefit of a frame identification system for a target language with as minimal target language annotations as possible. Therefore, we give ourselves an "annotation budget" and ask, in our frame selection method, how to best fill that budget.

6.3.1. Experiment 1: S only Baseline

Experiment 1 evaluates how well a frame identification system will work in a target language using only data from available, supplementary language(s). This experiment provides a lower bound for the target language, where the hope is that certain frames will be similar enough in their usage that data in a different language is still useful for the target.

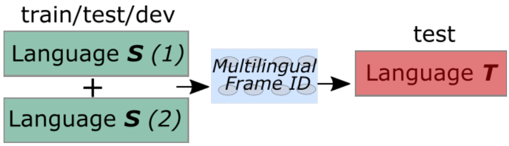


Figure 6.4.: *S only* baseline setup; Multilingual frame ID model is trained on the entire data set of the supplementary, S language(s) and evaluated on the test dataset of the target T language

Experiment 1 Methods

For our multilingual frame identification system, we use the best performing architecture from experiments described in Chapter 5: the top-down, prototype model. To extend the system to a multilingual setup, we replace the English pre-trained embedding model with BERT's multilingual pre-trained model (mBERT).

We then ask how well a frame identification system for a target language can perform using *only* data from S language(s); specifically, we can ask whether we can get by with training a frame identification system without needing annotations from the target language. To answer this question, we take all the frames that overlap between language pairs $\langle S, T \rangle$ and only use the S annotations for training. We then evaluate over the entire test set for the target, providing a zero-shot, *S only* baseline for cases in which no target language annotations are seen in training. Results of the *S only* set up were taken over a single run.

Experiment 1 Setup

Datasets Our multilingual frame identification experiments use data from English FrameNet, German SALSA, and French FrameNet projects. In Experiment 1, we use only the frames that overlap between the target and the supplementary language(s) for training in the target language. These include frames that have been modified for French and German from English; although language-specific modifications means that the overlap in frame definition is not perfect, we are imagining a scenario in our experiments where we want to use all possible labeled frame data in a supplementary language(s) available. Therefore, we include modified frames in hopes that, for example, a frame modified for French might still be useful for learning about its counterpart in English. Table 6.1 gives the numbers of overlapping frames and respective annotations for the supplementary and target languages.

Evaluation Metrics In all three experiments, we evaluate performance of frame identification models with the standard classifier accuracy, which is the performance of the model where each frame in the lexicon is considered at test time. This evaluation metric was chosen above the lexicon accuracy metric (described in Section 5.3) as we did not want to assume the availability of a fully developed frame lexicon for each language. More specifically, while the lexicon accuracy assumes knowledge of all the frames that a single predicate can evoke, the classifier accuracy only assumes that there is an established set of frames for the target language.

We also report on Macro F1 score for frames in the test data, which we use for analyzing frame performance. The F1 score of each frame is used for our frame selection prediction process (described later in Section 6.3.2) as it provides a measure of how each individual frame is performing in the frame

T	*S*	*# Frames*	*# Annotations*	*Classifier*	*Macro F1*
EN	DE	271	25k	17.88	5.91
	FR	68	11k	3.99	2.6
	DE+FR	294	36k	14.27	6.16
DE	EN	271	14k	38.79	20.97
	FR	34	8k	5.57	1.87
	EN+FR	271	22k	27.99	14.21
FR	DE	34	7k	12.42	6.44
	EN	68	2k	17.38	14.83
	DE+EN	70	9k	18.55	18.11

Table 6.1.: Classifier accuracy (*Classifier*) and Macro F1 scores (*Macro F1*) for *S only* (Experiment 1) baselines, where data from the *Supplementary* language(s) is used to train a *Target* language frame identification system.

identification model.

Data used for training/development/testing is described in Chapter 2, Table 2.1.

Experiment 1 Results

Results for the *S only* baseline are shown in Table 6.1. For the most part, results show that a frame identification system without any target language data for training does not perform well. This is especially the case for certain languages where the available, supplementary language data is quite low (esp., *Target* is EN/DE, *Supplementary* is FR). The most promising results comes from the English-German language pairs, where German is the *Target* language and English is the *Supplementary* language. This is likely due to the fact that this language pair also has the highest amount of overlapping frames, and therefore has significantly more training data; we discuss this further when we address the upper bound below (Section 6.3.1). Additionally, the higher diversity of frames available for classification in English and German (both have a frame inventory of over 1k compared to 100 in French,) with a higher number of frames with similar definitions across the EN-DE pair (see Table 2.1) likely contribute to higher *S only* scores for the language pair.

However, it is interesting that simple concatenation of frames from multiple supplementary languages (DE+FR, EN+FR, DE+EN) does not yield the highest results in 2 out of the 3 cases. In fact, in these cases, it appears as though the addition of more supplementary language frames makes it more difficult for the model to learn boundaries of certain frame classes.

Upper bound for S only frames Table 6.1 shows that there are only a subset of frames that have lexicographic overlap between the target and supplementary languages in the *S only* model. This suggests that there should be some hypothetical upper bound in *S only* performance, since a number of target language frames are never seen in training due to the fact that they don't exist in the supplementary language annotations. Table 6.2 gives this *hypothetical*, best-case scenario upper bound for the Macro F1 score over the test set for each target language. In this setup, we compute an upper bound score for the frames that have been seen in the training data (i.e., those that exist in both the S and T languages and have been seen in *S only* training). At test time, we take the T test set and we assume all other frames not in that set are misclassified, while the frames in that set are correctly classified. This score therefore reflects an upper bound for the target language frames that are also in the supplementary language.

Target	*Supplementary*		
	DE	*FR*	*DE+FR*
EN	69.04	6.85	70.07
	EN	*FR*	*EN+FR*
DE	76.19	8.98	76.19
	DE	*EN*	*DE+EN*
FR	27.59	37.77	37.77

Table 6.2.: Results of *S only* upper bound, where we score the test set with a maximum recall; that is, the highest possible score when only the frames that overlap across the *Supplementary* language(s) and the *Target* are recalled

Results from Table 6.2 show that the upper bound for certain language pairs (specifically, cases where FR is the *S* language) is actually quite low. Therefore, the results of our *S only* baseline in Table 6.1, despite the low performance, suggest that frames are indeed learned from the supplementary language annotations.

In two cases, frames were combined from two supplementary language pairs (EN+FR and DE+FR), yet the upper bound is identical to *EN* as the only supplementary language. Not surprisingly, this result shows that all French and German frames that overlap are found in the English frame set. The frame annotations do, however, impact the *S only* score in training; we see from the *S only* results that the French frames degrade the performance of a model where German is the target language (with *EN* as the only *S* language, results are 38.79; with *EN+FR* as the *S* languages, results are 27.99), while the German frames improve the French model (with *EN* as the only *S* language, results are 17.38; with *DE+EN* as the *S* languages, results are 18.55).

We explore further the frames that caused the most and least improvement in the *S only* training for different language pairs.

Frames with high/low performance in S only training Table 6.3 shows frames that have the highest and lowest F1 scores across language pairs in the *S only* baseline. The F1 scores for these frames are either a perfect classification score in the *High* columns, or frames that have *Low* F1 scores had no examples of correctly classified frames and therefore scored zero.

Frames that perform best are those that are strongly related semantically, which we would expect form a tight semantic cluster; for example, the KINSHIP frame has predicates all relating to familial relationships (*brother, sister, grandfather*, etc.), and the PEOPLE frame has LUs relating to human beings (*man, woman, child*, etc.). Frames that have the lowest cross-lingual F1 scores are those whose definitions have been modified, such as the JUSTIFYING frame shown in Figure 6.2, and the GRANT_PERMISSION frame shown in Figure 6.3. Additionally, certain frames with low F-scores have highly ambiguous LUs, such as the COMING_TO_BELIEVE frame which is evoked by *find.v* – a LU that can evoke 8 different frames. Surprisingly, despite the differences in the frame's LUs cross-lingually, the COMMERCE_BUY frame, shown in Example 28, nonetheless had one of the highest F1 scores in the *S only* training for French when German was the *S* language.

German			
English		*French*	
High	*Low*	*High*	*Low*
Membership Expertise Reason	People Filling Coming_to_believe	Calendric_unit Communication_response Questioning	Topic Text_creation Judgment_direct_address
English			
German		*French*	
High	*Low*	*High*	*Low*
Part_whole People_by_age Kinship	Taking_time Grant_permission Justifying	Exporting Commercial_transaction Attributed_information	Encoding Deserving Regard
French			
English		*German*	
High	*Low*	*High*	*Low*
Deciding Importing Commercial_transaction	Proving Cause_earning Contacting	Commerce_buy Commerce_sell Referring_by_name	Justifying Coming_to_believe Communication_response

Table 6.3.: Frames with top F1 scores from the supplementary only model (*High*) and the lowest F1 scores (*Low*). Each table (English, French, German) shows the target (T) language, and columns underneath are the supplementary (S) language which was used in training.

Experiment 1: Summary

Experiment 1 establishes a baseline for our further experiments in cross-lingual frame identification. Our work in building the *S only*, zero-shot baseline in Experiment 1 has also illustrated how well multilingual embeddings learned over large-scale language models can be used for learning frames across languages. We find that, although the performance of the *S only* setup is not at the level of a production setting, the results suggest there are frames that are learned in this setting, many of which are those with nominal predicates. Finally, it comes as some surprise that the addition of frame annotations from multiple languages (instead of a single, supplementary language), at times degrades rather than improves performance. This result strongly suggests that, when building a frame identification system for a target language without annotated data, it is worthwhile to select the frames from a supplementary language to maximally benefit performance, as certain frame annotations from target languages will not lead to improvements in the target language model. We build on these results in Experiments 2 and 3, where we propose a metric for predicting which frame annotations will be most beneficial to a target language frame identification system.

6.3.2. Experiment 2: Frame Selection with Cross-Lingual Predictors

The goal of a multilingual frame selection system is to maximize the benefit to a target language frame identification system using only a small number of target language annotations, given the existing frame annotations from other, supplementary languages. In this experiment, we improve upon the baseline system from Experiment 1 by adding a modest amount of data from the target language in training. Our selection procedure in Experiment 2 involves a cross-lingual frame selection metric, meaning that our metric assumes the availability of existing frame annotations in both the target and supplementary language(s). Essentially, the metric is an estimate of how beneficial a certain frame should be for multilingual training, inspired by work on the separability of word senses (McCarthy, Apidianaki, and Erk, 2016). In this experiment, a frame is selected if it a) has good *within-language* separability, that is, within a language, instances of the same frame cluster together closely and are well distinguished from instances of other frames, and b) it has poor *cross-lingual* separability, that is, instances of a frame in one language are not well distinguished from instances of the same frame in another language.

Experiment 2 Methods

Multilingual Frame Selection Frame selection consists of three steps: 1) building a baseline system for evaluating cross-lingual frame performance, 2) predicting frames to select from the target language annotations to add back for cross-lingual training, and 3) using the selected target language frames plus supplementary language annotations to train a frame identification system for the target language and evaluate the results. Figure 6.5 shows the three steps of the selection procedure.

Step 1: Building the Baseline System for Frame Selection The first step in the frame selection method is establishing frame performance in multilingual training. This step involves first contrasting two frame identification models. We take one model from Experiment 1 - that is, a target language frame identification system trained solely on the supplementary language data (*S only*), and in the second, we train a system on supplementary data plus

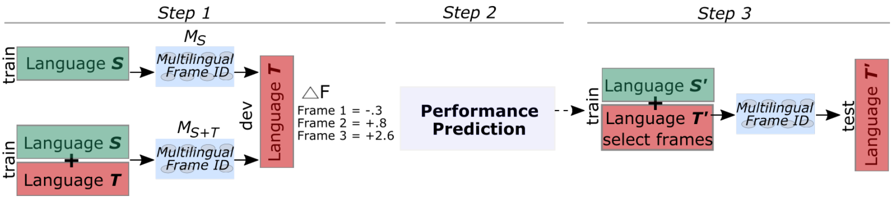


Figure 6.5.: Overview of frame selection procedure. Step 1: build a baseline, Step 2: learn the frame selection prediction model, Step 3: add selected frames to cross-lingual training and evaluate. Language pairs $\langle S, T \rangle$ are the supplementary and target languages used to train the frame selection performance predictor model, and a different language pair $\langle S', T' \rangle$ are used in testing.

all the training data in the target language. Comparing the performance of the supplementary+target and S *only* models gives us a concrete measure of a frame's improvement when target language annotations are added to the training process, which we call ΔF – a frame's transfer potential. A frame with a high ΔF score has a high improvement with additional target data, and is therefore a candidate we would want to select for target annotations. A frame with a low ΔF score can mean one of two things: either a) the S *only* model sufficiently learned the frame and therefore target language annotations provide little benefit, or b) the frame is possibly difficult to learn, even with target language data. It is also possible for a frame to have a negative ΔF score, in which case the addition of the frame's annotations causes the frame identifier to perform worse than its S *only* counterpart. This ΔF score is then used for the next step in the procedure: the learning of a performance prediction model that can automatically estimate frame performance for new target language frames.

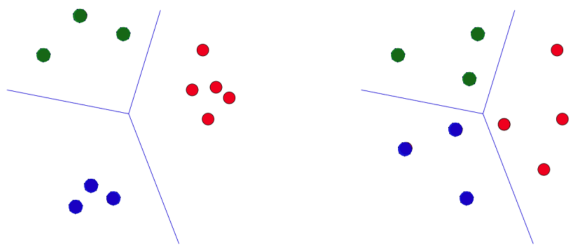
Frame performance in Step 1 is taken over a frame's F1 score. While we report our final evaluation results (Table 6.4) on classifier accuracy, which is a standard evaluation metric for frame identification systems, using F1 scores in our selection process ensures that the frame selection is not biased by highly frequent frames.

Step 2: Performance Prediction The goal of building a performance prediction model is to use the *properties* of a frame to predict its performance in cross-lingual training. Specifically, we want to be able to estimate how much improvement a frame will make when its target language annotations are added to existing supplementary annotations when training our frame identification model. We can then select frames that are predicted to have the highest improvement and only use those in training.

When we apply the performance prediction model to frames in a new target language (T' in Figure 6.5), we get an estimate of the frames' transfer potential ($\tilde{\Delta}F$) that we can use for selecting the frames that should be added back for annotation. We rank the target language frames by their predicted $\tilde{\Delta}F$ score and take the frames that have the highest $\tilde{\Delta}F$. These selected frames are added to supplementary language data and a frame identification model is trained over this dataset (see Step 2 in Figure 6.5). The frame identification system is then evaluated over the test data for the target language.

One important aspect of this design is that we fit the performance prediction model on a single $\langle S, T \rangle$ language pair (a supplementary language and a target language) for which we assume annotations already exist, but we apply this model as-is to all the other language pairs. Applying the performance prediction model on other language pairs speaks to the generalization of the model to other, unseen languages. For fitting the performance prediction model, we use the development set of English and German, where English is the target language and German is the supplementary language. All other language pairs assume that the target language is unseen and has no available annotations prior to selection.

Frame Properties: Frame Clusterability Recall that our approach to estimating how beneficial a certain frame should be for multilingual training, i.e., its transfer potential, is inspired by clusterability of word senses (McCarthy, Apidianaki, and Erk, 2016). The original study quantifies how easily the instances of different senses of a word can be separated into clusters via a measure called the **variance ratio**.



(a) More clusterable data

(b) Less clusterable data

Figure 6.6.: Clusterability; figure modified from (McCarthy, Apidianaki, and Erk, 2016). In our implementation, each dot represents an instance of a frame and colors represent a frame.

The variance ratio (Zhang, 2001) assumes that a good clustering includes instances of a word that are close to a centroid, and clusters that are far apart from one another. A centroid is measured by a weighted average of all the instances of a word in corpora:

$$centroid(Y) = \frac{1}{|Y|} \sum_{y \in Y} y \tag{6.1}$$

Using this centroid, the variance of the cluster Y can be measured with Equation 6.2:

$$\sigma^2 = \frac{1}{|Y|} \sum_{y \in Y} ||y - centroid(Y)||^2 \tag{6.2}$$

The second part of this metric involves the variance of the clusters - that is, how far apart the clusters are from one another. The distance of two clusters (Y_i and Y) is measured by Equation 6.3:

$$B(C) = ||centroid(Y_i) - centroid(Y)||^2 \tag{6.3}$$

The final variance ratio over a dataset with k clusters is the average of the $B(C)$ over the average of the σ^2 for all the clusters in the data:

$$VarianceRatio = \sum_{i=1}^{k} B(C) / \sum_{i=1}^{k} \sigma^2 \tag{6.4}$$

While these measures were first proposed to assess the clusterability of word senses in a corpus, we extend this idea to the case of frames and frame annotated data. When applying the above metrics to frames, we can then revise Equations 6.1-6.2 s.th. the centroid of a frame F is the weighted average of all of its annotated instances f, and Equation 6.3 is the distance of one frame (F_i) to another frame (F) in the annotated corpus. Consequently, a frame's clusterability is defined as (1) the variance of the instances of a frame f, where instances of the same frame should form a tight cluster, and (2) the distance from one frame F_i to another frame F. A frame that forms a coherent cluster, such as example (a) in Figure 6.6, should have a low variance amongst its instances and a high distance from other frames.

In Experiment 2, we apply this idea to cross-lingual frame selection, where we hypothesize that a frame with a high ΔF score (that is, a frame predicted to maximally benefit the target language) will ideally (1) have good **intra-lingual** clusterability, that is, within a language, instances of the same frame cluster with form a tight group which is easily separable from instances of other frames, and (2) it has poor **inter-lingual** clusterability, that is, instances of a frame in one language are not well distinguished from instances of the same frame in another language. If frames have poor inter-lingual clusterability, it signals that the frame actually has a high usage-based comparability; that is, its instances appear to be similar across both languages.

Intra-lingual Clusterability We adapt Equations 6.2-6.4 for three intra-lingual clusterability metrics. These measure the frame's clusterability within the target language. The IWC score revises Equation 6.2 where we measure the distance between a frame cluster F and its frame instances f:

$$IWC = \frac{1}{|F|} \sum_{f \in F} ||f - centroid(F)||^2 \tag{6.5}$$

The second feature is the intra-lingual, between-frame clusterability which adapts Equation 6.3 s.th. the distance of the frame (F_i) to its nearest neighboring frame (F):

$$IBC = ||centroid(F_i) - centroid(F)||^2 \tag{6.6}$$

Finally, the ratio of IBC and IWC scores is a final intra-clusterability score:

$$ICl(F) = IBC(F)/IWC(F) \tag{6.7}$$

This measure is computed for each frame in the target language. If the score is high, it would indicate that a specific frame forms a tight, cohesive cluster in the target language. The three factors that we use for intra-lingual clusterability are the IWC, IBC, and ICl.

Inter-lingual Clusterability Inter-lingual factors are properties of the frames' clustering across languages, where properties of the same frame across the target and supplementary languages are considered. We modify the variance score in Equation 6.2 for the cross-lingual case by combining the variance of a frame in the target language F and the variance of its counterpart in the other, supplementary language F' for an inter-lingual, within-frame clusterability score XWC:

$$XWC = \frac{|F|}{|F| + |F'|} WC(F) + \frac{|F'|}{|F| + |F'|} WC(F') \tag{6.8}$$

For the between-frame clusterability, the intra-lingual IBC score compared a target language frame F to its nearest neighbor frame in the same target language. The cross-lingual, or inter-lingual clusterability of the frame, XBC, modifies the between-frame clusterability score BC in Equation 6.9 s.th. F is again a frame in the target language, but instead of a nearest neighbor, the frame we compare F against (F') is the same frame in the other, supplementary language:

$$XBC(F) = |centroid(F) - centroid(F')|^2 \tag{6.9}$$

Finally, both the XWC and XBC scores are combined for a final, inter-lingual score:

$$XCl(F) = XBC(F)/XWC(F) \tag{6.10}$$

Altogether, the performance prediction model takes three features from the inter-lingual clusterability scores: the XBC, XWC, and XCl.

Step 3: Applying Frame Selection for Frame Identification The result of Step 2 is a list of frames ranked by their $\tilde{\Delta}F$ score. The final step of the frame selection process is to add the annotations of the frames with the highest $\tilde{\Delta}F$ from the target language to the existing supplementary annotations. Recall that the goal of the frame selection process is to define a procedure which will ultimately maximize the benefit of a frame identification system for a target language with as minimal target language annotations as possible. Therefore, we set an "annotation budget" of x number of training instances to add back from the target language annotations. We fill this budget by taking frame annotations from the top of the list and working downwards until the budget has been filled.

After the selected frames from the target language have been added to the supplementary language annotations, we train a frame identification model and evaluate on the (thus far unseen) target language test set. Recall that the language pair we use for fitting the performance predictor model is English and German, where English is the target language and German is supplementary. For the entire selection process, we only use the development set to make predictions for English. Therefore, even in the case of English as a target language, the evaluation is conducted over the test data, which has still not been previously seen.

Experiment 2 Setup

As we described in Section 6.3, we start with a *S only* baseline and add a moderate budget of annotations from T. For this budget, we choose sizes of 5k and 10k instances to balance the amount of data needed to properly train a system with as little data as possible. For all the target languages, there are frames with a large number of annotations, some of which have over 1k instances (POLITICAL_LOCALES in the German SALSA annotations has nearly 1k annotations for a single predicate, *Land.n*). In this case, we want to restrict the number of instances the classifier has seen for each frame so that the 5k/10k annotation budget is not filled almost exclusively with annotations from a single frame. We therefore restrict the instances of a frame to 200, in order to have a balance of a high number of instances in training with a decent variety of frames. We randomly select this 200 instances from the available instances of the frame, where the number 200 was selected to balance the goals of adding a substantial number of frames and a substantial number of instances per frame.

We train the frame selection model on the language pair $\langle S, T \rangle$ with the largest number of overlapping frames: ⟨German, English⟩. The frame selection model is then applied as-is to all other language pairs $\langle S', T' \rangle$ for frame selection. More specifically, this means that the model predicts frames for other language pairs (such as ⟨French,German⟩, ⟨English,German⟩) based on clusterability metrics from the ⟨German, English⟩ pair. This demonstrates the generalizability of frame selection to new, unseen languages. Frame identification models are then trained with this modified (S'+selected T') data and evaluated over unseen T' test data. In addition to a single, supplementary language, mBERT's multilingual embeddings enable us to build models using data from multiple supplementary languages simultaneously. Therefore, we also combine S' language data for a T' language, where we combine the ranked list of frames from each individual $\langle S', T' \rangle$ pairs and take the top predicted frames from this combined set as our selected T' frames. Each of these models were trained over a single run (as opposed to averaged over multiple runs).

Random Frame Selection Baselines We compare the frame selection approach in Experiment 2 with a random frame selection baseline. In this *Random* selection baseline, frames are randomly sampled from a pool of all possible target language frames. Identical to the constraints applied to our performance prediction selection method, we only take 200 instances of each frame from the target language training data for learning the model, and we concatenate this set of randomly selected frame annotations to the *supplementary-only* dataset for training. Random results are averaged over two runs for each condition.

Experiment 2 Results

Table 6.4 shows the results of our frame selection with cross-lingual predictors. It is clear from these results that the selection is worthwhile, as performance is significantly improved next to a random selection baseline.

Inter-lingual Clusterability Scores Table 6.5 shows the cross-lingual clusterability scores for frames with the highest and lowest *XCl* scores (shown in Equation 6.10). These are frames whose clusterability is highest and lowest in the cross-lingual vector space, meaning they either appear far apart from one another in the cross-lingual space (highest clusterability scores), or they appear overlapping or highly similar in the cross-lingual space (lowest clusterability scores). These scores tell us how (dis)-similar the frames appear in the multilingual space; a low cross-lingual clusterability implies that the frames are *not* separable in the cross-lingual space, indicating that they are appearing in similar contexts in their annotated data. Frames that have a

		Cross-lingual Frame Selection			
		Random		*Cross-lingual*	
T	*S*	+5k	+10k	+5k	+10k
EN	DE	33.99	55.61	**47.32**	**67.24**
	FR	27.74	52.13	**27.91**	**59.09**
	DE+FR	30.80	59.14	**53.87**	**64.86**
DE	EN	23.33	37.06	**30.68**	**45.20**
	FR	18.45	35.27	**21.72**	**41.93**
	EN+FR	22.55	39.54	**41.04**	**45.55**
FR	DE	37.88	54.97	**48.65**	**64.34**
	EN	25.58	59.26	**48.82**	**63.57**
	DE+EN	25.76	59.25	**60.41**	**64.98**

Table 6.4.: Results for Experiment 2 cross-lingual frame selection, where performance in the selection method (*Cross-lingual*) is consistently (and, in many cases, significantly) improved over a random frame selection baseline (*Random*).

high cross-lingual clusterability are those that are separable, and therefore distinct, in the cross-lingual space.

Results in Table 6.5 are based on scores in the development set, which were used to train the performance prediction model in Step 2 (see Section 6.3.2 of the frame selection procedure). The Justifying frame, shown in Figure 6.2, has one of the highest *XCl* scores across English/French and German/French. Other frames that are more dissimilar in cross-lingual space include Trial. This corresponds nicely to observations in linguistic research, where the language of criminal processes has been discussed in prior work on cross-lingual frame semantics as being particularly challenging. In fact, "legal corpora...face a double challenge: (i) the equivalence of lexical units, and (ii) the equivalence of legal concepts" ((Bertoldi and Chishman, 2012), pg. 5). We find that the cross-lingual clusterability results reveal that criminal processes indeed have a high degree of separability in the multilingual BERT embedding space.

Alternatively, several frames with the lowest *XCl* scores, that is, frames that are most similar in the cross-lingual space, are more general frames (CAUSATION, STATEMENT) which include a broad range of lexical units (*put.v*, *raise.v*, *result.n*, *make.v* for CAUSATION, and *exclaim.v*, *remark.v*, and *say.v* for STATEMENT).

T only Results Following our selection procedure, we then ask whether the results we see in Table 6.5 would be the same in set up where only 5k/10k of the target language annotations were used; more specifically, we ask whether the supplementary language annotations are providing any additional gains to our target language frame identification. In Table 6.6, we report results on frame identification systems trained only on target language annotations, where the target language frame annotations are identical to those shown for the *Cross-lingual* selection in Table 6.5. Results show that, when we do a direct comparison, the supplementary language annotations do indeed improve the overall performance of the system – in many cases, the performance is significantly boosted with the addition of supplementary annotations: with +5k target annotations, the addition of both supplementary (DE+FR) languages shows 54.40 accuracy for English versus 24.92 accuracy without; similarly, both supplementary (EN+FR) annotations for German shows a 43.24 accuracy versus 18.64 without; most significantly, for French we see that the addi-

Language Pair		*Frames*	
		Highest XCl Score	*Lowest XCl Score*
EN	*DE*	SEEKING	STATEMENT
		ACHIEVING_FIRST	POLITICAL_LOCALES
		TRIAL	KILLING
EN	*FR*	QUARRELING	CAUSATION
		JUSTIFYING	PURPOSE
		COMMUNICATION_RESPONSE	REQUEST
DE	*FR*	JUDGMENT_DIRECT_ADDRESS	REQUEST
		JUSTIFYING	CAUSATION
		REFERRING_BY_NAME	JUDGMENT_COMMUNICATION

Table 6.5.: Frames from the development set with the highest and lowest Interlingual clusterability scores (XCl). Recall that frames with the highest scores are most distinct, that is, dissimilar, in the cross-lingual space. Frames that have the lowest *XCl* scores have the most overlap in the cross-lingual space.

EN		*DE*		*FR*	
+5k	+10k	+5k	+10k	+5k	+10k
24.92	47.75	18.64	34.56	25.12	48.05

Table 6.6.: *Target-only* model results: test set classifier accuracy when training only on target language.

tional (DE+EN) annotations leads to accuracy of 59.99 versus 25.12 without. Therefore, we see in these results that the addition of supplementary annotations are not inconsequential to the final performance of the target language model.

Experiment 2 Summary

Experiment 2 outlines our frame selection method for multilingual frame identification. In this experiment, we predict frames that will maximally improve a target language frame identification system, given the availability of supplementary language annotations. Our selection method involves clusterability metrics, both within (intra-lingual) and across (inter-lingual) languages. Because Experiment 2 uses cross-lingual factors in assessing which frames to select for target language training, it assumes that both target and supplementary data is available at the time of the selection. To this end, we pose this experiment as an upper bound for our frame selection method.

Results demonstrate that frame selection with cross-lingual clusterability outperforms a random selection of frames, and in most cases, performance is markedly improved over this random baseline. Frames that are selected for target language training include frames that have diverging lexicographic definitions (such as JUSTIFYING), and frames that reflect culturally nuanced topics, such as TRIAL. Though the objective of the cross-lingual clusterability metric is to predict frames for cross-lingual frame identification, the results of frames with high and low cross-lingual clusterability scores validate that (dis)similarities in frames across languages are indeed captured by the multilingual, pre-trained BERT embeddings.

We continue our experiments in frame selection with a scenario that is more plausible than using cross-lingual factors; that is, a scenario in which frame annotations are not available for the target language, and only clustering metrics from the supplementary language can be used for selection.

6.3.3. Experiment 3: Frame Selection with Monolingual (S only) Predictors

Experiment 2 established our setup for frame selection; however, its premise was ultimately a best case scenario - one which assumed that frame annotations would be available for both the supplementary and target languages. In practice, this scenario is not entirely realistic; researchers looking to build a frame identification system in a new, target language will have to grapple with the lack of annotations in that target language. Experiment 3 therefore removes the expectation that frame annotations will be available for the target language. In this frame selection procedure, we predict frame performance solely based on properties of the frame annotations that we assume already exist - that is, annotations from the supplementary language.

The multilingual frame selection setup described in Section 6.3.2 outlined three steps of our frame selection procedure: building the baseline system, predicting frame performance, and applying frame selection to the frame identification model for the target language (shown in Figure 6.5). Experiment 3 follows the same three steps with a sole change to the frame performance prediction in step 2. In this step, we replace the cross-lingual frame properties (Section 6.3.2) that were used to predict the frames with the highest transfer potential with monolingual frame properties from the supplementary language. The performance prediction model is then trained over these frame properties and tested as described in step 3.

Experiment 3 Methods

Monolingual (S only) Frame Selection We use three features from the supplementary language annotations to compute our monolingual clusterability frame selection. First, a frame's monolingual within-frame variance,

$MWC(F)$, is computed following Equation 6.2 where the frame is from the supplementary language. The monolingual between-frame distance, MBC, compares a frame in with its nearest neighbor, following Equation 6.3. Finally, the overall monolingual clusterability of the frame, $MCl(F)$, is the ratio of the MWC and MBC:

$$MCl(F) = MBC(F)/MWC(F) \tag{6.11}$$

We use three monolingual features from the supplementary language, MWC, MBC, and MCl to select frames and train our performance prediction model in Experiment 3. Frame identification models for the target language (with the added frames that were predicted to improve the classifier) were trained over single runs.

Experiment 3 Results & Analysis

The results of our frame selection with monolingual features is shown in Table 6.7. It is clear from these results that monolingual frame selection (*Monolingual* in Table 6.7) is worthwhile, as performance with frames that are selected through our method described in Section 6.3.3 consistently outperform a random selection of frames. This means, that when researchers seek to build a new frame-annotated resource in their target language, it is worth their time to determine from the onset which frames they ought to annotate given frame performance on other, supplementary language(s).

We then compare the results of the monolingual frame selection from the results of cross-lingual frame selection in Experiment 2, where the results are taken directly from Table 6.4. We expect that the results from the cross-lingual selection to be better than results using supplementary (monolingual) language frame selection, as the selection method for the cross-lingual case

		Frame Selection					
		Random		*Cross-lingual*		*Monolingual*	
T	*S*	+5k	+10k	+5k	+10k	+5k	+10k
EN	DE	33.99	55.61	**47.32**	**67.24**	35.76	60.77
	FR	27.74	52.13	27.91	**59.09**	**28.29**	57.94
	DE+FR	30.80	59.14	53.87	**64.86**	**54.40**	62.36
DE	EN	23.33	37.06	**30.68**	**45.20**	24.16	42.75
	FR	18.45	35.27	**21.72**	**41.93**	21.66	40.25
	EN+FR	22.55	39.54	41.04	**45.55**	**43.24**	43.88
FR	DE	37.88	54.97	**48.65**	**64.34**	47.03	62.09
	EN	25.58	59.26	**48.82**	**63.57**	46.65	59.66
	DE+EN	25.76	59.25	**60.41**	**64.98**	59.99	61.77

Table 6.7.: Results for Experiment 3 monolingual frame selection, where performance in the selection method (*Monolingual*) is compared to a random selection of frames (*Random*), as well as results from Experiment 2 (*Cross-lingual*). *Monolingual* results that are higher than *Cross-lingual* counterparts are in bold for visibility as we would predict the *Cross-lingual* results are higher across all conditions; however, it can be observed that in nearly all cases, the *Cross-lingual* frame selection does perform better.

is an upper bound for frame selection. Although this is the case, we find strong performance using supplementary language annotations alone for target language frame identification.

Coefficients of Frame Properties in the Frame Selection Model We have defined in our frame selection procedure several frame properties, that is, features of the frame's clusterability (see Section 6.3.2), for predicting how well a frame will improve a target language frame identification system. Analyzing the coefficients of these properties allows us to better understand their relationship to the frame identification task, where we can directly determine which factors benefit the multilingual training the most. Table 6.9 shows the coefficients learned for the various frame properties and their significance (p values). We initially hypothesized that 1) the more dissimilar the instances of the frame are to one another, the more it will profit from target language annotation, and 2) the smaller the distance between a frame and its nearest neighbor, the more it will profit from target language annotation. The coefficients confirm only the first hypothesis, where a high within-frame variance is very significant in predicting a higher ΔF. The other two properties are not significantly correlated with ΔF, although the nearest neighbor distance is negatively correlated with a high ΔF score, which is consistent with our second hypothesis. However, when we do an analysis of the collinearity of the predictors, we see that the factors are highly collinear. This is no surprise as the Clusterability (MCl(F)) measure is defined by the other two factors in the model.

When we re-run the model using only the nearest neighbor distance (MBC(F)) and within-frame variance (MWC(F)), shown below in Table, we find that again the model supports our first hypothesis, where the more dissimilar the instances are to one another the better it predicts a frame's potential improvement to a multilingual frame ID system. The second hypothesis is not supported in this model, indicating that variance is a more relevant factor in predicting a frame's transfer potential.

Predictor	*Coeff.*	*Std. Error*	*p value*
Clusterability (*MCl(F)*)	0.17	0.20	>0.15
nearest neighbor distance (*MBC(F)*)	-0.0009	0.001	>0.10
within-frame variance (*MWC(F)*)	0.0008	0.0002	<0.05
Correlation	Clusterability	Distance	Variance
	7.42	9.497	1.862

Table 6.8.: Estimated coefficients, standard error, and p-values for embedding-based predictors, as learned by the performance prediction linear regression model. Multicollinearity of the factors is given, showing that the MCl and MBC scores are highly correlated in the model.

Predictor	*Coeff.*	*Std. Error*	*p value*
nearest neighbor distance (*MBC(F)*)	0.005	0.07	>0.10
within-frame variance (*MWC(F)*)	0.21	0.07	<0.01

Table 6.9.: Estimated coefficients, standard error, and p-values for nearest neighbor distance and within-frame variance predictors, as learned by the performance prediction linear regression model.

Frame Selection and Lexicographic Frame Comparability The results of our frame selection experiment show that a deliberate selection of frames improves results over a random selection of frames for the target language. Our selection method is based on the embeddings of a frame's instances - namely, their variance to one another and distance to other frames in the training data. This essentially means that our selection method is a usage-based measure of frames across languages, where the embeddings are representations of frames based on their use in context.

We now ask whether our usage-based selection method corresponds to the lexicographic similarity of the frames across languages. First, we ask whether the frames that perform well in the *S only* baseline are also the frames with high lexicographic comparability. We hypothesize that the frames in the *S only* baseline will have a high lexicographic comparability as they should be more similar, and therefore more useful, in *S only* training. Next, we can ask whether the frames selected for multilingual training have a lower lexicographic comparability, indicating that more language-specific frames improve the *S only* baseline more than frames with high lexicographic similarity.

Much of our discussion of frame semantic resources in English, German, and French has focused on the available annotations and annotation methods. Below, we will cover some of the similarities and differences that can be found in the lexicographic definitions of the frames in English, German, and French, where the English lexicon was the starting point for the German and French resources. Following this overview, we proceed to the results of our study where we compare the performance of frames with high and low lexicographic comparability in the *S only* and frame selection models.

We compare the frame performance of frames that have high lexicographic comparability ("same") and frames with low lexicographic comparability ("modified") from the supplementary-only model and add the average improvement of each type of frame when selected for training using our monolingual frame selection method.

Figure 6.7 shows the difference in frames with high/low lexicographic comparability and their improvement to a cross-lingual (usage-based) frame identification system. These are only the frames that have been selected by our +10k annotation model, where language pairs are shown in the form $\langle T - S \rangle$

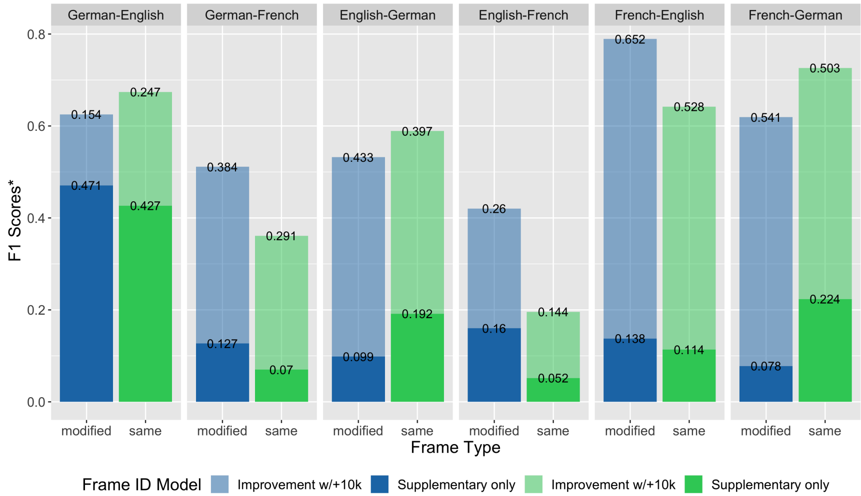


Figure 6.7.: Selected Frame Performance: correlation between similarities in the frame definitions ("Frame Type") and model performance ("F1 Scores")

(e.g., DE-EN is DE as T and EN as S) where results are tested over target language test data. For the *supplementary only* condition (dark bars), we report absolute F1 scores for performance, while "Improvement w/+10k" shows the average increase in F1 score (light bars) after the frame type was added. For frames in the supplementary-only condition, where we test to determine whether the frames with high/low lexicographic comparability performed better or worse with supplementary data alone, we find that performance is mixed.

The results from the monolingual frame selection process showed that, for all language pairs except one (EN-DE), the frames with a greater improvement were those that had lower lexicographic comparability. This result is consistent with what we would predict for the frame selection: that frames whose lexicographic definitions diverge would provide greater gains to the target language when additional, (target) language-specific annotations are added to the training process. Because frames with low lexicographic comparability are those whose definitions are more closely catered to the target language, we would expect that the system would require more language-specific data to learn these frames.

However, results of the *S only* condition in Figure 6.7 were not as clear; we would normally expect that the frames whose definitions diverged would perform worse in the supplementary only training. This was only true for some of the frame pairs, whereas other language pairs (EN-FR, DE-FR, FR-EN and DE-EN) showed a greater improvement over modified frames. There are several potential explanations for this result; one explanation is that the *S only* condition did not control for the number of training instances per frame - allowing some frames to have a large number of annotations in training while others had less (recall from Section 6.3.2 that the frame selection process restricts the number of annotations to 200 instances per selected frame). A second explanation is that, in certain cases, the modified version includes semantic roles whose names are different across the resources but function similarly in the annotations themselves. For instance, the French COMMERCIAL_TRANSACTION is the highest performing frame in the EN-FR language pair and the frame differs lexicographically across English and French. However, the modification involved creating two new roles, PARTICIPANT_1 and PARTICIPANT_2, for the BUYER and SELLER roles in cases where the participants are not syntactically marked as to who the SELLER or BUYER is. While this is relevant for semantic role labeling, it is less likely to affect frame prediction across languages.

Experiment 3 Summary

Experiment 3 presents our study on frame selection using data from supplementary language(s) for an unseen target language. The goal of this study is to determine whether the clusterability of supplementary frames can be an indicator of frames that will improve the performance of a target language which lacks annotated data.

We find that the monolingual frame selection method in Experiment 3 outperforms a random selection of frames, validating the clustering metrics as an approach to frame selection. We compare the lexicographic status of the selected target frames, namely whether the frames had a high or low comparability w.r.t. their definitions. In this process, we find that frames that are predicted to maximally improve a target language are those with lower lexicographic comparability. Overall, results in this study indicate that a) knowledge of lexicographic frame comparability is beneficial when predicting which frames will maximally improve a target language frame identification, and b) clustering properties of existing, supplementary languages can be useful in predicting frames for target language annotation. We surmise that a) can be addressed via linguistic analysis, where certain frames are more likely to diverge due to differences in valency, part-of-speech, or other typological differences (Torrent et al., 2018). In terms of clustering properties, our results suggest that frames with high within-frame variance and small distance across frames in supplementary languages should be targeted for annotation.

6.4. Summary

Experiments 1-3 have each addressed different aspects of multilingual frame identification. Experiment 1 demonstrated that, in a zero-shot, multilingual frame identification setup, performance is impacted by the size of the frame inventory and diversity of frames annotated in the supplementary language. We find that the upper bound for a classifier trained only on supplementary language data can be quite low in some cases, but supplementary annotations are nonetheless useful in predicting frames for the target language given the existence of supplementary training data for those frames. Overall, we find that there are frames that have enough overlap to demonstrate applicability for a different, target language; specifically, we find that the frames that perform well in this setup are those whose LUs form a more tight semantic cluster, as opposed to frames with LUs that are presumably more spread out. Interestingly, performance of a frame in cross-lingual training doesn't show a clear correlation to its lexicographic comparability – suggesting that, given annotations that are conducted for different languages by different research groups, a frame's usage in context is, for frame identification, not directly correlated with similarities in its cross-lingual definition. We conclude that, in a *S only* baseline, *S only* results are not sufficient for any production setting.

We then move forward with our experiments on frame selection where we add a minimal amount of target language frame annotations back to target language training. In Experiments 2 and 3, we asked whether we could find an explicit, usage-based metric for selecting frames for annotation in a target language when existing annotations are available from other, supplementary languages. The ultimate goal of this work was to determine whether this selection could maximally improve the performance of a frame identification system for a target language while minimizing the annotation requirements for that language by assuming that certain frame classes are 'learnable' from other languages. We find that this is largely true; while results from a model trained only on supplementary data (Experiment 3) is not equal to performance when target language data is available for estimating frame potential (Experiment 2), using the available frame annotations from different languages does provide benefit to a system that has a reduced set of annotations in the target language. Additionally, when we intentionally select for the frames that we want to add back to the target language, we find that results are further improved, suggesting that certain frames are more useful for training a cross-lingual frame identification system than others. When we take the selected frames and correlate their frame performance with their lexicographic status – that is, whether the frame has supposedly "high" or "low" cross-lingual lexicographic comparability, we find that frames with a lower lexicographic comparability improve the target language model more. This indicates that it is worthwhile for the linguists that are creating resources for a new target language to focus more of their efforts on annotation for language-specific frames and to use the frame annotations from other languages when possible.

It should be noted that the makeup of the annotations in the languages we study in this chapter, that is, the method of frame annotation (frame-by-frame and lemma-by-lemma approaches are discussed in Chapter 3) and its subsequent effect on the number of annotations per frame and the variety of predicates that are annotated, could likely all have had an effect on the performance of the final frame identification model. As much as possible, we do control for these effects in our random sampling of 200 frame instances during our selection process.

Part IV.

Coping with Low Frame Comparability

7. Frame Paraphrases

In this chapter, we describe pairs of frames which have different lexicographic definitions and yet can appear similar in specific contexts (i.e., translations). These cases present specific challenges to cross-lingual frame semantics in NLP, as they can create difficulties in aligning frames across translations (see Section 3.3). More specifically, these cases speak to a deeper issue in frame comparability; that is, cases of high usage-based comparability - which indicate that two frames should be the same - but actually are different in other contexts, requiring their own lexicographic definitions. Explicitly identifying the factors that explain these cases is an important step in understanding frame comparability, both monolingually (paraphrases) and cross-lingually (translations).

We look at cases of frames that are mismatched over paraphrases, where sentences intend to convey the same meaning but the manner in which they do so evoke different frames. Because there is a broad spectrum of phenomena that could be considered a paraphrase, we begin our investigations by focusing on a specific type of paraphrasing called **concept-based paraphrases**. These concept-based paraphrases are paraphrases that cannot be reduced to a classic linguistic relation but rather require real-world knowledge to interpret. While paraphrasing is essentially different ways of describing the same state of affairs (Leech, 1974; Katz and Fodor, 1963), there is a range of views in terms of the conceptual knowledge that constitutes a paraphrase (Schreyer, 1978). In his early works, paraphrases were at the center of Fillmore's 'semantics of understanding', which formed the basis of early works in frame semantics (Fillmore, 1985). While the concept-based paraphrases we investigate in this chapter appear in parallel corpora, as we will attest, there is relatively little linguistic literature that attempts to formally define them.

Concept-based paraphrases can evoke so-called **frame paraphrases** (Padó, 2007; Ruppenhofer et al., 2006): frames that are different across two paraphrased sentences, where the entirety of the frame and semantic roles seems equivalent in certain contexts. Frame paraphrases can occur in both monolingual and cross-lingual texts, as we will demonstrate below (Section 7.1). In fact, many of the same monolingual frame paraphrases can also be found in translated, cross-lingual data. These cases are demonstrating a key finding, which we analyze in more detail over monolingual, English frames: that parallel text does not always equate to parallel frames.

Contributions of the analyses in this chapter include the following: 1) a classification of frame pairs that are related over a FrameNet frame-to-frame relation (***Using***) 2) a demonstration of paraphrasing over frame pairs within this classification, and 3) a formal definition of concept-based paraphrases, including side conditions that define concretely how frame alignment can be achieved across these pairs.

We pose this chapter as part of an effort to explain certain cases of poor comparability, as the linguistic analyses we conduct in this chapter illustrate how frame paraphrases can be resolved through modifications to the frame's lexicon. This chapter is thus the only place where we focus exclusively on linguistic properties of poor usage-based comparability. Work in this chapter has been published in the journal of Constructions and Frames, entitled

"FrameNet's using relation as a source of concept-based paraphrases" (Sikos and Padó, 2018a).

7.1. Frame Paraphrases

As we discussed earlier in Chapter 1, frame semantics is a particularly appealing framework for cross-lingual semantics as frames are an abstraction over surface-level variation. The ability to capture similarity in meaning over paraphrases (*Y bought a laptop from X – X sold a laptop to Y*) is one of the appeals of frame semantics as a theory of language (Fillmore, 1985), where a paraphrase is explained by the fact that the sentences are evoking the same frame. (Hasegawa et al., 2011; Ruppenhofer et al., 2006) discuss cases of paraphrasing within the same frame that are triggered by specific linguistic relations, such as antonymy (*we continued doing it – we didn't stop doing it*), voice alternations (*the management rewarded Susan – Susan was rewarded by management*), or support verb constructions (*they discussed it – they had a long discussion about it*).

In certain cases, these linguistic relations can also occur cross-lingually. (Hasegawa et al., 2011) discuss causation in Japanese and English, where many transitive verbs in English are translated into the intransitive in Japanese, largely due to the fact that Japanese is thought to prefer describing events as state-changes instead of actions with actors (Ikegami, 1991). (Hasegawa et al., 2011) gives the example of the CAUSATION frame that has been paraphrased across English and Japanese:

(29) [Better diagnosis CAUSE] [*has made* CAUSATION]
[Shindan hoohoo ga shinpo-shita koto CAUSE] [*ni.yotte* BECOMING]
[experts aware that Parkinson's can attack those under 40 EFFECT]
[40-sai.miman demo paakinson-byoo o hasshoo-suru koto ga wakatte.kita EFFECT]

"Due to the fact that diagnostic methods advanced, we've become aware that even those under 40 can have symptoms of Parkinson's."

The *has made–become* paraphrase is an instance of two lexical units that trigger frame paraphrase between Causation–Becoming across English and Japanese. The translation not only elicits two different frames for the causation, but the utterance as a whole evokes several different frames across the two languages. Therefore, there are many cases that can be found in text of paraphrases where the total meaning expressed in one sentence over multiple frames can be paraphrased with one or more separate, but related, frames. Analyzed as a whole, the translation in Example 29 is actually several different frames paraphrased (Hasegawa et al., 2011), as shown in Figure 7.1:

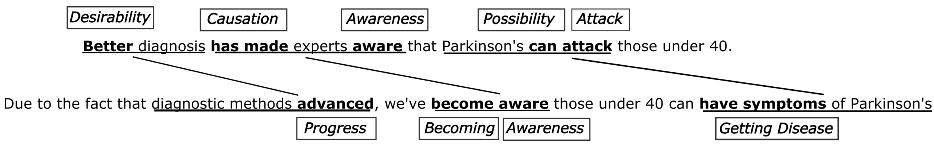


Figure 7.1.: Frame paraphrases across entire sentences, where multiple frames are paraphrased

The frames (Desirability, Attack, Getting_disease) are cases of lower usage-based comparability, where there is a poor match between the frames cross-lingually, despite the fact that the sentences are actually paraphrases. Many of the paraphrased frames (Desirability: *Better diagnosis* – Progress: *diagnostic methods advanced*) are closely related in the frame lexicon. In fact, many frame paraphrases evoke frames that are actually related via a specific frame-to-frame relation (for a description of FrameNet's frame-to-frame relations, see Section 2.4.1).

7.2. Monolingual Frame Paraphrases

While we keep the cross-lingual cases in mind for future work (see Section 7.5), this chapter focuses on monolingual frame paraphrases in English as a starting point. To identify monolingual frame paraphrases, there are two choices one could make: 1) take paraphrased data and search for frame paraphrases automatically by labeling both sentences in a paraphrase and taking cases of frame mismatch, or 2) look at existing frames in the frame lexicon and analyze whether these frames could appear in a frame paraphrase. We adopt strategy #2, as we found that much of the paraphrasing datasets do not trigger the kind of full-sentence paraphrases that would enable us to do a deeper linguistic analysis of frames.

We find that paraphrases which evoke different frames are in fact evoking frames that are closely related in the FrameNet hierarchy; importantly, we find that the frame-to-frame relations are particularly promising structures for investigating the phenomenon of frame paraphrases. Because the Berkeley FrameNet hierarchy clearly defines many frame-to-frame relations and its lexicon has been curated by linguists for decades, it is a natural starting point for characterizing and understanding when frame paraphrases can occur.

7.2.1. Frame-to-Frame Relations and Frame Paraphrases

The most straightforward case of frame paraphrases are frames that can be connected by a specific linguistic relation such as the ***Inheritance*** frame-to-frame relation, which is analogous to "is-a" ontological relationships. Frames that are related via an ***Inheritance*** relationship can be paraphrased with hyponym/hypernym swapping:

(30) a. [Myeloski $_{\text{BUYER}}$] had insisted on [*buying* $_{\text{COMMERCE\_BUY}}$] [a pizza $_{\text{GOODS}}$]

at Dominos.

b. [Myeloski SOURCE] had insisted on [*getting* GETTING] [a pizza THEME] at Dominos.

Other frame-to-frame relations that can account for frame paraphrases include the ***Inchoative_of*** and ***Causative_of*** relations, where one frame in the ***Causative_of*** relation can express a CAUSE and the other is a stative event. More nuanced frame paraphrases include frames that describe the same situation from different perspectives. These include the GET_A_JOB and HIRING frames, which are related via the ***Perspective_on*** relation:

(31) a. [I EMPLOYEE] [*signed on* GET_A_JOB] [with YouTube EMPLOYER] [to make them a new GUI TASK].

b. [YouTube EMPLOYER] [*hired* HIRING] [me EMPLOYEE] [to make them a new GUI TASK].

Perspective_on is a frame-to-frame relation that incorporates more human intuition about frame relatedness, where frames are more conceptually related instead of relying on an explicit linguistic or ontological explanation. Other types of these relations include ***Subframe*** and ***Using***, where paraphrases between frames in this group tend to rely on inference on the part of the reader:

(32) a. [They SUSPECT] were [*arrested* ARREST] [for robbery CHARGES].

b. [They SUSPECT] were put on [*trial* TRIAL] [for robbery CHARGES].

The frames in Examples 32a and 32b are connected by the ***Subframe*** relation to the CRIMINAL_PROCESS frame, and are essentially paraphrases relating to a sequence of events. These types of paraphrases are more conceptual, as they draw upon world knowledge for interpretation (see Section 2.3). In our analysis, we look at these concept-based paraphrases, triggered by the ***Using*** FrameNet relation. We focus specifically on the ***Using*** relation, which we will outline below in Section 7.3. Essentially, the ***Using*** relation is used in FrameNet as a pragmatic device to indicate partially-shared conceptual structures without having to specify a new, more abstract frame that connects them. As a consequence, this relation is quite heterogeneous and is composed of many frame pairs that can evoke conceptual paraphrases.

7.3. The Using Frame-to-Frame relation

Using is a frame-to-frame relation that explicitly connects frames that are conceptually related at an abstract level. More concretely, two definitions of ***Using*** have been proposed: the first states ***Using*** as a relation that connects "a particular frame [that] makes reference in a very general kinds of way to the structure of a more abstract, schematic frame" (Ruppenhofer et al., 2006). An example of this would be TRANSLATING – MENTAL_ACTIVITY, where MENTAL_ACTIVITY is an abstract, non-lexicalized frame covering events related to cognition. The second definition of the ***Using*** relation refers to the frames as more of a parent-child relationship where "only some of the semantic roles in the parent have a corresponding entity in the child" (Petruck and De Melo, 2012). An example is the JUDGMENT_COMMUNICATION - LABELING frame pair (paraphrased in Example 33 below), where the LABELING frame includes predicates such as *call*, *brand*, and *term*, and refers to a speaker using a label to characterize an entity.

(33) a. [He SPEAKER] **called** [him ENTITY] [a hero LABEL].

b. [He COMMUNICATOR] **praised** [him EVALUEE] [for being a hero REASON].

In the second sentence, *praise* evokes JUDGMENT_COMMUNICATION, a frame in which a communicator expresses an opinion about another person or phenomena. JUDGMENT_COMMUNICATION includes predicates such as *condemn* and *praise*, where each conveys a sentiment that can be either positive or negative. Since there is no proper mapping across rolesets in these frames, no ***Inheritance*** relation can be established. At the same time, there is a considerable intersection between frames in terms of the states of affairs that can be verbalized within either frame. Thus, FrameNet falls back to ***Using*** to capture this relationship.

These two definitions are not contradictory, but they differ enough to make the ***Using*** relation a confused mix of frames. The first definition focuses on the difference in specificity between two frames without prescribing any specific relation between their respective frame elements. The second definition is centered on the presence of a partial mapping between the frame elements without imposing specific constraints on the two frames' relative specificity. As a consequence of these two definitions, the relationship between frames that are connected via this frame-to-frame relation are, as we will demonstrate below in Section 7.3.1, heterogeneous.

Many of the frames that are connected via ***Using*** can be paraphrased even when the lexicographic comparability between the frames is very low, with no matching semantic roles and lexical units that are quite different. Example 33 clearly shows the ***Using*** relation in a paraphrase where no semantic roles or lexical units match across sentences:

Below, we analyze each of the frame pairs connected via the ***Using*** relation, in total 490 frame pairs. We find that these pairs fall into one of 5 categories, which we discuss w.r.t. the frame paraphrases that can be found in each[1].

[1]The classification of all 490 pairs can be downloaded in text and PDF from `http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/FN-using.html`

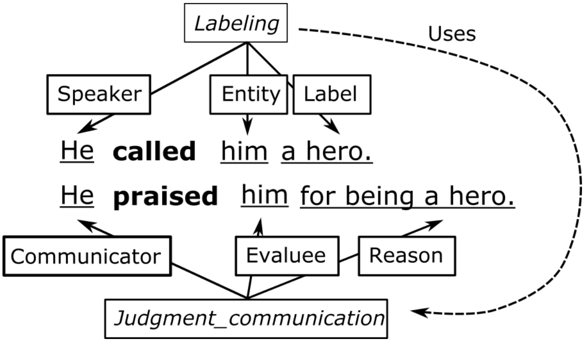


Figure 7.2.: Labeling and Judgment_Communication frame paraphrase

7.3.1. Classifying Frame Pairs Connected via Using

Our guiding principle in classifying the frame pairs connected through the ***Using*** relation was whether the frame pairs can lead to frame paraphrases; more specifically, if Frame 1 uses Frame 2, we ask whether (and how) Frame 2 can be used to paraphrase Frame 1. For instance, in Figure 7.2, the Judgment_Communication frame is Frame 2, which uses the Labeling frame, Frame 1. The procedure for determining whether, and how, Frame 1 and Frame 2 can be paraphrased was as follows:

1. Sample a predicate p from the set of predicates in Frame 1.

2. Sample a sentence s from the example sentences for p provided by FrameNet, which are mostly drawn from the British National Corpus (BNC)[2].

3. Test whether s can be paraphrased with Frame 2 by manually generating paraphrases. If this is possible without introducing additional frames,

[2] http://www.natcorp.ox.ac.uk/

we call this a *minimal paraphrase*. Figure 7.2, for example, is a minimal paraphrase.

- If there are minimal paraphrases, record semantic properties that must be met to make the paraphrase felicitous, if there are any (see Section 7.4 about side conditions)
- If there are no minimal paraphrases, but paraphrasing is possible with the introduction of other frames, record additional frames that can be used to produce a felicitous paraphrase.
- If no paraphrases are possible, record why not.

4. Repeat the process of 2-3 other s and p to obtain a comprehensive understanding of the relation between Frame 1 and Frame 2.

This process led to the classification diagrammed in Figure 7.3. As we discussed in earlier in chapter 2, frames have predicates of mixed classes, where nominal, adverbial, adjectival, and verbal predicates are all possible within the same frame. During this process, we found that a distinction can be made amongst frames that are 'eventualities' (Bach, 1981), that is, verbs, deverbal nominalizations, adjectives, and adverbs, and 'objects', expressed by common nouns.

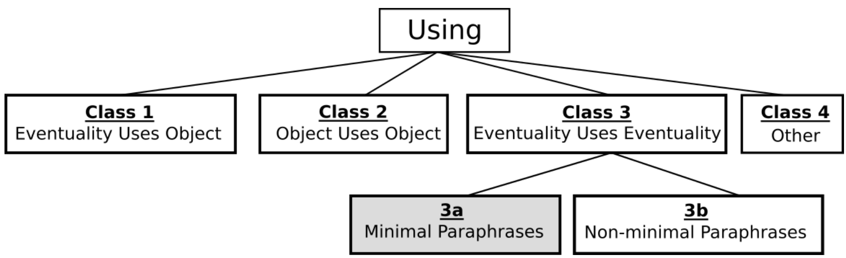


Figure 7.3.: Classification of frame pairs in the ***Using*** relation

Table 7.1 gives the frequency of each class, where class 3 (described below in Section 7.3.1) is the predominate class in the relation. A full list of the frame pairs and their classification is given in Appendix A. Examples of the types of possible paraphrases for each class is described below.

Class 1	Class 2	Class 3a	Class 3b	Class 4
Eventuality Uses Object	Object Uses Object	Eventuality Uses Eventuality: Minimal Paraphrase	Eventuality Uses Eventuality: Non-Minimal Paraphrase	Other
95	42	95	197	61

Table 7.1.: Frequency of classes in ***Using***.

Class 1: Eventuality uses Object

A common class of paraphrases in the ***Using*** relation is between two frames where one frame (Frame 2) fills a semantic role of another (Frame 1). Generally speaking, Frame 1 and Frame 2 can co-exist in the same sentence in these cases. For example, the DRESSING frame uses the ACCOUTREMENTS frame, s.th. the sentence *Jack put on his watch* can evoke both frames simultaneously: [*put on* DRESSING] and [*his watch* ACCOUTREMENTS]. Paraphrasing is also possible between these two frames when Frame 1 is expressed as a stative and an additional frame is added to Frame 2; for instance, the WEARING frame is expressed with the ACCOUTREMENTS frame in Figure 7.4, essentially making the paraphrase *being dressed in X – wearing X*.

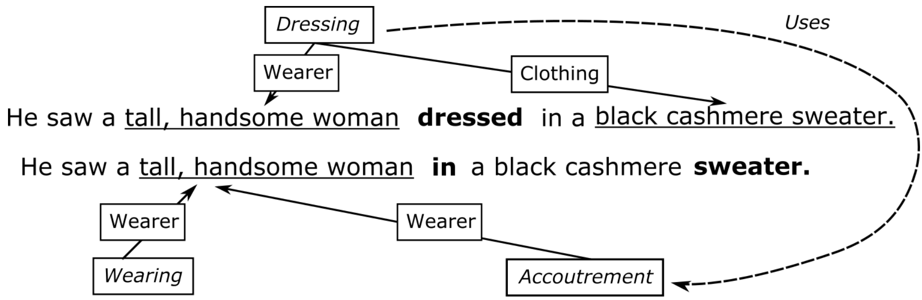


Figure 7.4.: Class 1 of ***Using*** paraphrases: Eventuality uses Object

Class 2: Object uses Object

In this class, the Frame 1 and Frame 2 are both common nouns, and the relationship between the frames is similar to the types of relations found in other ontologies such as WordNet (Fellbaum, 1998). Paraphrasing with this class includes lexical substitution, such as the frame CLOTHING_PART, which uses the frame CLOTHING in Figure 7.5

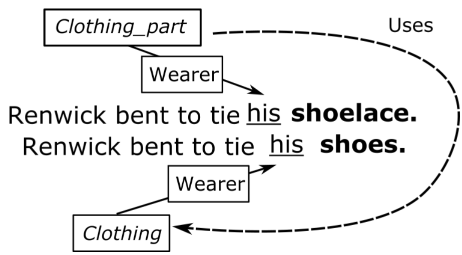


Figure 7.5.: Class 2 of ***Using*** paraphrases: Object uses Object

Class 3: Eventuality uses Eventuality

The largest class of ***Using*** frame pairs are those in which both have eventuality predicates. These can be further divided into frame pairs that are minimal paraphrases, that is, paraphrases that don't require the introduction of another frame, and those that do, which we call *non-minimal paraphrases*.

Minimal Paraphrases In Section 7.4 below, we discuss in further detail minimal paraphrases and the *side conditions* that enable paraphrasing between these two frames. Figure 7.2 is an example of minimal paraphrases between eventuality frames LABELING and JUDGMENT_COMMUNICATION. This class of frame paraphrases are of particular interest for us, as they are concept-based paraphrases. Understanding the conceptual relationship between labeling and communication in Figure 7.2, for example, requires an understanding that labeling an individual 'a hero' involves a judgment on their behavior or character, which are necessary when giving praise. Thus, common world knowledge is essential to recognizing these sentences as semantically equivalent.

Non-Minimal Paraphrase Aside from minimal paraphrases, two eventuality frames can also be paraphrased with the addition of other frames. An example of these frame paraphrases is shown in Figure 7.6 below, where the frame PRAISEWORTHINESS uses the frame JUDGMENT. JUDGMENT has the core role of COGNIZER, which is the agent that is making the judgment. However, because the PRAISEWORTHINESS frame is primarily composed of adjectival predicates such as *commendable.adj, laudable.adj,* and *admirable.adj*, the COGNIZER is expressed in the added OPINION frame.

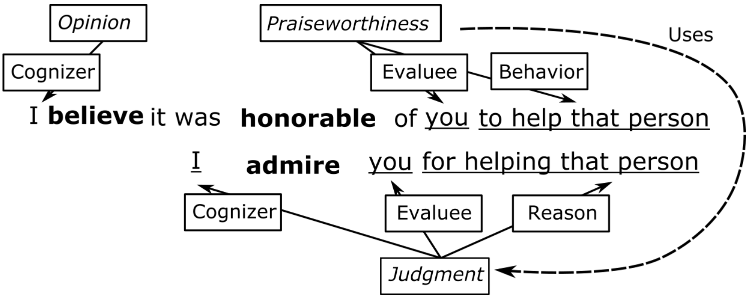


Figure 7.6.: Non-minimal Paraphrase

Class 4: Other (non-Eventuality, non-Object)

The final category of frame pairs related via ***Using*** are not lexicalized frames, and are therefore not capable of paraphrasing. In Section 2.4.1, we describe lexical and non-lexical frames in the FrameNet resource, and this final group of ***Using*** frames have at least one frame that is abstract and not realized overtly.

7.4. Side Conditions and Minimal Paraphrases

Recall from Section 7.3.1 above that minimal paraphrases are paraphrases between two eventuality frames that do not require additional frames/frame groups. By homing in on this group of ***Using*** frames, we can look specifically at the conditions that allow for a felicitous paraphrase between two distinct but conceptually related frames. These concept-based paraphrases are particularly interesting as they are not frames that are linked by a standard linguistic or ontological explanation (such as the paraphrases discussed in Section 7.2.1).

7.4.1. Side condition 1: Sentiment

The paraphrase between the JUDGMENT_COMMUNICATION and LABELING frames in Figure 7.2 is possible when the LABELING frame expresses a LABEL role with a negative or positive sentiment. For instance, the LABELING frame applies in cases where no sentiment need to be expressed, as below:

(34) [they SPEAKER] [*called* LABELING] [him ENTITY] [the "mayor" of Stuttgart LABEL]

Example 34 is not paraphrased by the JUDGMENT_COMMUNICATION frame in this case, since the JUDGMENT_COMMUNICATION frame has LUs such as *praise.v, condemn.v,* and *accuse.v*, which necessarily express sentiment. Therefore, paraphrasing between the two frames is only possible when the LABEL role in the LABELING frame expresses sentiment that matches in polarity with the JUDGMENT_COMMUNICATION frame:

(35) a. [They SPEAKER] [*called* LABELING] [her ENTITY] [a saint LABEL].

b. [They SPEAKER] [*praised* JUDGMENT_COMMUNICATION] [her EVALUEE] [for being a saint REASON].

The sentiment side condition can be noted over the LABEL role, where the sentiment polarity (positive/negative) of the LABEL role must match with the sentiment polarity of the frame-evoking JUDGMENT_COMMUNICATION predicate. This side condition can be represented more formally as an Attribute Value Matrix (AVM), as in Figure 7.7.

Additional frame pairs that have this side condition include DESIRING uses EXPERIENCER_FOCUS, RELIANCE_ON_EXPECTATION uses AWARENESS, and EVENTIVE_COGNIZER_AFFECTING uses SUBJECTIVE_INFLUENCE.

$$\begin{bmatrix} \textsc{Labeling} & \\ \text{Entity} & \boxed{1} \\ \text{Speaker} & \boxed{2} \\ \text{Label} & \boxed{3}\ [\text{Polarity}\ \ \boxed{4}\,] \end{bmatrix} \longrightarrow \begin{bmatrix} \textsc{Judgment\_Communication} & \\ \text{Predicate} & [\text{Polarity}\ \ \boxed{4}\,] \\ \text{Communicator} & \boxed{1} \\ \text{Evaluee} & \boxed{2} \\ \text{Reason} & \boxed{3} \end{bmatrix}$$

Figure 7.7.: Side Condition 1: Presence of Sentiment.

7.4.2. Side condition 2: Granularity of semantic roles

It is also the case that certain frame pairs in the ***Using*** relation can be paraphrased by expressing the filler of a single semantic role in one frame as multiple semantic roles in the other frame. For instance, the ADOPT_SELECTION frame, with LUs such as *adopt.v, assume.v* uses the CHOOSING frame, which has LUs *choose.v, elect.v*, and *pick.v.* In a paraphrase such as Example 7.4.2 below, the ATTRIBUTE and VALUE of the ADOPT_SELECTION frame can be expressed as the CHOSEN role in the CHOOSING frame:

(36) a. It is true that [baroque $_{\text{VALUE}}$] had long been [*adopted* $_{\text{ADOPT\_SELECTION}}$] [as the style $_{\text{ATTRIBUTE}}$] [for state capitols in the United States $_{\text{PURPOSE}}$].

b. It is true that [baroque style $_{\text{CHOSEN}}$] had long been [*chosen* $_{\text{CHOOSING}}$] [for the state capitols in the United States $_{\text{INHERENT\_PURPOSE}}$].

The CHOSEN role in 36b expresses the VALUE (i.e. *baroque*) as a modifier to the head noun, *style*, which is the ATTRIBUTE in 36a. This side condition differs from the first in that, the semantics of the filler for the role is not what changes but rather the condition is syntactic-semantic.The semantic role side condition can be represented as an AVM, shown in Figure 7.8:

$$\begin{bmatrix} \textsc{Adopt\_Selection} & & \\ \text{Agent} & \boxed{1} & \\ \text{Value} & \boxed{3} & [\ \text{Head}\ \ \boxed{2}] \\ \text{Attribute} & \boxed{2} & \end{bmatrix} \longrightarrow \begin{bmatrix} \textsc{Choosing} & \\ \text{Cognizer} & \boxed{1} \\ \text{Chosen} & \boxed{3} \end{bmatrix}$$

Figure 7.8.: Side Condition 2: Difference in Granularity of Semantic Roles (Split ATTRIBUTE/VALUE)

Additional frame pairs that require granularity of their role sets to be specified include ADDUCING uses STATEMENT and BEYOND_COMPARE uses SURPASSING.

7.4.3. Side condition 3: Presence of semantic roles

Finally, we found cases where the overt realization of a semantic role was necessary for a felicitous paraphrase between Frame 1 and Frame 2. For instance, the BEAT_OPPONENT frame uses the WIN_PRIZE frame, where the BEAT_OPPONENT frame can express a PRIZE as a non-core semantic role (core versus non-core semantic roles are discussed in Section 2.4.1). The PRIZE role is, however, a requirement if paraphrasing with the WIN_PRIZE frame is desired:

(37) a. [He $_{\text{WINNER}}$] [*beat* $_{\text{BEAT\_OPPONENT}}$] [the challenger $_{\text{LOSER}}$] [for the title $_{\text{PRIZE}}$].

b. [He $_{\text{COMPETITOR}}$] [*won* $_{\text{WIN\_PRIZE}}$] [the title $_{\text{PRIZE}}$] [over the challenger $_{\text{OPPONENT}}$]

The side condition that allows for the paraphrase in Example 7.4.3 is the requirement of the presence of the non-core PRIZE role, where the WINNER

$$\begin{bmatrix} \text{B}\small{\text{EAT\_OPPONENT}} & \\ \text{Winner} & \boxed{1} \\ \text{Loser} & \boxed{2} \\ \text{Prize} & \boxed{3} \neq \emptyset \end{bmatrix} \longrightarrow \begin{bmatrix} \text{W}\small{\text{IN\_PRIZE}} & \\ \text{Competitor} & \boxed{1} \\ \text{Opponent} & \boxed{2} \\ \text{Prize} & \boxed{3} \end{bmatrix}$$

Figure 7.9.: Semantic Side Condition 3: Presence of (non-core) Semantic Role.

and COMPETITOR roles are filled by the same entity, and the LOSER and OPPONENT roles are also filled by the same entity. This side condition is represented in an AVM in Figure 7.9.

7.4.4. Summary of Side Conditions

Our analyses above formally describe how certain paraphrases across two distinct frames can be achieved when specific side conditions are met. These side conditions are especially relevant for understanding cases of poor frame alignment across seemingly parallel sentences, as these cases *do not* warrant a change to the frame definitions themselves. Exploring further side conditions, and ideally identifying these side conditions computationally (see Future Work below), is a critical step in our understanding of frame comparability. Side conditions are particularly important to understanding frame comparability, as they capture the reasoning behind frame pairs with high usage-based comparability and lower lexicographic comparability.

7.5. Future Work: Cross-lingual Paraphrases

While we find that a monolingual investigation is a more straightforward starting point to investigate frame paraphrases, the knowledge we gain vis-a-vis the side conditions (Section 7.4) could readily apply to resolving alignment issues in cross-lingual frame semantic parsing.

7.5.1. Resolving Alignment Issues in Cross-lingual Frame Paraphrases

When encountering cross-lingual paraphrases in translated text, one straightforward strategy in aligning frame structures would include heuristics which check for the presence of a side condition and trigger an alignment across frames. For example, the presence of semantic roles (side condition 3 in Section 7.4 would be a simple heuristic to implement, and a lexicon-based sentiment analysis (Khoo and Johnkhan, 2018; Wilson, Wiebe, and Hoffmann, 2005) which recognizes predicates with positive or negative valence would largely be sufficient in detecting side condition 1. Shifts in the granularity of semantic roles (side condition 2) could be handled in the lexicon, where a system would check if the semantic role was defined as equivalent to multiple semantic roles in another frame. It would therefore benefit a cross-lingual NLP system to have side conditions for frame paraphrases to be specified within the frame lexicon of at least one, but preferably both, the languages of interest.

7.5.2. Cross-lingual Frame Paraphrases in Encoder-Decoders

One promising recent direction in cross-lingual semantic parsing involves the use of encoder-decoders for labeling semantic roles across languages (Cai and Lapata, 2020; Blloshmi et al., 2021). In this approach, a model is trained with data from a source language and applied directly to new languages using cross-lingual word embeddings or Universal POS tags (McDonald et al., 2013). In the process of translation, semantic role labels are also predicted for the target language (Daza and Frank, 2019b). However, to date these encoder-decoder approaches have only been attempted on PropBank-style role labels (Kingsbury and Palmer, 2003) which are more closely tied to syntactic dependencies and do not involve predicting general semantic categories such as frames.

One future direction could be to design an encoder-decoder to predict frames in a target language in a similar fashion. Certain side conditions, such as sentiment, could be designed as additional input layers to the encoder, and additional side conditions, such as a difference in granularity of semantic roles, could also be discovered in this approach. The final advantage of this setup is that the model would be predicting frame paraphrases across languages without requiring parallel, aligned data as input.

7.6. Summary

This chapter addresses the question of low frame comparability from a linguistic perspective, where we are interested in cases of mismatch between frame usage in text and their lexicographic relationships. Frame paraphrases occur when frames differ across paraphrased text, where the text seems parallel but in practice actually evokes non-parallelism in frame structures. They have been attested in cross-lingual data, and they lead to mismatches in frame semantic parsing across languages (Padó, 2007). However, there has yet to be a study to concretely identify difficult cases of frame paraphrases - that is, paraphrases that are more conceptual in nature and are not explained via existing FrameNet structures. This work presents the first attempt to link frames in the FrameNet hierarchy that evoke parapharases outside of the classical linguistic relations under which they are primarily examined. As frame-to-frame relations are studied almost exclusively in English, we began our study with looking at monolingual cases, targeting side conditions we observed in English frame paraphrases. However, we would speculate that side conditions can also be found in cross-lingual paraphrases, and mismatches in cross-lingual frame semantic parsing can be accounted for when there are explicit lexicographic devices that connect these frames conceptually. Alternatively, future directions could explore ways to computationally identify side conditions that could be incorporated into the FrameNet lexicon. In Part 7.5 we outlined how a computational system can cope with cross-lingual frame paraphrases and proposed using either side conditions to improve alignment of frame paraphrases, or encoder-decoder architectures to predict cross-lingual frame paraphrases without the need for an aligned, parallel corpus.

The work in this chapter presented a linguistic analysis of frame paraphrases sketched different technical approaches to address frame paraphrases based on observations from this analysis. It was also the first attempt to formally define, via frame-to-frame relations, concept-based paraphrases across frames that require real-world knowledge for their understanding. Concretely, we have identified cases of frame comparability where the usage of the frame has poor alignment across parallel sentences; however, instead of modifying the definition of the frame, certain side conditions actually explain this poor usage-based comparability. These are exceptions to the standard approach to coping with poor usage-based comparability, where often the approach involves an alteration to the definition of the frame. Instead, these side conditions explain how such usage-based mismatches might occur, and instead of more substantial shifts in the definition of the frame itself, these paraphrases warrant small additions (side conditions) to the frame lexicon.

Part V.

Conclusions

8. Conclusions

The beginning of this thesis introduced the concept of frame comparability, where frames can be compared across languages from a lexicographic aspect - that is, a frame's definition, or a usage-based aspect - that is, how a frame appears in corpora. Each part of this thesis touches upon both lexicographic and usage-based aspects from computational and linguistic perspectives.

We also introduced three main research questions, addressed in separate parts of this thesis: 1) how can we measure frame comparability, 2) how much is low frame comparability an issue for models of frame identification across languages, and how can we quantify comparability to improve models of cross-lingual frame identification, and 3) how can we enrich the frame lexicon to explain low usage-based frame comparability. Below, we will review the contributions we make to the above research questions, and distinguish specific contributions to the lexicographic and usage-based aspects of frames across languages.

8.1. Contributions

This thesis addresses frame comparability from computational and linguistic perspectives, and we make several key contributions to both of these perspectives. Below, we describe contributions from Parts II-IV of this thesis, where each part addressed one of the three main research questions above.

8.1.1. Embeddings and Frame Comparability

Part II described the first work on using embeddings of frames to directly compare a frame in one language with its counterpart in another language. Our work relied on existing frame-annotated corpora for English and German, where we built embeddings based on monolingual corpora and projected those embeddings to a shared space via a linear mapping (see Chapter 4). This work was novel in the sense that 1) we built computational representations of frames based on corpora originally written in the target languages (not translations,) which enabled us to study computational representations of frames without introducing issues relating to translationese (Salkie, 2002); 2) our approach used a linear mapping to project vectors into a shared, multilingual space, where we could directly compare the similarity of a frame's usage in English and German text, which was the first work to compare embeddings of frames across languages. Our results linked similarities in the shared German-English vector space to findings in the linguistics literature relating to frames across languages, and we found that, while the approach enabled us to directly compare frames across languages, the embeddings also reflected differences in annotation choices including differences in frequency and variety of the frame's LU annotations.

8.1.2. Comparison of Model Designs for Frame Identification

Part III addressed the impact of system design in the frame identification task. Chapter 5 introduced a system for frame identification where we compared four designs along 2 dimensions for the classification of frames. The first dimension was a comparison of prototype and exemplar models, and the second dimension was a comparison of performance over pre-trained and (pre-trained+) fine-tuned embeddings. Our results set a new state-of-the-art in the frame identification task, and we found that a prototype-based model for frame classification produced the best results over the existing available frame annotations. Additionally, we found that although fine-tuning (predictably) provides the best results, the pre-trained embeddings nonetheless perform surprisingly well.

8.1.3. Frame Selection for Multilingual Frame Identification

Following the success of our BERT-based frame identification system, we contributed the first multilingual frame identification system using multilingual BERT (described in Chapter 6). We tested our frame identification with a zero-shot, *S only*, baseline and found that certain language pairs worked better than others. Although results in the *S only* condition showed that relying on frame annotations from other languages alone is not sufficient for any production-ready system for frame identification in the target language, the model was able to generalize about certain frames for a target language by only seeing frame annotations from a different, source language.

Chapter 6 also introduced two experiments in selecting frames for multilingual frame identification. In these experiments, the objective was to predict the source language frames that would maximally improve the performance of a frame identification system for a target language. We based our selection criteria on clustering properties of the frame within (Experiment 3) and across (Experiment 2) languages. These experiments resulted in several contributions. First, we found that, across Experiments 2 and 3, selecting frames based on their clustering properties led to greater gains than a random selection of frames. Practically speaking, in Experiment 3, we show that this method is even applicable when no target language annotations are available; meaning one can base their frame selection on available, source language annotations alone and still see gains in their target language performance. Second, an important design choice in these experiments was to train the selection method on a single pair of source-target languages (EN-DE), which could then be generalized to other language pairs in our test. The result of this process validated that parameters learned in the frame selection model can, in fact, be applied as-is to other language pairs - confirming the generalizability of our approach. Finally, we compared the gains in the cross-lingual frame identification model when the frames selected had high lexicographic comparability or low lexicographic comparability, and we found that a majority of the language pairs showed greater gains when the frames had lower lexicographic comparability (Figure 6.7).

8.1.4. Side Conditions and Categorization of Frame Paraphrases

Part IV describes our linguistic analysis of frame paraphrases: monolingual or cross-lingual paraphrases that evoke different frames. Frame paraphrases arise when the comparability across parallel text - data in which we might expect to find high frame comparability - is actually low enough to cause a mismatch in aligning frame structures across languages. Our contributions of this work were a) a categorization of different frame pairs in the ***Using*** frame-to-frame relation, where each class is capable of evoking different types of frame paraphrases; and b) a description of the existing side conditions which explain how a frame in one utterance is translated as another frame in its paraphrase. We would suggest adding these side conditions to existing frame lexicons to enable computational systems to predict frame paraphrases in the future; alternatively, side conditions can be learned automatically given parallel frame annotations, which are emerging as part of the Multilingual FrameNet project [1].

8.2. Improvements to our Existing Work

Subsequent research emerged following our studies which could contribute to improvements in our design or analyses.

8.2.1. Cross-lingual Frame Embeddings

While Part III utilized the most recent advancements in large-scale language models for monolingual and cross-lingual models of frame identification, these models emerged after the publication of our study in Part II. Our approach to aligning frame embeddings into a shared, multilingual vector space required a linear map, which would not be necessary if we were to use existing multilingual embeddings that were trained in a joint, multilingual space. In fact, there is some evidence to suggest that assuming a linear map across two vector spaces yields poorer results in comparison to non-linear transformations - thus weakening the assumption of linearity across those spaces (Grave, Joulin, and Berthet, 2019). Additionally, there have been follow-up studies using large-scale, multilingual embeddings to evaluate frames across languages, although these studies evaluated cross-lingual frames at the level of translations of the lexical units (Baker and Lorenzi, 2020).

An improvement to our current design would consist of evaluating frame embeddings in a multilingual space by nearest neighbor comparison. Concretely, frame embeddings could be constructed from existing frame-annotated cor-

[1] https://www.globalframenet.org/

pora by taking the centroid of the lexical unit embeddings (similar to the frame prototypes described in Section 5.2.1) and one could then directly compare the embedding of a frame in one language to another by cosine similarity. This would be an improvement to our reported study, as a) no effort is required to map embeddings into a shared space, and b) since large-scale, multilingual embeddings are trained on a larger corpus, we could determine how much the (dis-)similarity of cross-lingual frames was due to factors related to the annotated data in our approach.

8.2.2. Frame Selection by Lexical Unit

Chapter 5 demonstrated the efficacy of frame selection using the clustering properties of frames. A clear extension of this work would be to select instances for target language training at the level of the lexical unit instead of the frame. In our approach, we combined annotations from different lexical units within the same frame and added them back to the training process. While this enabled us to make observations about frame performance in a cross-lingual frame identification system, it is possible that a model which selected more promising lexical units could achieve an even greater performance. This could be the case, as some lexical units within a frame are more likely to transfer across languages then others. To implement this, only moderate modifications would be necessary to our current design; the clustering metric, instead of measuring frame clusters, would measure clustering properties of lexical units. Benefits to this approach would be a greater sensitivity to translation issues in cross-lingual frame identification, where certain lexical units that map to many senses in the target language could be more likely to be selected for training than those with more direct, one-to-one mappings.

8.2.3. Side Conditions for Frame Paraphrases

Part IV introduced side conditions, which accounted for frame paraphrases across parallel sentences. We speculated that these side conditions would benefit models of frame alignment across translated texts, as they could account for certain alignment errors that are likely to occur in the process. Expanding our existing list of side conditions, and enriching the frame lexicon with these side conditions, would greatly improve our ability to account for cases of poor frame alignment in computational models.

It is also possible that future systems could detect additional side conditions from parallel frame annotated corpora. This work would involve identification of paraphrases that evoke different frames - either by taking translated, parallel sentences and predicting their frames, or taking sentences that have frame annotations in one language and translating them into another. A frame paraphrase corpus could then be composed of cases of frames that are and are not aligned across these parallel, translated sentences. A computational model trained on this data would implicitly learn the properties of translations that lead to frame mismatch, thus potentially resulting in the discovery of new side conditions.

8.3. Lexicographic and Usage-based Frame Comparability in Linguistic and Computational Perspectives

We have described in this thesis two dimensions along which a frame can be compared across languages: a lexicographic dimension and a usage-based dimension, and we describe throughout the thesis how these dimensions appear from linguistic and computational perspectives. The computational perspective provides quantitative comparison of frames over large corpora, while the linguistic perspective brings qualitative analyses of lexicographic and usage-based frames on a more case-by-case basis. Below, we summarize key findings that illustrate how nuanced the relationship between lexicographic and usage-based comparability is for frames across languages.

8.3.1. Computational Perspectives

Our computational models were built on word embeddings, which in our case are representations of a frame in context that were learned by a neural network. We would expect that many frames with similar definitions across languages would also correspond to having a high usage-based comparability, and therefore a connection could be drawn between the embeddings of a frame and the similarity in the definitions of the frame across languages. We find in our experiments that this assumption is only weakly supported; in Chapter 6, we showed that the frames that were selected for transfer to different languages were not consistently those with high lexicographic comparability. Mirroring this finding, in Chapter 4, we compare embeddings across English and German where a predicate (*ankündigen – to herald*) belongs in different frames in the English and German lexicons but actually appears similar in its usage, thereby indicating that the lexicographic divergence is not corresponding to its actual use in text. These findings further weaken a hypothesized relationship between the lexicographic definition of a frame and its usage in cross-lingual embedding space, and suggest that computational approaches to improve cross-lingual frame semantic parsing ought to focus on more embedding and usage-based metrics over incorporating lexicographic similarities to improve cross-lingual systems.

One potential reason for the lack of a consistent relationship between frame embeddings and lexicographic definitions of frames across languages is that our frame embeddings are based on annotated corpora. Because the corpora are different texts written in the native language (and not translations), they diverge in topics and content. Those divergences could also be influencing (dis-)similaritites in the frame embeddings. While this further motivates using usage-based metrics for improving cross-lingual computational models, it also motivates further linguistic inquiry into lexicographic and usage-based comparability for a more complete understanding of this relationship.

8.3.2. Linguistic Perspectives

In Chapter 7, we considered paraphrases across different frames, a phenomenon where two sentences that seem parallel actually evoke different frame structures. Frame paraphrases are cases of low usage-based frame comparability and low lexicographic comparability, which can be found in translations and monolingual, parallel corpora. Our linguistic analyses suggest that there is an implicit relationship between these frames that could be drawn explicitly in the lexicon (i.e. side conditions), thereby recovering the connection between the frame's usage-based and lexicographic comparability. Our linguistic analysis deepens our understanding of lexicographic and usage-based comparability by examining exactly the cases where that breakdown occurs, why they are occurring, and suggests what the reconciliation between these comparability dimensions could look like.

On the whole, we find that assessing frame comparability is a challenge as there is a good possibility of a disconnect between the usage of a frame in text and its definition in the lexicon - either due to idiosyncrasies in the corpora or semantic relationships that are missing in the lexical entry. We address ways to improve both of these in the next section, where we can explore new corpora for assessing usage-based frame comparability, and further refinement and development of a lexicon that incorporates frames from different languages as a place to explicitly capture lexicographic-based frame comparability.

8.4. Future Directions

Work in this thesis has described the differences that can be found when studying the lexicographic versus the usage-based comparability of frames across languages. There are many paths forward that will reconcile these mismatches and further support the application of frames to different languages. More broadly, these efforts will lead to a deeper understanding of the universal applicability of frames and their structures.

There is an ongoing annotation project for frames across languages currently underway across several research groups in the multilingual FrameNet community (Torrent et al., 2018). This work involves annotations of frames in different languages across the same text, where the text (originally in English) has been translated into the target languages of interest. To date, these annotations have not been publicly released for computational models, but linguistic analyses have emerged in preliminary studies that contribute to the landscape of how we understand frames across languages. Observations from initial studies in Brazilian Portuguese have reported optimistic results that as much as 80% of lexical units in Portuguese were found to fit into frames defined from the English Berkeley FrameNet (Torrent et al., 2018). Analyzing these multilingual frame annotations over parallel text will ideally shed more light on instances of high frame comparability in the usage of frames, and will also be the ideal corpus for further studies on frame paraphrases in a cross-lingual setting.

The most recent trend in NLP is experimentation with large-scale language models such as BERT. Cross-lingual frame semantic research stands to benefit from the development of these models and the ongoing research in tuning these models for cross-lingual systems, such as the work described in Section 6.1. As the Multilingual FrameNet project publishes their corpus of parallel, cross-lingual frame annotations, it would be worth investigating the extent to which these large-scale, multilingual language models could predict utterances that evoke different frames. While our studies in cross-lingual frame identification with BERT illustrate the efficacy of these models in learning frames for different languages, it is still an open question as to whether the current trend in model transfer (Daza and Frank, 2019b; Cai and Lapata, 2020; Blloshmi et al., 2021) successfully predicts frames in frames with lower lexicographic or usage-based comparability.

8.4.1. FrameNets Across Languages

We have described in this thesis frame comparability in two broad terms: 1) comparability of frames in terms of how they are defined in the lexicon, and 2) comparability of the frames in terms of their usage in corpora. Further annotation of parallel text might improve our understanding of usage-based frame comparability by potentially shedding light on the effect of translationese on frame comparability, thus strengthening the understanding of whether frame incomparability is due to shifts in typological differences across languages, or whether these shifts can be seen as differences in the choices in language use (specifically, how the same scene is conceived of differently across languages).

There is much work that would contribute to our understanding of frames across languages from the lexicographic perspective. Current FrameNets are released and maintained independently from one another (see Section 2.4.2), leaving much of the work of cross-lingual frame comparison to the linguist. Implementing multilingual structures directly into FrameNet would enable future computational research to incorporate lexicographic frame comparability into models of frames. We demonstrated in Chapter 6 that the lexicographic comparability of a frame is one indicator of its potential benefit to a target language frame identification system, so from the computational perspective there is significant motivation to improve the accessibility of any established research relating to a frame's lexicographic comparability.

Ideally, a fully multilingual FrameNet would include several additions to the existing FrameNet structures (see Figure 8.1). First, a multilingual FrameNet would enable cross-lingual descriptions of a single frame, and in the case of one-to-many alignments, would explicitly outline how a frame in one language maps to one or many counterparts in another. This would involve translations of the lexical units across languages, as well as mappings of semantic roles across languages. Second, frame paraphrases (discussed in Chapter 7) and their side conditions would be represented via additional relations and constraints in the lexicon (see the *Frame Paraphrase* relation in Figure 8.1), enabling research in computational frame semantics to better cope with paraphrasing in translations. Instituting these changes would make the FrameNet resource more compatible with cross-lingual frame comparisons and improve computational models of frames across languages in the following ways:

- Lexicographic comparability of the frame would be transparent, enabling future computational systems to adapt frame semantic parsing more readily to different languages
- Usage-based comparability over different corpora could be linked to frame entries in the lexicon. This would enable better generalizations of

the relationship between usage-based and lexicographic-based comparability of frames across languages

- There would be a single FrameNet resource that new languages could contribute to, in contrast to the current state of operations where FrameNets are developed independently. Frames could be language-specific in this resource, therefore having only a single language in the *cross-lingual* property, and could also be shared, as in Figure 8.1

- As more languages contribute to a cross-lingual FrameNet resource, further observations about frames across language families could be made. Much of the observations about frame comparability have been made over single language pairs; having a resource with multiple languages and their evocation of certain frames would illustrate clearly whether related families show similar patterns in how they conceptualize different frames

- A unified, cross-lingual FrameNet would facilitate more informed and more inclusive understanding of the universality of frames, where researchers who have not considered frames in their own language would ideally be able to participate in the discussion and contribute to the research more quickly and effectively

Judgment_Communication

A Communicator communicates a judgment of an Evaluee to an Addressee.
Cross-lingual: English (EN), German (DE), French(FR)

FEs

Communicator
Semantic Type: Sentient
Cross-ingual: EN, DE, FR

Evaluee
Cross-ingual: EN, DE, FR

Expressor
Cross-ingual: EN, DE, FR

Frame-frame Relations:
Is Used by:
Labeling
cross-lingual:EN-DE
Inherits from:
Judgment_communication_no_reason
cross-lingual: FR
Frame Paraphrase:
Labeling
side condition: sentiment
cross-lingual: EN-DE-FR

Lexical Units

EN	DE	FR
reject.v	ablehnen.v	
accuse.v	vorwerfen.v	accuser.v
criticism.n	Kritiker.n	critiquer.v
praise.v	preisen.v	louer.v

Figure 8.1.: Cross-lingual frame lexicon entry. Structures include a definition of the frame as it applies to specific language(s), semantic roles (FEs) that apply cross-lingually, frame relations, and lexical units for each language that can evoke the frame. Modifications can be expressed via a *cross-lingual* key, where each of the structures (FEs, frame relations, lexical units) defines the language it applies to. Language-specific modifications can readily be represented in this framework; for instance, the ***Inherits from*** relation only applies to the French (FR) language.

Part VI.

Appendices

Appendix A

Frame Pairs in the Using Frame-to-Frame Relation and their Classification

Class 1	
Frame 1	**Frame 2**
Word_relations	Linguistic_meaning
Wearing	Observable_body_parts
Wearing	Clothing
Wearing	Accoutrements
Undressing	Clothing
Undressing	Accoutrements
Terrorism	Intentionally_act
Temporal_pattern	Event
Tasting	Food
Studying	Education_teaching
Store	Storing
Sound_movement	Sounds

Social_event	Eventive_affecting
Simple_naming	Simple_name
Sign	Evidence
Severity_of_offense	Offenses
Sent_items	Sending
Roadways	Motion
Remainder	Change_position_on_a_scale
Relational_natural_features	Locative_relation
Referring_by_name	Being_named
Recovery	Medical_conditions
Reason	Intentionally_act
Protecting	Run_risk
Prohibiting	Law
Progress	Process
Prison	Inhibit_movement
Point_of_dispute	Discussion
Point_of_dispute	Be_in_agreement_on_assessment
Performers	Performers_and_roles
Performers_and_roles	Performing_arts
People_by_residence	Residence
People_by_religion	Religious_belief
People_by_morality	Morality_evaluation
People_by_age	Age
Path_traveled	Motion

Offenses	Compliance
Offenses	Committing_crime
Occupy_rank	Rank
Medical_specialties	Cure
Medical_professionals	Cure
Medical_instruments	Cure
Manipulation	Observable_body_parts
Making_faces	Facial_expression
Locative_relation	Existence
Locale	Locative_relation
Locale_by_ownership	Possession
Locale_by_event	Event
Linguistic_meaning	Simple_name
Intoxication	Intoxicants
Ingredients	Creating
Individual_history	Importance
Inclusion	Part_whole
Going_back_on_a_commitment	Commitment
Frequency	Event
Food	Ingestion
Fining	Commerce_collect
Fastener	Closure
Facial_expression	Body_movement
Expensiveness	Commerce_scenario

Expected_location_of_person	Custom
Exemplar	Judgment
Exclude_member	Membership
Exchange_currency	Commerce_scenario
Elusive_goal	Purpose
Earnings_and_losses	Commerce_pay
Dressing	Clothing
Dressing	Accoutrements
Documents	Grant_permission
Deserving	Reason
Defend	Attack
Create_representation	Physical_artworks
Create_physical_artwork	Physical_artworks
Correctness	Information
Containers	Containing
Connectors	Attaching
Completeness	Part_whole
Communication	Topic
Communication	Information
Clothing	Closure
Chemical-sense_description	Appearance
Change_event_time	Calendric_unit
Change_direction	Direction
Categorization	Instance

Body_movement	Observable_body_parts
Being_located	Locative_relation
Being_in_operation	Gizmo
Behind_the_scenes	Performing_arts
Bearing_arms	Weapon
Awareness	Information
Attributed_information	Statement
Attention_getting	Referring_by_name
Artifact	Using
Arson	Setting_fire
Arson	Intentionally_affect

Class 2

Frame 1	**Frame 2**
Wagering	Run_risk
Time_period_of_action	Possibilities
Terms_of_agreement	Documents
Temporal_subregion	Part_orientational
Temporal_pattern	Process
Purpose	Means
Public_services	Institutions
Project	Purpose
Political_locales	Leadership
People_by_origin	Origin
People_by_jurisdiction	Political_locales
Ordinal_numbers	Cardinal_numbers

Frame 1	Frame 2
Opportunity	Possibilities
Network	Set_of_interrelated_entities
Namesake	Being_named
Medical_conditions	Observable_body_parts
Make_agreement_on_action	Commitment
Luck	Destiny
Law	Text
Isolated_places	Locale
Institutions	Infrastructure
Indigenous_origin	Foreign_or_domestic_country
Identicality	Instance
Hospitality	Guest_and_host
Hair_configuration	Observable_body_parts
Gizmo	Artifact
Fighting_activity	Hostile_encounter
Fields	People
Economy	Political_locales
Diversity	Type
Craft	Fields
Craft	Custom
Clothing_parts	Clothing
Businesses	Commerce_scenario
Buildings	Locale_by_use
Building_subparts	Locale

Frame 1	**Frame 2**
Body_mark	Observable_body_parts
Body_description_part	Observable_body_parts
Body_decoration	Observable_body_parts
Be_in_agreement_on_action	Make_agreement_on_action
Alliance	Organization
Alliance	Competition

Class 3a

Frame 1	**Frame 2**
Withdraw_from_participation	Participation
Waiting	Change_event_time
Undressing	Removing
Traversing	Path_shape
Touring	Visiting
Tolerating	Experiencer_focus
Text_creation	Communication
Tasting	Ingestion
Talking_into	Suasion
Sufficiency	Capability
Successfully_communicate_ message	Communication
Suasion	Eventive_cognizer_affecting
Statement	Communication
Spelling_and_pronouncing	Text_creation

Frame 1	**Frame 2**
Speak_on_topic	Communication
Shoot_projectiles	Use_firearm
Scrutiny	Perception_active
Robbery	Theft
Remembering_to_do	Remembering_information
Reliance	Contingency
Reliance_on_expectation	Certainty
Reliance_on_expectation	Awareness
Regard	Judgment
Reforming_a_system	Cause_change
Reasoning	Communication
Ratification	Grant_permission
Quarreling	Be_in_agreement_on_assessment
Prevent_from_having	Possession
Prevarication	Communication
Predicting	Expectation
Partiality	Taking_sides
Obscurity	Fame
Needing	Have_as_requirement
Motion_noise	Motion
Mass_motion	Abounding_with
Launch_process	Cause_to_start
Labeling	Judgment_communication
Justifying	Communication

Frame 1	Frame 2
Jury_deliberation	Discussion
Judgment_communication	Statement
Installing	Placing
Importing	Import_export
Importing	Commerce_buy
Imitating	Similarity
Holding_off_on	Change_event_time
Hit_target	Hit_or_miss
Hit_target	Cause_impact
Historic_event	Importance
Having_or_lacking_access	Arriving
Have_associated	Existence
Have_as_translation_equivalent	Translating
Getting_up	Change_posture
Fullness	Containing
Front_for	Posing_as
First_rank	Prominence
Finish_competition	Success_or_failure
Fairness_evaluation	Social_interaction_evaluation
Extreme_value	Measurable_attributes
Exporting	Import_export
Explaining_the_facts	Evidence
Exchange_currency	Exchange
Eventive_cognizer_affecting	Subjective_influence

Frame 1	Frame 2
Eventive_cognizer_affecting	Influence_of_event_on_cognizer[2]
Eclipse	Perception_experience
Earnings_and_losses	Commerce_collect
Dressing	Placing
Dominate_situation	First_rank
Disembarking	Departing
Discussion	Communication
Desiring	Experiencer_focus
Desirable_event	Required_event
Deny_permission	Communication
Degree_of_processing	Processing_materials
Criminal_investigation	Seeking
Create_representation	Create_physical_artwork
Court_examination	Questioning
Cooking_creation	Apply_heat
Conduct	Intentionally_act
Communication_noise	Make_noise
Communicate_categorization	Categorization
Committing_crime	Compliance
Colonization	Residence
Cause_to_fragment	Destroying
Carry_goods	Storing
Bringing	Motion
Bringing	Cause_motion

Frame 1	**Frame 2**
Beyond_compare	Surpassing
Being_in_category	Categorization
Being_detached	Being_attached
Beat_opponent	Win_prize
Attempt_suasion	Influence_of_event_on_cognizer
Arranging	Placing
Adopt_selection	Choosing
Adjusting	Cause_change
Adducing	Statement

Class 3b

Frame 1	**Frame 2**
Win_prize	Finish_competition
Willingness	Choosing
Wealthiness	Possession
Waver_between_options	Choosing
Want_suspect	Appearance
Waiting	Intentionally_act
Volubility	Communication
Verification	Correctness
Verdict	Communication
Usefulness	Using
Usefulness	Capability
Undergoing	Eventive_affecting

Frame 1	Frame 2
Unattributed_information	Statement
Typicality	Similarity
Turning_out	Coming_to_believe
Toxic_substance	Cause_harm
Time_vector	Direction
Thriving	Desirability
Taking_time	Duration_attribute
Taking_time	Being_necessary
Taking_sides	Opinion
Taking_sides	Desirable_event
Suspicion	Criminal_investigation
Surrendering	Want_suspect
Surrendering	Arrest
Supporting	Cause_change_of_strength
Subjective_influence	Intentionally_act
Suasion	Communication
Suasion	Attempt_suasion
Strictness	Compliance
Storing	Placing
Stage_of_progress	Progress
Sole_instance	Instance
Sidereal_appearance	Motion_directional
Shopping	Commerce_buy
Shoot_projectiles	Cause_motion

Frame 1	Frame 2
Sentencing	Communication
Sending	Bringing
Secrecy_status	Awareness
Scrutiny	Becoming_aware
Ruling_legally[3]	Communication
Rite	Intentionally_act
Resurrection	Dead_or_alive
Respond_to_proposal	Communication_response
Research	Cogitation
Required_event	Being_necessary
Request	Communication
Representing	Awareness
Reporting	Communication
Repel	Being_strong
Renunciation	Statement
Removing	Motion
Remembering_to_do	Purpose
Remembering_to_do	Intentionally_act
Remembering_information	Awareness
Remembering_experience	Cogitation
Religious_belief	Awareness
Redirecting	Motion
Range	Capability
Questioning	Communication

Frame 1	**Frame 2**
Purpose	Desiring
Progress[4]	Undergo_change
Probability	Position_on_a_scale
Preventing	Event
Praiseworthiness	Judgment
Placing	Motion
Place_weight_on[5]	Importance
Piracy	Operate_vehicle
Perception_body	Perception_experience
Perception_active	Attention
Path_shape	Locative_relation
Participation	Event
Operational_testing	Operating_a_system
Operate_vehicle	Motion
Openness	Traversing
Offering	Giving
Notification_of_charges	Communication
Needing	Desiring
Name_conferral	Communication
Misdeed	Morality_evaluation
Memory	Eventive_affecting
Meet_with	Discussion
Meet_specifications	Sufficiency
Make_cognitive_connection	Cognitive_connection

Frame 1	Frame 2
Luck	Likelihood
Losing_it	Mental_property
Locating	Seeking
Lively_place	Activity_ongoing
Light_movement	Motion
Legality	Morality_evaluation
Legality	Law
Legality	Compliance
Left_to_do	Purpose
Judgment_communication	Judgment
Institutionalization	Cure
Ingestion	Cause_motion
Impression	Awareness
Import_export	Intentionally_affect
Holding_off_on	Forgoing
Hit_target	Shoot_projectiles
Hindering	Event
Hear[6]	Communication
Health_response	Response
Guilt_or_innocence	Misdeed
Grooming	Desirability
Grant_permission[7]	Communication
Giving_in	Taking_sides
Gathering_up	Cause_motion

Frame 1	**Frame 2**
Front_for	Prevarication
Friction	Impact
Forging	Artificiality
Fluidic_motion	Motion
Feigning	Conduct
Fame	Awareness
Expressing_publicly	Communication
Exporting	Commerce_sell
Experience_bodily_harm	Intentionally_act
Expensiveness	Abounding_with
Excreting	Motion
Excreting	Cause_motion
Examination	Awareness
Evoking	Memory
Evading	Motion
Estimated_value	Estimating
Entering_of_plea	Communication
Enforcing	Being_in_effect
Encoding	Communication
Emotions_of_mental_activity	Attention
Emanating	Motion
Dough_rising	Expansion
Distinctiveness	Similarity
Disembarking	Ride_vehicle

Frame 1	Frame 2
Difficulty	Hindering
Detaining	Inhibit_movement
Desirability	Experiencer_focus
Deserving	Response
Deserving	Required_event
Departing	Motion
Delivery	Sending
Delimitation_of_diversity	Position_on_a_scale
Delimitation_of_diversity	Diversity
Degree	Importance
Deciding	Intentionally_act
Custom	Frequency
Cotheme	Motion
Correctness	Similarity
Convey_importance	Communication
Continued_state_of_affairs	State_continue
Contacting	Communication
Concessive	Communication
Competition	Intentionally_act
Commutation	Clemency
Communication_means	Communication
Committing_crime	Legality
Commitment	Communication
Coming_up_with	Eventive_affecting

Frame 1	Frame 2
Claim_ownership	Communication
Chatting	Statement
Change_of_quantity_of_possession	Possession
Change_of_quantity_of_possession	Change_position_on_a_scale
Change_direction	Motion
Certainty	Awareness
Cause_to_move_in_place	Manipulation
Cause_to_move_in_place	Activity_ongoing
Cause_harm	Experience_bodily_harm
Catastrophe	Eventive_affecting
Carry_goods	Commerce_sell
Capability	Likelihood
Candidness	Communication
Bungling	Intentionally_affect
Bungling	Intentionally_act
Breathing	Fluidic_motion
Bond_maturation	Repayment
Body_movement	Motion
Board_vehicle	Ride_vehicle
Being_rotted	Being_wet
Being_operational	Render_nonfunctional
Being_active	Activity_ongoing
Be_translation_equivalent	Translating

Frame 1	**Frame 2**
Be_in_agreement_on_assessment	Opinion
Bail_decision	Communication
Attempt_suasion	Communication
Atonement	Forgiveness
Assistance	Intentionally_act
Assemble	Intentionally_act
Arrest	Inhibit_movement
Aging	Age
Adding_up	Amounting_to
Activity_stop	Eventive_affecting
Activity_start	Eventive_affecting
Activity_prepare	Eventive_affecting
Activity_pause	Eventive_affecting
Achieving_first	First_experience
Accuracy	Success_or_failure
Accuracy	Measurable_attributes
Accomplishment	Intentionally_act
Abusing	Cause_harm

Class 4

Frame 1	**Frame 2**
Volubility	Social_behavior_evaluation
Visitor_and_host	Guest_and_host
Use_vehicle	Ride_vehicle
Translating	Mental_activity

Frame 1	**Frame 2**
Subordinates_and_superiors	Relation_between_individuals
State_continue	State
Sounds	Perception
Sociability	Social_behavior_evaluation
Shooting_scenario	Bearing_arms
Shapes	Bounded_entity
Separating	Transitive_action
Searching_scenario	Attention
Reliance	Needing
Releasing_from_custody	Detaining
Relation_between_individuals	People
Purpose	Mental_activity
Product_delivery	Commerce_goods-transfer
Predicament	Emotions
Personal_relationship	Relation_between_individuals
People_by_residence	Relation_between_individuals
Make_noise	Perception
Lodging_scenario	Guest_and_host
Location_of_light	Perception
Locale	Bounded_entity
Kinship	Relation_between_individuals
Judgment	Emotions
Into	Source_path_goal
Into	Containment

Frame 1	**Frame 2**
In	Containment
Hit_target	Shooting_scenario
Goal	Source_path_goal
Forgiveness	Emotions
First_rank	Gradable_attributes
Fields	Employment_scenario
Feeling	Emotions
Extreme_point	Gradable_attributes
Expertise	Resolve_attempt_scenario
Experiencer_obj	Emotions
Exchange	Transfer_scenario
Estimating	Mental_activity
Emotion_heat	Emotions
Emotion_directed	Emotions
Emotion_active	Emotions
Documents	Obligation_scenario
Differentiation	Mental_activity
Desiring	Emotions
Cycle_of_existence_scenario	Entity
Contrition	Emotions
Containment_relation_IS	Containment
Containers	Bounded_entity
Conduct	Social_behavior_evaluation
Compliance	Obligation_scenario

Frame 1	**Frame 2**
Coming_to_believe	Mental_activity
Cogitation	Mental_activity
Categorization	Mental_activity
Bounded_entity	Boundary
Bounded_entity	Being_located
Being_in_effect	Obligation_scenario
Attempt_distant_interaction_ scenario	Manipulation
Appeal	Criminal_process
Activity_abandoned_state	Process_stopped_state

References

Abelson, Robert P (1976). "Script processing in attitude formation and decision making." In:

Akata, Zeynep et al. (2015). "Evaluation of output embeddings for fine-grained image classification". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2927–2936.

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

Akbik, Alan et al. (2015). "Generating high quality proposition banks for multilingual semantic role labeling". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 397–407.

Alec, Radford et al. (2018). *Improving language understanding with unsupervised learning*. Tech. rep. Tech. Rep., Technical report, OpenAI.

Anne, Abeillé and Barrier Nicolas (2004). "Enriching a French treebank." In: *LREC*.

Arora, Sanjeev, Yingyu Liang, and Tengyu Ma (2016). "A simple but tough-to-beat baseline for sentence embeddings". In:

Artetxe, Mikel, Gorka Labaka, and Eneko Agirre (2017). “Learning bilingual word embeddings with (almost) no bilingual data”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 451–462.

Atzler, Judith Kerstin (2011). “Twist in the list: Frame semantics as vocabulary teaching and learning tool”. PhD thesis.

Bach, Emmon (1981). “On Time, Tense, and Aspect: An Essay in English Metaphysics”. In: *Radical Pragmatics*. Ed. by P. Cole. New York: Academic Press, pp. 63–81.

Baker, Collin, Charles Fillmore, and John Lowe (1998). “The berkeley framenet project”. In: *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pp. 86–90.

Baker, Collin and Arthur Lorenzi (May 2020). “Exploring Crosslinguistic Frame Alignment”. English. In: *Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet*. Marseille, France: European Language Resources Association, pp. 77–84. ISBN: 979-10-95546-58-0. URL: `https://www.aclweb.org/anthology/2020.framenet-1.11`.

Baker, Collin F (2014). “FrameNet: A knowledge base for natural language processing”. In: *Proceedings of Frame Semantics in NLP: A workshop in honor of Chuck Fillmore (1929-2014)*, pp. 1–5.

Baker, Collin F et al. (2018). “Frame semantics across languages: Towards a multilingual framenet”. In: *Proceedings of the 27th International Conference on Computational Linguistics: Tutorial Abstracts*, pp. 9–12.

Bertoldi, Anderson and Rove Chishman (2012). “Frame Semantics and legal corpora annotation”. In: *Linguistic Issues in Language Technology* 7.

Biehl, Michael, Barbara Hammer, and Thomas Villmann (2016). “Prototype-based models in machine learning”. In: *Wiley Interdisciplinary Reviews: Cognitive Science* 7.2, pp. 92–111.

Blei, David, Andrew Ng, and Michael Jordan (2003). “Latent dirichlet allocation”. In: *Journal of machine Learning research* 3.Jan, pp. 993–1022.

Blloshmi, Rexhina et al. (2021). “Generating Senses and RoLes: An end-to-end model for dependency-and span-based Semantic Role Labeling”. In: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 3786–3793.

Boas, Hans C (2001). “Frame Semantics as a framework for describing polysemy and syntactic structures of English and German motion verbs in contrastive computational lexicography”. In: *Proceedings of corpus linguistics*. Vol. 113. University Centre for computer corpus research on language Lancaster, UK, pp. 64–73.

— (2005a). *From theory to practice: Frame Semantics and the design of FrameNet*. na.

— (2005b). “Semantic frames as interlingual representations for multilingual lexical databases”. In: *International Journal of Lexicography* 18.4, pp. 445–478.

— (2009). *Multilingual FrameNets in computational lexicography – Methods and applications*. De Gruyter.

— (2013). “Frame Semantics and translation”. In: *Iraide Ibarretxe-Antuñano/Ana Rojo (Hg.), Cognitive Linguistics and Translation, Berlin/New York*, pp. 125–158.

— (2020a). “A roadmap towards determining the universal status of semantic frames”. In: *New Approaches to Contrastive Linguistics*, p. 21.

Boas, Hans C (2020b). “A roadmap towards determining the universal status of semantic frames”. In: *New Approaches to Contrastive Linguistics*. De Gruyter Mouton, pp. 21–52.

Boas, Hans C and Ryan Dux (2013). “5) Semantic Frames for Foreign Language Education: Towards a German frame-based online dictionary”. In: *Veredas-Revista de Estudos Linguísticos* 17.1-.

Boas, Hans C, Ryan Dux, and Alexander Ziem (2016). “Frames and constructions in an online learner’s dictionary of German”. In: *Applied construction grammar* 32, p. 303.

Boas, Hans Christian (2002). “Bilingual FrameNet Dictionaries for Machine Translation.” In: *LREC*. Citeseer.

Bojanowski, Piotr et al. (2017). “Enriching word vectors with subword information”. In: *Transactions of the Association for Computational Linguistics* 5, pp. 135–146.

Borin, Lars et al. (2010). “The past meets the present in Swedish FrameNet++”. In: *14th EURALEX international congress*, pp. 269–281.

Botschen, Teresa, Hatem Mousselly-Sergieh, and Iryna Gurevych (2017). “Prediction of frame-to-frame relations in the FrameNet hierarchy with frame embeddings”. In: *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pp. 146–156.

Botschen, Teresa et al. (2018). “Multimodal frame identification with multilingual evaluation”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1481–1491.

Brants, Sabine et al. (2002). “The TIGER treebank”. In: *Proceedings of the workshop on treebanks and linguistic theories*. Vol. 168.

Budanitsky, Alexander and Graeme Hirst (2006). "Evaluating wordnet-based measures of lexical semantic relatedness". In: *Computational linguistics* 32.1, pp. 13–47.

Burchardt, Aljoscha et al. (2006). "The SALSA corpus: a German corpus resource for lexical semantics". In: *LREC*, pp. 969–974.

— (2009). "FrameNet for the semantic analysis of German: Annotation, representation and automation". In: *Multilingual FrameNets in Computational Lexicography: methods and applications* 200, pp. 209–244.

Cai, Rui and Mirella Lapata (2020). "Alignment-free Cross-lingual Semantic Role Labeling". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3883–3894.

Chang, Angel et al. (May 2016). "A comparison of Named-Entity Disambiguation and Word Sense Disambiguation". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), pp. 860–867. URL: `https://www.aclweb.org/anthology/L16-1139`.

Chang, Ming-Wei, Lev Ratinov, and Dan Roth (2007). "Guiding semi-supervision with constraint-driven learning". In: *Proceedings of the 45th annual meeting of the association of computational linguistics*, pp. 280–287.

Chen, Yun-Nung, William Yang Wang, and Alexander I Rudnicky (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*. IEEE, pp. 120–125.

Cho, Kyunghyun et al. (2014). "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734.

Chomsky, Noam (1955). "Logical syntax and semantics: Their linguistic relevance". In: *Language* 31.1, pp. 36–45.

— (1957). "Syntactic structures". In: *The Hague: Mouton.*

— (1964). *Current Issues in Linguistic Theory.* The Hauge: Mouton.

Chow, Ian C and Jonathan J Webster (2007). "Integration of linguistic resources for verb classification: FrameNet Frame, WordNet verb and suggested upper merged ontology". In: *International Conference on Intelligent Text Processing and Computational Linguistics.* Springer, pp. 1–11.

Conneau, Alexis and Guillaume Lample (2019). "Cross-lingual Language Model Pretraining". In: *Advances in Neural Information Processing Systems.* Vol. 32. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2019/file/c04c19c2c2474dbf5f7ac4372c5b9af1-Paper.pdf.

Conneau, Alexis et al. (2017). "Word Translation Without Parallel Data". In: *arXiv preprint arXiv:1710.04087.*

Costa, Diniz da et al. (2018). "Framenet-based modeling of the domains of tourism and sports for the development of a personal travel assistant application". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pp. 6–12.

Croft, William and D Alan Cruse (2004). *Cognitive linguistics.* Cambridge University Press.

Curran, James R. and Marc Moens (July 2002). "Improvements in Automatic Thesaurus Extraction". In: *Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition.* Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 59–66. DOI: 10.3115/1118627.1118635. URL: https://www.aclweb.org/anthology/W02-0908.

Das, Dipanjan et al. (2010a). “Probabilistic frame-semantic parsing”. In: *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pp. 948–956.

— (2010b). “SEMAFOR 1.0: A probabilistic frame-semantic parser”. In: *Language Technologies Institute, School of Computer Science, Carnegie Mellon University.*

Das, Dipanjan et al. (Mar. 2014). “Frame-Semantic Parsing”. In: *Computational Linguistics* 40.1, pp. 9–56. DOI: 10.1162/COLI_a_00163. URL: https://www.aclweb.org/anthology/J14-1002.

Daza, Angel and Anette Frank (2019a). “Translate and Label! An Encoder-Decoder Approach for Cross-lingual 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)*, pp. 603–615.

— (Nov. 2019b). “Translate and Label! An Encoder-Decoder Approach for Cross-lingual 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).* Hong Kong, China: Association for Computational Linguistics, pp. 603–615. DOI: 10.18653/v1/D19-1056. URL: https://www.aclweb.org/anthology/D19-1056.

— (Nov. 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).* Online: Association for Computational Linguistics, pp. 3904–3914. DOI: 10.18653/v1/2020.emnlp-main.321. URL: https://www.aclweb.org/anthology/2020.emnlp-main.321.

Deerwester, Scott et al. (1990). "Indexing by latent semantic analysis". In: *Journal of the American society for information science* 41.6, pp. 391–407.

Devlin, Jacob et al. (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://www.aclweb.org/anthology/N19-1423.

Djemaa, Marianne et al. (May 2016). "Corpus Annotation within the French FrameNet: a Domain-by-domain Methodology". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), pp. 3794–3801. URL: https://www.aclweb.org/anthology/L16-1601.

Du, Jingcheng et al. (2019). "Gene2vec: distributed representation of genes based on co-expression". In: *BMC genomics* 20.1, p. 82.

Ellsworth, Michael et al. (2006). "Frame-semantic analysis of motion scenarios in English, German, Spanish, and Japanese". In: *Fourth International Conference on Construction Grammar (ICCG4)*.

Erk, Katrin (2005). "Frame assignment as word sense disambiguation". In: *Proceedings of IWCS*. Vol. 6.

Erk, Katrin and Sebastian Pado (2006). "Shalmaneser-A Toolchain For Shallow Semantic Parsing." In: *LREC*. Citeseer, pp. 527–532.

Erk, Katrin and Sebastian Padó (2009). "Paraphrase assessment in structured vector space: Exploring parameters and datasets". In: *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pp. 57–65.

Erk, Katrin and Sebastian Pado (2010). "Exemplar-based models for word meaning in context". In: *Proceedings of the acl 2010 conference short papers*, pp. 92–97.

Fei, Hao, Meishan Zhang, and Donghong Ji (July 2020). "Cross-Lingual Semantic Role Labeling with High-Quality Translated Training Corpus". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 7014–7026. URL: https://www.aclweb.org/anthology/2020.acl-main.627.

Fellbaum, Christiane, ed. (1998). *WordNet An Electronic Lexical Database*. Cambridge, MA ; London: The MIT Press. ISBN: 978-0-262-06197-1. URL: http://mitpress.mit.edu/catalog/item/default.asp?ttype=2&tid=8106.

Ferrández, Oscar et al. (2010). "Aligning FrameNet and WordNet based on Semantic Neighborhoods." In: *LREC*. Vol. 10, pp. 310–314.

Fillmore, Charles and Collin F. Baker (2001). "Frame semantics for text understanding". In:

Fillmore, Charles J (1968). "The case for case". In: *Universals in Linguistic Theory*, pp. 1–88.

— (1982). "Frame Semantics". In: *Linguistics in the morning calm*, pp. 111–137.

— (1985). "Frames and the semantics of understanding". In: *Quaderni di semantica* 6.2, pp. 222–254.

Fillmore, Charles J (1992). "Corpus linguistics" or "Computer-aided armchair linguistics"". In: *Directions in corpus linguistics. Proceedings of Nobel Symposium.* Vol. 82, pp. 35–60.

Fillmore, Charles J et al. (1976). "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.* Vol. 280. 1. New York, pp. 20–32.

Frege, Gottlob (1884). "The foundations of arithmetic. Die Grundlagen der Arithmetik". In: *JL Austin (trans.)(Oxford: Blackwell and Mott, 1950 [1884]).*

Gaddy, David M et al. (2016). "Ten pairs to tag-multilingual POS tagging via coarse mapping between embeddings". In: Association for Computational Linguistics.

Gellerstam, Martin (1986). "Translationese in Swedish novels translated from English". In: *Translation studies in Scandinavia* 1, pp. 88–95.

Gerber, Matthew and Joyce Y Chai (2012). "Semantic role labeling of implicit arguments for nominal predicates". In: *Computational Linguistics* 38.4, pp. 755–798.

Gildea, Daniel and Daniel Jurafsky (2002). "Automatic labeling of semantic roles". In: *Computational linguistics* 28.3, pp. 245–288.

Goldstone, Robert L (1995). "Effects of categorization on color perception". In: *Psychological Science* 6.5, pp. 298–304.

Goldstone, Robert L, Yvonne Lippa, and Richard M Shiffrin (2001). "Altering object representations through category learning". In: *Cognition* 78.1, pp. 27–43.

Grave, Edouard, Armand Joulin, and Quentin Berthet (2019). "Unsupervised alignment of embeddings with wasserstein procrustes". In: *The 22nd In-*

ternational Conference on Artificial Intelligence and Statistics. PMLR, pp. 1880–1890.

Gries, Stefan Th and Nick C Ellis (2015). “Statistical measures for usage-based linguistics”. In: *Language Learning* 65.S1, pp. 228–255.

Hampton, James A (1995). “Testing the prototype theory of concepts”. In: *Journal of Memory and Language* 34.5, pp. 686–708.

Hanks, Patrick (2009). “The linguistic double helix: Norms and exploitations”. In: *After half a century of Slavonic natural language processing*, pp. 63–80.

Hanks, Patrick and K Allan (2013). “Lexicography from earliest times to the present”. In: *The Oxford Handbook of the History of Linguistics*. Oxford University Press Oxford, pp. 503–536.

Harris, Randy Allen (1995). *The linguistics wars*. Oxford University Press on Demand.

Harris, Zellig (1954). “Distributional structure”. In: *Word* 10.2-3, pp. 146–162.

Hartmann, Silvana et al. (2017). “Out-of-domain FrameNet semantic role labeling”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 471–482.

Hasegawa, Yoko et al. (2011). “FrameNet as a resource for paraphrase research”. In: *Constructions and Frames* 3.1, pp. 104–127.

Hermann, Karl Moritz and Phil Blunsom (June 2014). “Multilingual Models for Compositional Distributed Semantics”. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland: Association for Computational Linguistics, pp. 58–68. DOI: 10.3115/v1/P14-1006. URL: https://www.aclweb.org/anthology/P14-1006.

Hermann, Karl Moritz et al. (2014). “Semantic Frame Identification with Distributed Word Representations”. In: *Proceedings of the 52th Annual Meeting of the Association for Computational Linguistics.*

Hintzman, Douglas L (1986). “”Schema abstraction” in a multiple-trace memory model.” In: *Psychological review* 93.4, p. 411.

Huang, Eric H et al. (2012). “Improving word representations via global context and multiple word prototypes”. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 873–882.

Huang, Haoyang et al. (2019). “Unicoder: A Universal Language Encoder by Pre-training with Multiple Cross-lingual Tasks”. 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)*, pp. 2485–2494.

Hwang, Jena D et al. (2010). “Propbank annotation of multilingual light verb constructions”. In: *Proceedings of the Fourth Linguistic Annotation Workshop*, pp. 82–90.

Iacobacci, Ignacio, Mohammad Taher Pilehvar, and Roberto Navigli (July 2015). “SensEmbed: Learning Sense Embeddings for Word and Relational Similarity”. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China: Association for Computational Linguistics, pp. 95–105. DOI: 10.3115/v1/P15-1010. URL: https://www.aclweb.org/anthology/P15-1010.

Ikegami, Yoshihiko (1991). "DO-language'and 'BECOME-language': Two contrasting types of linguistic representation". In: *The empire of signs: Semiotic essays on Japanese culture* 8, pp. 285–327.

Jäkel, Frank, Bernhard Schölkopf, and Felix A Wichmann (2008). "Generalization and similarity in exemplar models of categorization: Insights from machine learning". In: *Psychonomic Bulletin & Review* 15.2, pp. 256–271.

Janssen, Theo (2011). "Montague semantics". In:

Johannsen, Anders, Héctor Martínez Alonso, and Anders Søgaard (2015). "Any-language frame-semantic parsing". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2062–2066.

Johansson, Richard, Karin Friberg Heppin, and Dimitrios Kokkinakis (2012). "Semantic Role Labeling with the Swedish FrameNet". In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12); Istanbul, Turkey; May 23-25*, pp. 3697–3700. ISBN: 978-2-9517408-7-7.

Johansson, Richard, Karin Friberg Heppin, and Dimitrios Kokkinakis (2012). "Semantic Role Labeling with the Swedish FrameNet." In: *LREC*, pp. 3697–3700.

Johansson, Richard and Pierre Nugues (2007). "LTH: semantic structure extraction using nonprojective dependency trees". In: *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)*, pp. 227–230.

Karthikeyan, Kaliyaperumal et al. (2019). "Cross-lingual ability of multilingual bert: An empirical study". In: *International Conference on Learning Representations*.

Katz, Jerrold J (1964). "Semantic theory and the meaning of'good'". In: *The Journal of Philosophy* 61.23, pp. 739–766.

Katz, Jerrold J and Jerry A Fodor (1963). "The structure of a semantic theory". In: *language* 39.2, pp. 170–210.

Khoo, Christopher SG and Sathik Basha Johnkhan (2018). "Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons". In: *Journal of Information Science* 44.4, pp. 491–511.

Kibler, Dennis and David W Aha (1987). "Learning representative exemplars of concepts: An initial case study". In: *Proceedings of the Fourth International Workshop on Machine Learning*. Elsevier, pp. 24–30.

Kingsbury, Paul and Martha Palmer (2003). "Propbank: the next level of treebank". In: *Proceedings of Treebanks and lexical Theories*. Vol. 3. Citeseer.

Klare, Brendan F. and Anil K. Jain (2013). "Heterogeneous Face Recognition Using Kernel Prototype Similarities". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.6, pp. 1410–1422. DOI: `10.1109/TPAMI.2012.229`.

Koppel, Moshe and Noam Ordan (2011). "Translationese and its dialects". In: *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 1318–1326.

Kozhevnikov, Mikhail and Ivan Titov (2013). "Cross-lingual transfer of semantic role labeling models". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1190–1200.

Kripke, Saul A (1972). "Naming and necessity". In: *Semantics of natural language*. Springer, pp. 253–355.

Lakoff, George (1999). "Cognitive models and prototype theory". In: *Concepts: Core Readings*, pp. 391–421.

Lample, Guillaume et al. (2017). "Unsupervised Machine Translation Using Monolingual Corpora Only". In: *arXiv preprint arXiv:1711.00043*.

Leech, Geoffrey (1974). "Semantics". In:

Lenci, Alessandro et al. (2021). "A comprehensive comparative evaluation and analysis of Distributional Semantic Models". In: *arXiv preprint arXiv:2105.09825*.

Lewis, David D (1995). "A sequential algorithm for training text classifiers: Corrigendum and additional data". In: *Acm Sigir Forum*. Vol. 29. 2. ACM New York, NY, USA, pp. 13–19.

Li, Jihong, Ruibo Wang, and Yahui Gao (2010). "Sequential Tagging of Semantic Roles on Chinese FrameNet". In: *23rd International Conference on Computational Linguistics*. Citeseer, p. 22.

Li, Yuxuan and Xiuzhen Zhang (2011). "Improving k nearest neighbor with exemplar generalization for imbalanced classification". In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, pp. 321–332.

Lin, Dekang and Patrick Pantel (2001). "Dirt@ sbt@ discovery of inference rules from text". In: *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 323–328.

Lowe, Will (2001). "Towards a theory of semantic space". In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 23. 23.

Luong, Thang, Hieu Pham, and Christopher D. Manning (June 2015). "Bilingual Word Representations with Monolingual Quality in Mind". In: *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*. Denver, Colorado: Association for Computational Linguistics, pp. 151–159. DOI: 10.3115/v1/W15-1521. URL: https://www.aclweb.org/anthology/W15-1521.

Malt, Barbara C (1989). "An on-line investigation of prototype and exemplar strategies in classification." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 15.4, p. 539.

Manning, Christopher D, Hinrich Schütze, and Prabhakar Raghavan (2008). *Introduction to information retrieval.* Cambridge university press.

Marie, Candito and Seddah Djamé (June 2012). "Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical". In: *TALN 2012 - 19e conférence sur le Traitement Automatique des Langues Naturelles.* Grenoble, France. URL: `https://hal.inria.fr/hal-00698938`.

Marie, Candito et al. (May 2014). "Developing a French FrameNet: Methodology and First results". In: *LREC - The 9th edition of the Language Resources and Evaluation Conference.* Reykjavik, Iceland. URL: `https://hal.inria.fr/hal-01022385`.

Marzinotto, Gabriel (May 2020). "FrameNet Annotations Alignment using Attention-based Machine Translation". English. In: *Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet.* Marseille, France: European Language Resources Association, pp. 41–47. ISBN: 979-10-95546-58-0. URL: `https://www.aclweb.org/anthology/2020.framenet-1.6`.

McCarthy, Diana, Marianna Apidianaki, and Katrin Erk (2016). "Word sense clustering and clusterability". In: *Computational Linguistics* 42.2, pp. 245–275.

McDonald, Ryan et al. (2013). "Universal dependency annotation for multilingual parsing". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 92–97.

Michalon, Olivier et al. (Dec. 2016). "Deeper syntax for better semantic parsing". In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan: The COLING 2016 Organizing Committee, pp. 409–420. URL: `https://www.aclweb.org/anthology/C16-1040`.

Mikolov, Tomas, Quoc V. Le, and Ilya Sutskever (2013). "Exploiting Similarities among Languages for Machine Translation". In: *CoRR* abs/1309.4168. arXiv: `1309.4168`. URL: `http://arxiv.org/abs/1309.4168`.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig (June 2013). "Linguistic Regularities in Continuous Space Word Representations". In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Georgia: Association for Computational Linguistics, pp. 746–751. URL: `https://www.aclweb.org/anthology/N13-1090`.

Mikolov, Tomas et al. (2013). "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781*.

Mitchell, Jeff and Mirella Lapata (2008). "Vector-based models of semantic composition". In: *proceedings of ACL-08: HLT*, pp. 236–244.

Montague, Richard (1970). "Universal grammar". In: *Theoria* 36.3, pp. 373–398.

Mulcaire, Phoebe, Swabha Swayamdipta, and Noah A. Smith (July 2018). "Polyglot Semantic Role Labeling". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 667–672. DOI: `10.18653/v1/P18-2106`. URL: `https://www.aclweb.org/anthology/P18-2106`.

Nosofsky, Robert M (1986). "Attention, similarity, and the identification–categorization relationship." In: *Journal of experimental psychology: General* 115.1, p. 39.

Ohara, Kyoko (2020). "Finding Corresponding Constructions in English and Japanese in a TED Talk Parallel Corpus using Frames-and-Constructions Analysis". In: *Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet*, pp. 8–12.

Ohara, Kyoko Hirose et al. (2003). "The Japanese FrameNet project: A preliminary report". In: *Proceedings of pacific association for computational linguistics*. Citeseer, pp. 249–254.

Osswald, Rainer and Robert D Van Valin (2014). "FrameNet, frame structure, and the syntax-semantics interface". In: *Frames and concept types*. Springer, pp. 125–156.

Padó, Sebastian (2007). "Cross-lingual annotation projection models for role-semantic information". PhD thesis. Saarland: Saarland University.

Padó, Sebastian and Mirella Lapata (2005). "Cross-linguistic projection of role-semantic information". In: *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pp. 859–866.

Padó, Sebastian (2007). "Cross-lingual Parallelism and Translational Equivalence: The Case of FrameNet Frames". In: *Proceedings of the NODALIDA Workshop on Building Frame Semantics Resources for Scandinavian and Baltic Languages*. Tartu, Estonia. URL: `https://www.nlpado.de/~sebastian/pub/papers/nodalida07_pado.pdf`.

Pavlick, Ellie et al. (2015). "Framenet+: Fast paraphrastic tripling of framenet". In: *Proceedings of the 53rd Annual Meeting of the Association for Compu-*

tational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pp. 408–413.

Peters, Matthew et al. (June 2018). "Deep Contextualized Word Representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 2227–2237. DOI: 10.18653/v1/N18-1202. URL: https://www.aclweb.org/anthology/N18-1202.

Petruck, Miriam RL (1996). "Frame semantics". In: *Handbook of pragmatics* 1, p. 13.

Petruck, Miriam RL and Hans C Boas (2003). "All in a day's week". In: *Proceedings of CIL*. Vol. 17.

Petruck, Miriam RL and Gerard De Melo (2012). "Precedes: A semantic relation in FrameNet". In: *Proceedings of the Workshop on Language Resources for Public Security Applications*. European Language Resource Association Istanbul, Turkey, pp. 45–49.

Petruck, Miriam RL et al. (2004). "Reframing framenet data". In: *Proceedings of The 11th EURALEX International Congress*. Université de Bretagne Sud Lorient, France, pp. 405–416.

Pires Telmo andcSchlinger, Eva and Dan Garrette (July 2019). "How Multilingual is Multilingual BERT?" In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 4996–5001. DOI: 10.18653/v1/P19-1493. URL: https://www.aclweb.org/anthology/P19-1493.

Plas, Lonneke Van der, Paola Merlo, and James Henderson (2011). "Scaling up automatic cross-lingual semantic role annotation". In: *Proceedings of the*

49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 299–304.

Popov, Alexander and Jennifer Sikos (2019). “Graph Embeddings for Frame Identification”. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pp. 939–948.

Posner, Michael I and Steven W Keele (1968). “On the genesis of abstract ideas.” In: *Journal of experimental psychology* 77.3p1, p. 353.

Quattoni, Ariadna, Michael Collins, and Trevor Darrell (2008). “Transfer learning for image classification with sparse prototype representations”. In: *2008 IEEE Conference on Computer Vision and Pattern Recognition.* IEEE, pp. 1–8.

Rehbein, Ines et al. (2016). “Adding nominal spice to SALSA – frame-semantic annotation of German nouns and verbs”. en. In: ed. by Jeremy Jancsary. 11 th Conference on Natural Language Processing (KONVENS). Empirical Methods in Natural Language Processing. Eigenverlag ÖGAI, pp. 89 –97. ISBN: 3 - 85027 - 005 - X. URL: `http://nbn-resolving.de/urn:nbn:de:bsz:mh39-52542`.

Reisinger, Joseph and Raymond Mooney (2010). “Multi-prototype vector-space models of word meaning”. In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 109–117.

Resnik, Philip (1996). “Selectional constraints: An information-theoretic model and its computational realization”. In: *Cognition* 61.1-2, pp. 127–159.

Rosch, Eleanor H (1973a). “Natural categories”. In: *Cognitive psychology* 4.3, pp. 328–350.

— (1973b). “On the internal structure of perceptual and semantic categories”. In: *Cognitive development and acquisition of language*. Elsevier, pp. 111–144.

Roth, Michael and Anette Frank (2013). “Automatically identifying implicit arguments to improve argument linking and coherence modeling”. In: *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pp. 306–316.

Roth, Michael and Mirella Lapata (2015). “Context-aware frame-semantic role labeling”. In: *Transactions of the Association for Computational Linguistics* 3, pp. 449–460.

Roy, Nicholas and Andrew McCallum (2001). “Toward optimal active learning through sampling estimation of error reduction. Int. Conf. on Machine Learning”. In:

Ruppenhofer, Josef et al. (2006). “FrameNet II: Extended theory and practice”. In:

Ruppenhofer, Josef et al. (2010). “Semeval-2010 task 10: Linking events and their participants in discourse”. In:

Salkie, Raphael (2002). “How can linguists profit from parallel corpora?” In: *Parallel corpora, parallel worlds*. Brill Rodopi, pp. 91–109.

Salton, G (1971). “The SMART system”. In: *Retrieval Results and Future Plans*.

Samardžic, Tanja et al. (2010). “The scope and the sources of variation in verbal predicates in English and French”. In: *Ninth International Workshop on Treebanks and Linguistic Theories*, p. 199.

Schank, Roger C and Robert P Abelson (1988). “Scripts, plans, goals, and understanding.” In:

Scheffczyk, Jan, Adam Pease, and Michael Ellsworth (2006). “Linking FrameNet to the suggested upper merged ontology”. In: *Frontiers in artificial intelligence and applications* 150, p. 289.

Schiller, Anne et al. (1999). *Guidelines für das Tagging deutscher Textcorpora mit STTS.* Tech. rep. Technical Report. Institut für maschinelle Sprachverarbeitung, Stuttgart.

Schreyer, Rüdiger (1978). “On paraphrase relations”. In: *InWortstellung in Bedeutung: Akten des* 12, pp. 225–235.

Schütze, Hinrich (1998). “Automatic word sense discrimination”. In: *Computational linguistics* 24.1, pp. 97–123.

Shen, Dan and Mirella Lapata (June 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)*. Prague, Czech Republic: Association for Computational Linguistics, pp. 12–21. URL: `https://www.aclweb.org/anthology/D07-1002`.

Sikos, Jennifer and Sebastian Padó (2018a). “FrameNet’s Using relation as a source of concept-based paraphrases”. In: *Constructions and Frames* 10.1, pp. 38–60.

— (2018b). “Using embeddings to compare framenet frames across languages”. In: *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, pp. 91–101.

— (2019). “Frame identification as categorization: Exemplars vs prototypes in embeddingland”. In: *Proceedings of the 13th International Conference on Computational Semantics-Long Papers*, pp. 295–306.

Sikos, Jennifer, Michael Roth, and Sebastian Padó. “Improving multilingual frame identification by estimating frame transferability”. ”Under review”.

— (2022). "Improving Multilingual Frame Identification by Estimating Frame Transferability". In: *Linguistic Issues in Language Technology* 19. DOI: 10.33011/lilt.v19i.939. URL: https://journals.colorado.edu/index.php/lilt/article/view/939.

Sikos, Jennifer, Yannick Versley, and Anette Frank (2016). "Implicit semantic roles in a multilingual setting". In: *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pp. 45–54.

Smith, Edward E and Steven A Sloman (1994). "Similarity-versus rule-based categorization". In: *Memory & Cognition* 22.4, pp. 377–386.

Socher, Richard et al. (2013). "Zero-shot learning through cross-modal transfer". In: *Advances in neural information processing systems*, pp. 935–943.

Søgaard, Anders, Barbara Plank, and Hector Martinez Alonso (2015). "Using frame semantics for knowledge extraction from twitter". In: *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Spreyer, Kathrin and Anette Frank (2008). "Projection-based acquisition of a temporal labeller". In: *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

Surdeanu, Mihai et al. (July 2003). "Using Predicate-Argument Structures for Information Extraction". In: *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. Sapporo, Japan: Association for Computational Linguistics, pp. 8–15. DOI: 10.3115/1075096.1075098. URL: https://www.aclweb.org/anthology/P03-1002.

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le (2014). "Sequence to Sequence Learning with Neural Networks". In: *Advances in Neural Information Processing Systems* 27, pp. 3104–3112.

Swayamdipta, Swabha et al. (2017). "Frame-semantic parsing with softmax-margin segmental rnns and a syntactic scaffold". In: *arXiv preprint arXiv:1706.09528*.

Talmy, Leonard (1985). "Lexicalization patterns: Semantic structure in lexical forms". In: *Language typology and syntactic description* 3.99, pp. 36–149.

Tan, Sang-Sang and Jin-Cheon Na (2019). "Positional Attention-based Frame Identification with BERT: A Deep Learning Approach to Target Disambiguation and Semantic Frame Selection". In: *arXiv preprint arXiv:1910.14549*.

Tannen, Deborah (1993). "What's in a frame? Surface evidence for underlying expectations". In: *Framing in discourse* 14, p. 56.

Thater, Stefan, Georgiana Dinu, and Manfred Pinkal (2009). "Ranking paraphrases in context". In: *Proceedings of the 2009 Workshop on Applied Textual Inference (TextInfer)*, pp. 44–47.

Tomas, Mikolov et al. (2013). "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems*, pp. 3111–3119.

Tonelli, Sara and Emanuele Pianta (2008). "Frame Information Transfer from English to Italian." In: *LREC*.

Torrent, Tiago Timponi et al. (2014). "Copa 2014 FrameNet Brasil: a frame-based trilingual electronic dictionary for the Football World Cup". In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pp. 10–14.

Torrent, Tiago Timponi et al. (2018). "The Multilingual FrameNet Shared Annotation Task: A Preliminary Report". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) edited by Tiago Timponi Torrent, Lars Borin, and Collin F. Baker*, pp. 7–12.

Turney, Peter D and Patrick Pantel (2010). "From frequency to meaning: Vector space models of semantics". In: *Journal of artificial intelligence research* 37, pp. 141–188.

Upadhyay, Shyam et al. (Aug. 2016). “Cross-lingual Models of Word Embeddings: An Empirical Comparison”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 1661–1670. DOI: 10.18653/v1/P16-1157. URL: https://www.aclweb.org/anthology/P16-1157.

Van Der Maaten, Laurens (2014). “Accelerating t-SNE using tree-based algorithms”. In: *The Journal of Machine Learning Research* 15.1, pp. 3221–3245.

VanNoy, Annika (2017). “Culture specific aspects of semantic frames in multilingual frame descriptions”. PhD thesis.

Vaswani, Ashish et al. (2017). “Attention is all you need”. In: *Advances in neural information processing systems*, pp. 5998–6008.

Voorspoels, Wouter, Wolf Vanpaemel, and Gert Storms (2009). “The role of extensional information in conceptual combination”. In: *Proceedings of the Annual Conference of the Cognitive Science Society*. Vol. 31. Citeseer, pp. 2432–2437.

Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann (2005). “Recognizing contextual polarity in phrase-level sentiment analysis”. In: *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pp. 347–354.

Wu, Shijie and Mark Dredze (Nov. 2019). “Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT”. 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)*. Hong Kong, China: Association for Computational

Linguistics, pp. 833–844. DOI: `10.18653/v1/D19-1077`. URL: `https://www.aclweb.org/anthology/D19-1077`.

Xing, Chao et al. (2015). "Normalized word embedding and orthogonal transform for bilingual word translation". In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1006–1011.

Yi, Dong, Zhen Lei, and Stan Z Li (2015). "Shared representation learning for heterogenous face recognition". In: *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. Vol. 1. IEEE, pp. 1–7.

Yuret, Deniz and Mehmet Ali Yatbaz (2010). "The Noisy Channel Model for Unsupervised Word Sense Disambiguation". In: *Computational Linguistics* 36.1, pp. 111–127. DOI: `10.1162/coli.2010.36.1.36103`. URL: `https://www.aclweb.org/anthology/J10-1004`.

Zhang, Bin (2001). "Dependence of clustering algorithm performance on clusteredness of data". In: *HP Labs Technical Report HPL-2001-91*.

Zhong, Huaping et al. (2015). "Aligning knowledge and text embeddings by entity descriptions". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 267–272.

Zubek, Julian and Ludmila Kuncheva (2018). "Learning from Exemplars and Prototypes in Machine Learning and Psychology". In: *arXiv preprint arXiv:1806.01130*.