Toward a Concept-Based Theory of Lexical Semantics

by

Bradley Hauer

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Computing Science

University of Alberta

Abstract

This work aims to address the lack of clear theoretical foundations in computational lexical semantics, the sub-field of natural language processing pertaining to computing with the meaning of words. Semantic tasks are of interest for end-user applications (e.g. contextual translation), downstream tasks (e.g. semantic parsing), and evaluating language models and contextualized representations. Nevertheless, the linguistic phenomena on which semantic methods depend – such as senses, synonymy, and translation – lack a clear, coherent theory, consisting of explicitly stated assumptions, definitions, and proven theorems. Further, there is a deficiency of prior work empirically assessing the utility of such theoretical developments. This thesis, in short, argues for a theory of lexical semantics grounded in universal lexical concepts, and demonstrates, via experimental evidence, that such a theory is important for developing novel, useful, interpretable methods and resources.

The thesis begins with a novel theoretical model for *wordnets*, a class of resources commonly used in lexical semantics. Key definitions are grounded in lexical concepts, culminating in an empirically validated set of best practices for wordnet construction. Next is an investigation of word senses and translations, beginning at the level of *lexemes*, the most basic level of semantic distinction in a lexicon, and proceeding to more fine-grained sense distinctions. This ultimately yields a novel method semantic tagging method, with applications to *word sense disambiguation*. The thesis concludes with a novel analysis of semantic tasks themselves, borrowing from theoretical computing science the notion of *reducibility*, and finally proposing the first *taxonomy of semantic tasks*. Taken together, these contributions represent substantial progress toward the development of a theory of semantics, and for the development of interpretable methods and resources for semantic tasks.

Preface

This thesis is principally comprised of five academic papers, three of which have been published in peer-reviewed venues. Early versions of the other two papers are publicly available as pre-print documents. Each of these papers represents a collaborative effort between myself and my supervisor, Dr. Greg Kondrak. I was involved with all aspects of each paper, from the initial theoretical and empirical design, to implementation and programming, to writing the papers. Except where otherwise noted, I implemented and conducted all experiments presented in this thesis.

Modifications to the published papers have been made where necessary, and to improve formatting and ease of reading, but the content has generally been preserved. In particular, although I am listed as the sole author of this thesis, the use of first-person plural pronouns ("we", "our", etc.) has been retained throughout.

Chapter 2 is available in pre-print (Hauer and Kondrak, 2020b).

Chapter 3 was published in the Proceedings of the AAAI Conference on Artificial Intelligence in 2020 (Hauer and Kondrak, 2020a).

Chapter 4 is available in pre-print (Hauer and Kondrak, 2021).

Chapter 5 was published in the Proceedings of the 2022 Conference of NAACL HLT (Hauer and Kondrak, 2022).

Chapter 6 was published in the Findings of the Association for Computational Linguistics: ACL 2023 (Hauer and Kondrak, 2023).

Citations of these and other papers published during my doctoral program are included in the references section at the end of the thesis.

Material from these papers has in some cases been relocated, copied, or adapted to other parts of the thesis.

Acknowledgements

I am grateful for the assistance and support of my supervisor, Greg Kondrak, and the members of my supervisory and examining committees: David Beck, Colin Cherry, Roberto Navigli, and Robert Holte. Their valuable insights, suggestions, and feedback were vital to the preparation of this document. I would also like to thank the reviewers of the papers which comprise this thesis, for their insightful and helpful comments. This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Alberta Machine Intelligence Institute (Amii).

Contents

1 Introduction **1**

1.1 Prior Work on Lexical Semantics 3

1.2 Outline 7

1.2.1 A Theory of Sense, Synonymy, and Translation 7

1.2.2 Translation and Lexical Semantics 7

1.2.3 Analysis of Semantic Tasks 8

2 A Theory of Sense, Synonymy, and Translational Equivalence **10**

2.1 Introduction 10

2.2 Related Work 12

2.2.1 Princeton WordNet 12

2.2.2 Multilinguality and Multi-Wordnets 13

2.3 Theory 14

2.3.1 Words and Sentences 14

2.3.2 The Sender Axiom 15

2.3.3 Concepts and Senses 15

2.3.4 Synonymy 16

2.3.5 Translational Equivalence 17

2.3.6 Wordnets and Synsets 17

2.3.7 The Wordnet Theorem 18

2.3.8 Multilingual Wordnets 19

2.3.9 The Bitext Theorem 20

2.3.10 Synonymy and Translation of Words 21

2.4 Experimental Evidence 22

2.4.1 Automatic Sense Tag Correction 23

2.4.2 Automatic Wordnet Expansion 26

2.5 Discussion 30

2.5.1 Universality of Concepts 30

2.5.2 Glossability of Concepts 31

2.5.3 Universal Wordnet 32

2.5.4 English Hegemony 33

2.6 Conclusion 34

3 One Homonym per Translation 35

3.1 Introduction . . . 35

3.2 Homonym Hypotheses . . . 38

3.2.1 Preliminaries . . . 38

3.2.2 One Homonym per Translation . . . 39

3.2.3 One Homonym per Discourse . . . 40

3.2.4 One Homonym per Collocation . . . 41

3.2.5 One Homonym per Sense Cluster . . . 42

3.3 Homonym Data . . . 42

3.3.1 Typology of Homonyms . . . 43

3.3.2 List of Type-A Homonyms . . . 44

3.3.3 Mapping WordNet Senses to Homonyms . . . 45

3.4 Homonym Evidence . . . 46

3.4.1 SemCor and Translations . . . 47

3.4.2 WordNet . . . 47

3.4.3 One Homonym per Translation . . . 48

3.4.4 One Homonym per Discourse . . . 50

3.4.5 One Homonym per Collocation . . . 51

3.4.6 One Homonym per Sense Cluster . . . 52

3.5 Conclusion . . . 52

3.6 Addendum: Distinguishing Between Homonymy and Polysemy . . . 53

4 One Sense per Translation 55

4.1 Introduction . . . 55

4.2 Related Work . . . 57

4.3 Mapping Senses and Translations . . . 59

4.3.1 One Sense per Translation (OSPT) . . . 59

4.3.2 One Translation per Sense (OTPS) . . . 61

4.3.3 No Lexical Gaps (NoLG) . . . 61

4.4 Empirical Analysis . . . 62

4.5 Corpus Tagging with OSPT . . . 63

4.5.1 Corpus Tagging . . . 63

4.5.2 Method . . . 65

4.5.3 Intrinsic Evaluation . . . 67

4.5.4 Extrinsic Evaluation . . . 67

4.6 Discussion . . . 72

4.6.1 One Concept per Word: No Polysemy . . . 72

4.6.2 One Word per Concept: No Synonymy . . . 73

4.6.3 One Word $\equiv$ One Concept . . . 74

4.7 Conclusion . . . 75

5 WiC = TSV = WSD: On the Equivalence of Three Semantic Tasks 76

5.1 Introduction . . . 76

5.2 Theoretical Formalization . . . 78

5.2.1 Problem Definitions . . . 78

5.2.2 Problem Equivalence 79
5.2.3 Problem Reductions 79
5.3 WiC Datasets 82
5.3.1 WiC 82
5.3.2 WiC-TSV 82
5.3.3 MCL-WiC 83
5.4 Empirical Validation 84
5.4.1 Systems 84
5.4.2 Solving WSD with WiC 85
5.4.3 Solving WiC with WSD 86
5.4.4 Manual Annotation Experiment 90
5.5 Discussion 91
5.6 Conclusion 93

6 Taxonomy of Problems in Lexical Semantics 94
6.1 Introduction 94
6.2 Theoretical Formalization 96
6.2.1 Words 96
6.2.2 Contexts and Concepts 97
6.2.3 Monolexical Problems 97
6.2.4 Word-in-Context Problems 99
6.2.5 Lexical Substitution Problems 99
6.2.6 Word Synonymy Problems 101
6.3 Problem Reductions 102
6.3.1 *Syn $\leq$ *LexSub $\leq$ *WiC 102
6.3.2 Reductions to WSD 103
6.3.3 MultiWiC $\leq$ MultiLexSub 104
6.4 Relationship to Synsets 104
6.4.1 *Syn $\leq$ Sense $\leq$ WSD 105
6.4.2 Substitution Lemma 106
6.5 Empirical Validation 106
6.5.1 Solving MultiWiC with WiC 107
6.5.2 Solving MultiWiC with WSD 109
6.5.3 Solving MultiLexSub with WSD 110
6.6 Conclusion 112

7 Conclusion 113

References 116

List of Tables

2.1 Senses, words, and concepts. 16
2.2 Results for our sense tag correction method. 25
2.3 Evaluation of our wordnet expansion method. 29

3.1 Senses and homonyms of “bank”. 36
3.2 Sample of our homonym resource. 44
3.3 Corpora sizes. 48
3.4 Summary of evidence for our hypotheses. 48

4.1 The percentage of English polysemous words in BabelNet which exhibit each of the three sense-translation mapping properties with respect to three languages of translation. 63
4.2 Statistics on the sets of sense annotations generated using the three filtering procedures. 68
4.3 Number of instances in each of the subsets of each dataset and the concatenation of all five datasets. 69
4.4 F1-scores (in %) on the 10 WSD test sets. SE07 is the development set. The best results are in bold. 71
4.5 F1-scores (in %) on subsets of the concatenation of all R17 datasets. The best results are in bold. 72

5.1 Examples of senses not found in WordNet. 89

List of Figures

2.1 Pseudocode for our sense tag correction method. 23

3.1 An exception to the one translation per discourse hypothesis. 41

3.2 The diachronic distinction between two types of homonyms. 44

4.1 An example mapping between senses and translations. Each translation corresponds to at least one sense. 60

4.2 Pseudo-code for the sense tagging algorithm. 65

5.1 Three problem reductions. 80

6.1 Taxonomy of problems in lexical semantics. 96

Chapter 1

Introduction

Computational lexical semantics is the study of automated processes which depend on the meaning of words in human languages (Jurafsky and Martin, 2009). To enable machines to answer questions, translate text, retrieve information, or any of a variety of other important and interesting *natural language processing (NLP)* tasks, a computational representation of word meaning is required. This is complicated by the *semantic ambiguity* of natural languages, the uncertainty of word meaning caused by the ability for words to express different meanings depending on the context. For example, given the question "Do bats live nearby?" an automatic question answering system would need to infer that the user is most likely asking about the "animal" sense of *bat*, rather than the "club" sense, and answer accordingly. The task of computationally resolving this semantic ambiguity by mapping the meaning of a word to an entry in a *sense inventory* – that is, of automatically identifying the sense of a word in context – is known as *word sense disambiguation (WSD)*, and has long been a central focus of research in lexical semantics. Further, if it was necessary to automatically translate the aforementioned question into, for example, French, a translation program would need to be able to recognize that *bat* should be translated as *chauve-souris*, rather than *batte*. Thus, it should be clear that resolving semantic ambiguity, and the study of semantics in general, has been, and remains, vital to the progress of natural language processing (NLP), the field of artificial intelligence research involving human languages (Navigli, 2018).

Beyond applications to human end-users, as in the above examples of question answering and translation, semantic tasks have been shown to be useful as a precursor to tasks outside of lexical semantics. For example, the task of semantic parsing, in which the meaning of an input sentence must be represented in a structured machine-readable format, has been shown to benefit from the use of a WSD system (Martínez Lorenzo et al., 2022). While continuous vector embeddings of semantic knowledge have dominated the field in recent years, such findings demonstrate that the ability to link words in context to discrete entries in a database is still relevant to NLP research.

On the subject of contextualized embeddings, semantic tasks are also of interest as a means of evaluating representations of linguistic phenomena. For example, Loureiro et al. (2022) propose a novel method of creating embeddings of word senses, via pre-trained language models based on the transformer architecture. To demonstrate the utility of this method, they apply their embeddings to a variety of semantic tasks such as WSD, the word-in-context (WiC) task (see Chapter 5), and sense similarity, with stronger results being interpreted as evidence of better representations. Lexical semantics therefore remains vital for measuring progress in NLP, particularly with recent proposals of challenge datasets (Maru et al., 2022).

Despite its importance to NLP and AI in general, lexical semantics suffers from a lack of a sound, human-readable theoretical foundation. Contemporary semantic research is focused almost exclusively on improving performance on benchmark datasets (Tedeschi et al., 2023), leading to a "scientific debt" (Nityasya et al., 2023), a sacrifice of scientific understanding and predictability in favor of a purely engineering-based approach to achieve higher performance on benchmarks. Definitions are often inconsistent and unclear (Urešová et al., 2018); assumptions are often unstated and untested (Yao et al., 2012). Perhaps most vitally, there is no clearly articulated set of axioms and theorems for lexical semantics. Having a set of fundamental assumptions, and a sound exploration of what follows from them, would facilitate the establishment of a set of best practices, that is, what properties methods and resources ought to have. It would also set clear expectations arising from those practices: what we can reasonably expect to do with those methods and resources. The intuition is that scientific results and artifacts based on a clearly stated theory of lexical semantics will benefit from greater interpretability and predictability.

This thesis is comprised of five principal chapters, each one aimed at bolstering the theoretical foundations of computational lexical semantics. Each contains a discussion of pertinent concepts (tasks, resources, etc.) as well as relevant prior work. Each contains not only clearly stated theoretical claims, but also empirical validation on previously published datasets and resources, supporting the soundness of our theoretical arguments. The essential thesis statement of this work is as follows: **An empirically-validated theory of sense, synonymy, translation, and lexical concepts yields an improved understanding of lexical resources, methods and tasks, including novel evidence for linguistic hypotheses, and a taxonomy of semantic problems.**

1.1 Prior Work on Lexical Semantics

In this section, we will give a brief overview of relevant historical trends in computational lexical semantics, with a focus on its flagship task, word sense disambiguation (WSD), and, in particular, its relation to translation. We will follow the development of the field from its origins in the earliest days of natural language processing, to the present day, with methods reaching the noise ceiling imposed by human inter-annotator agreement.

As mentioned above, WSD is the task of automatically labeling a word in context with its *sense*, chosen from a given *sense inventory*. A classic example of an ambiguous word is *bank*, which may be used to refer to a financial institution (as in, “The bank hired a new manager.”), a building owned by such an institution, (as in, “The bank is near the supermarket.”), or sloping land near a river (as in, “The bank of the river was slippery.”). We will formulate a precise definition of sense, alongside other terms, in Chapter 2; there we will also discuss the assumptions we make regarding WSD. For now, it will suffice to think of a sense inventory as a list of the meanings of a word in a dictionary, each represented by a definition or *gloss* describing its meaning (indeed, early work in WSD used exactly this formulation).

The origins of lexical semantics, and word sense disambiguation in particular, relate to the necessity of understanding word meaning for machine translation (Weaver, 1949). Decades later, the proliferation of machine-readable dictionaries would enable the work of Lesk (1986), who suggested a dictionary-based WSD method which would be come the first true baseline method for WSD.

In the 1990s, the increasing availability of multilingual resources ushered in the era of *translations as sense inventories (TSI)*. Brown et al. (1991) and Dagan et al. (1991) developed statistical approaches to WSD, with the former presenting a direct application to statistical machine translation. The central idea is that different senses of a word translate differently; thus, knowledge of the sense of a word could facilitate its translation, and knowledge of the translation of a word could help identify its sense. Gale et al. (1992b) were the first to explicitly define WSD in terms of identifying the correct translation, for example, reducing the task of distinguishing the "tax" and "obligation" senses of *duty* to choosing the correct French translation (*droit* or *devoir*). The TSI paradigm influenced the landmark WSD work of Yarowsky (1995) and Schütze (1998). By the late 1990s, the alignment of sense distinctions with translation distinctions was directly proposed by Resnik and Yarowsky (1997).

Just as multilingual data and translation dominated the WSD literature in the 1990s, WSD in the 2000s was heavily influenced by the rising popularity of WordNet (Miller et al., 1990). WordNet is a lexico-semantic knowledge base for English which arranges words into sets of synonyms, or *synsets*. Each synset is associated with a single part of speech: noun, verb, adjective, or adverb. Each synset is also associated with a gloss, and, optionally, one or more example usages of the corresponding concept. For example, one synset contains the words *discipline*, *subject*, and *field*, with the gloss "a branch of knowledge", and examples such as "in what discipline is his doctorate?". A word may be in more than one synset. For example, *field* shares a different synset with *plain* and *champaign*. Synsets are further linked by semantic relations, such as *hypernymy*; the *plain* synset has as its hypernym a synset containing *land*, with the gloss "the solid part of the earth's surface".

Despite being originally designed for psycho-linguistics, it became popular as a freely-available knowledge-rich machine-readable dictionary. In particular, Word-

Net quickly became both the de facto sense inventory and a widely used knowledge-base for English WSD (Navigli, 2009). The notion was that the senses of a word correspond to the synsets it is an element of. Continuing the above example, *field* has WordNet senses meaning “a branch of knowledge” and “extensive tract of level open land”, among others, corresponding to the various synsets containing it. Thus, English WSD, in practice, became the task of identifying which WordNet sense a given word had in a given context.

While concerns were raised about the fine granularity of WordNet senses (Navigli, 2006; Hovy et al., 2006), it nevertheless served as the sense inventory for the first WSD shared tasks, international competitions which challenged contestants to devise novel approaches to WSD on shared datasets. Among the first were shared tasks at Senseval-2 (Edmonds and Cotton, 2001), Senseval-3 (Snyder and Palmer, 2004), and SemEval 2007 (Pradhan et al., 2007). SemCor (Miller et al., 1993), a subset of the English Brown Corpus[1] consisting of over 200,000 tokens annotated with WordNet senses quickly found use as a standard training corpus for supervised WSD[2].

While interest in translation continued throughout the 2000s (Ide, 2000; Chan et al., 2007; Apidianaki, 2008), such work had little apparent influence on the direction of WSD. WordNet and SemCor, purely English resources, were now the core of WSD research. As the 2010s began, Zhong and Ng (2010) published *It Makes Sense (IMS)*, a supervised WSD system which employed the *word expert* model, training a separate supervised machine learning classifier for each word in the training data. The release of IMS can be viewed as representing a firm end to the era of translations as sense inventories. This freely available state-of-the-art approach used exclusively contextual features, achieving remarkable results with no reference to multilingual information. Iacobacci et al. (2016) demonstrated that IMS could be enhanced through the addition of static word embeddings, such as

[1] `http://korpus.uib.no/icame/manuals/brown/`

[2] Following standard machine learning terminology, a computational method for solving a problem is said to be *supervised* if it depends on labelled data, i.e. data on which the task has already been completed, such that the method can be *trained* by example. Methods with no such dependency are *unsupervised*. A *corpus* refers to a large body of text, often organized and associated with supplementary data.

those created by WORD2VEC (Mikolov et al., 2013), a result which, being based on dense embeddings created via neural network models, can be viewed as the beginning of the modern era of WSD research. That IMS continued to see use toward, and beyond, the end of the decade (Scarlini et al., 2019; Hauer et al., 2021b) is a testament to its influence on the field.

In the late 2010s and early 2020s, the transformer architecture (Vaswani et al., 2017) came to dominate countless areas of NLP research, and lexical semantics was no exception. WSD methods leveraging transformer-based pre-trained language models such as BERT (Devlin et al., 2019) had approached, and finally exceeded, 80% accuracy on standard WSD datasets. These include BEM (Blevins and Zettlemoyer, 2020), ESCHER (Barba et al., 2021a), and ConSeC (Barba et al., 2021c). The last of these, ConSeC, remained the state of the art over a year after its publication. As WSD performance now appeared bounded only by the level of human inter-annotator agreement, researchers placed increasing focus on rare senses, with works such as Blevins and Zettlemoyer (2020) and Yoon et al. (2022) evaluating the performance of their models on relatively rare senses. Maru et al. (2022) presented novel datasets designed to test the ability of models to disambiguate challenging instances. Other works have focused on the development of new semantic tasks, with the goal of avoiding WSD and sense inventories entirely. Examples of this trend include the WiC task (Pilehvar and Camacho-Collados, 2019; Martelli et al., 2021), and work on gloss generation (Bevilacqua et al., 2020). Nevertheless, WSD remains useful for downstream applications such as linking tokens to lexical knowledge bases for semantic parsing (Martínez Lorenzo et al., 2022).

Arriving at the present day, we find that there remain vital gaps in the literature of computational lexical semantics. While WordNet and its successors – which we will discuss further in Chapter 2 – have undeniably had a positive impact on the field, there remains a lack of theoretical understanding of the phenomena underlying these resources, or even clear, useful definitions of the relevant terms. It is unclear how such resources should be defined, constructed, or evaluated. The paradigm of using translations as a means of defining sense distinctions, or as a source of knowledge, has been largely discarded, leaving open the question of how

– or if – multilingual knowledge can benefit contemporary lexical semantics. The proliferation of new tasks and benchmarks for language models raises the question of how those tasks relate to one another. These and other open questions make it clear that, while great progress has been made in lexical semantics over the past decades, there is still much more to be done.

1.2 Outline

Having established the history and current state of lexical semantics, this section briefly outlines the content and contributions of this thesis.

1.2.1 A Theory of Sense, Synonymy, and Translation

Chapter 2 begins by addressing by directly addressing the lack of sound theoretical foundations for lexical semantics. In particular, it examines *wordnets*, lexico-semantic resources patterned after the Princeton WordNet (see Section 1.1. Such resources are essential to modern lexical semantics, serving as sense inventories (Raganato et al., 2017), and sources of both structured linguistic knowledge and semantically-disambiguated text such as glosses and examples (Huang et al., 2019). Despite their importance, there are numerous outstanding issues with the definition, construction, and usage of wordnets. We work to resolve these problems by presenting a first-of-its-kind theory of sense, synonymy, and translation. This includes clearly stated definitions, axioms, and theorems, along with an empirical validation on two semantic tasks, and a discussion of the broader implications of our theory.

1.2.2 Translation and Lexical Semantics

It is well known that distinct senses of a word may translate differently. For example, the "flat ground" sense of *field* can be translated into French as *champ*, while the "area of study" sense is translated as *domaine*. On the other hand, this need not be the case: for example, Gale et al. (1992a) observe that the financial and psychological senses English word *interest* can both be translated into French as *intérêt*. Having established a theoretical foundation for reasoning about lexical semantics, we next investigate the relation between the important linguistic phenomena of *senses* and *translations*. This investigation spans Chapters 3 and 4.

In Chapter 3, we investigate the phenomenon of *homonymy*, a rare special case of semantic ambiguity, in contrast to the far more common *polysemy*. Our principal claim in this chapter is our *one homonym per translation* hypothesis (OHPT), which asserts that homonymous senses, senses which are semantically unrelated but nevertheless share an orthographic form, *must* be translated differently. We further advance the novel hypotheses that homonymous senses are not found together in discourses, collocations, or clusterings of fine-grained senses. We describe each of these hypotheses in detail, and demonstrate strong empirical support for each. In the course of these experiments, we also produce a novel resource: a database of English homonymous senses linked to WordNet.

Following that, Chapter 4 explores the historical, theoretical, and empirical use of translation information – including fine-grained sense distinctions found in contemporary semantic resources such as WordNet – to semantic tasks, namely WSD. Early WSD research (approximately speaking, prior to the year 2010) made extensive use of translations as a source of knowledge not only for WSD methods, but for defining senses themselves; indeed, Resnik and Yarowsky (1997) proposed outright "to restrict a word sense inventory to those distinctions that are typically lexicalized cross-linguistically." This chapter discusses the assumptions which would need to hold for this "translations as sense inventories" paradigm to be viable, and empirically demonstrates that these conditions do not obtain. However, we demonstrate that translation information remains useful for lexical semantics, by proposing and evaluating a novel corpus tagging method which exploits the minority of cases where the *one sense per translation* assumption holds. The results show that translation information can be used to improve the accuracy even of modern WSD systems on recently proposed datasets designed to be challenging.

1.2.3 Analysis of Semantic Tasks

Semantic tasks play a vital role in evaluating and comparing models and methods, in addition to providing information to human end-users or downstream tasks.

Nevertheless, they are often vaguely defined (Omarov and Kondrak, 2023), and the relations between them are poorly understood, making it difficult to interpret and analyze results. If two tasks are closely related, claiming state-of-the-art results on both would be less remarkable than obtaining such results on tasks which are independent. Chapters 5 and 6 work to address this deficiency by clearly defining and analyzing semantic tasks.

In Chapter 5, we propose the *sense-meaning hypothesis*: *different instances of a word have the same meaning if and only if they have the same sense.* In other words, we hypothesize that human judgements of sameness of meaning in practical datasets tend to align with sense distinctions in lexical resources. From this hypothesis, we argue that three semantic tasks – word sense disambiguation (WSD), word-in-context (WiC), and target sense verification (TSV) – are equivalent. By "equivalent", we mean that the tasks are pairwise reducible to one another: given an "oracle" method which perfectly solves one of these tasks, analogous methods can be created for the other two, following algorithms that we specify. We then empirically validate these reductions, and find that these experiments support the correctness of our reductions, and the hypothesis on which they are based.

Chapter 6 expands this theoretical analysis of semantic tasks from three problems to thirteen. We generalize the sense-meaning hypothesis to the *concept-meaning hypothesis*: *different word instances have the same meaning if and only if they express the same concept.* This generalization allows us to include cross-lingual and multilingual tasks in our analysis. Our investigation yields a first-of-its-kind taxonomy of problems in lexical semantics. Six of the thirteen included tasks form a set of *wordnet-complete* problems, which are all pairwise equivalent, and to which all other problems in the taxonomy can be reduced. Once again, we provide an empirical validation of our reductions, and the underlying hypothesis.

Chapter 2

A Theory of Sense, Synonymy, and Translational Equivalence

Synonymy and translational equivalence are the relations of sameness of meaning within and across languages. As the principal relations in wordnets, they are vital to computational lexical semantics, which would benefit from a common formal framework to define their properties and relationship. This chapter proposes a unifying theoretical treatment of sense, synonymy, and translational equivalence, along with an experimental validation. The theory establishes a solid foundation for critically re-evaluating prior work in cross-lingual semantics, and facilitating the creation, verification, and amelioration of lexical resources.[1]

2.1 Introduction

Lexical semantics is crucial to natural language understanding, a field which has been identified by Navigli (2018) as a cornerstone of artificial intelligence. Despite its importance and long history, there remains a lack of clear theoretical foundations for the field; instead, research is generally focused on achieving state-of-the-art results on benchmark datasets (Tedeschi et al., 2023). The Princeton WordNet is the prototypical example of the lexico-semantic knowledge bases we call *wordnets*, which are essential to modern lexical semantics. Wordnets often serve as sense inventories and sources of knowledge for methods of solving semantic tasks. However, there is no established theoretical framework for how such resources should

[1] This chapter is based on Hauer and Kondrak (2020b). See the preface for details.

be constructed or evaluated, or how to define and reason about the essential phenomena of sense, synonymy, and translation. Without a strong underlying theory, it is unclear what properties these resources should have, or how best to satisfy them in practice.

Wordnets are comprised of sets of synonymous words, or *synsets*, each associated with a gloss describing the meaning the words in the synset share. BabelNet, a popular example of a multilingual wordnet, contains a synset with the gloss "The child of your aunt or uncle", which (correctly) contains the English word *cousin*. However, it also contains the Spanish word *prima*, which refers specifically to a female child of one's aunt or uncle, as well as non-lexical constructions such as *primo o prima*. Should such cases be strictly regarded as errors, or should wordnets admit a degree of flexibility with respect to translations? This question, among others, remains open, with little prior work discussing the implications of the various possible definitions and design decisions.

In this chapter, we bolster computational lexical semantics with a theory that is sound, empirically validated, and immediately applicable. This theory consists of clearly stated definitions, assumptions, and theorems. In prior work, the notions of senses, synsets, and concepts are often confused, and theoretical assumptions are rarely stated. Our theory seeks to remedy such deficiencies; it provides an explanation of the relationship between synonymy and translational equivalence, as well as the role of these relations as the basis of wordnets. It also leads to the development of a set of best practices for creating multilingual lexical resources, which is currently lacking.

Our key theoretical results are two theorems which we prove follow from our assumptions and definitions. In the first, we show that a key property of synsets can be derived from a minimal set of definitions. This supports the consistency and minimality of our theory, and confirms (often unstated) intuitions. In the second, we show that bilingual dictionaries can be used to create a sufficient condition for identifying literal translations. To empirically validate our theory, we carry out extrinsic evaluations on two tasks: corpus sense tagging, and automatically expanding a wordnet. In the former, we show that automatically generated sense tags can be greatly improved, especially on non-English text, by applying our theory. In the latter, we show that our theory can be used to automatically expand a wordnet by adding senses for another language. We conclude with a discussion of the notion of universality in semantics, and the broader implications of our theory.

2.2 Related Work

In this section, we provide a brief overview of Princeton WordNet, its importance in lexical semantics and NLP, and its generalizations to the multilingual setting.

2.2.1 Princeton WordNet

The Princeton WordNet (Miller et al., 1990), often abbreviated to WordNet, PWN, or WN, is an English lexical knowledge base created to facilitate the study of psycholinguistics, specifically theories of lexical memory. The basic unit of WordNet is the *synset*, defined in the WordNet online documentation[2] as "a list of synonymous words or collocations (e.g., 'fountain pen', 'take in')". Each synset is associated with a specific part of speech: noun, verb, adjective, or adverb. A word may occur in multiple synsets, with each distinct synset occurrence corresponding to a *sense* of the word. For example, *bank* is in 10 noun synsets and 8 verb synsets. WordNet synsets are connected via various relations, such as hyponymy, the "is a" relation, and meronymy, the "has a" relation. Finally, each synset is associated with a gloss describing the shared meaning of the words it contains, and, optionally, one or more example usages.

From its origins in psychology, WordNet has found widespread use in natural language processing, particularly in lexical semantics. It has become the standard sense inventory for English word sense disambiguation (Raganato et al., 2017; Maru et al., 2022), or WSD, providing the base set of sense tags a WSD system must use to classify a given word in context. WordNet was used as the sense inventory for SemCor (Miller et al., 1993), a sense annotated corpus which remains in use decades after its creation as training data for WSD systems (Barba et al.,

[2] `https://wordnet.princeton.edu/documentation/wngloss7wn`

2021c). WordNet's example sentences were used to construct the first Word-in-Context (WiC) dataset (Pilehvar and Camacho-Collados, 2019). It has also been used as a source of knowledge for solving semantic tasks (Huang et al., 2019; Loureiro et al., 2022).

2.2.2 Multilinguality and Multi-Wordnets

Multilingual knowledge has played a key role in the development of lexical semantics (Brown et al., 1991; Dagan et al., 1991), shared tasks (Mihalcea et al., 2010; Lefever and Hoste, 2010), and modern WSD systems (Luan et al., 2020). This has motivated the development of WordNet-like resources — *wordnets* — for languages other than English, or which include information on multiple languages. There are two principal paradigms for creating multilingual wordnets: *expand* and *merge* (Vossen, 1996).

The *expand* model adds words in other languages to the synsets of an existing base wordnet (in practice, typically Princeton WordNet). Examples include MultiWordNet (Pianta et al., 2002), Open Multilingual WordNet (Bond and Foster, 2013), and BabelNet (Navigli and Ponzetto, 2012). Such wordnets tend to be bound by the structure of the Princeton WordNet, with no clear method of handling *lexical gaps*, that is, words or senses in the newly added language that have no equivalent in the base language.

The *merge* model constructs a wordnet independently for each language, and then links them via inter-lingual relations, similar to the intra-lingual relations which exist within monolingual wordnets. Examples include EuroWordNet (Vossen, 2004) and Polish WordNet (Rudnicka et al., 2012). While this method tends to mitigate linguistic bias, it has proven less successful in practice, as both the construction and mapping processes are complex and labor-intensive.

We observe a distinct lack of theoretical investigation or understanding of how multilingual wordnets should be constructed. Assumptions about cross-lingual semantics are often unstated and contradictory across different works (Yao et al., 2012). It is also unclear how to define cross-lingual synonymy (Urešová et al., 2018), and how to define and maintain the essential properties of synsets which contain elements from multiple languages (Kwong, 2018).

2.3 Theory

To facilitate formal description of and reasoning about semantic resources, we formulate the theoretical properties of wordnets, and propose a unified treatment of synonymy and translational equivalence. While these properties are often implicitly assumed in prior work, we precisely formulate their theoretical foundations, and ensure their consistency and minimality.

We view existing lexical resources as imperfect approximations of theoretical models. The divergence of contemporary resources from a hypothetical ideal should not preclude their theoretical analysis.

2.3.1 Words and Sentences

Our basic definitions follow the standard usage in computational lexical semantics. In particular, we define "word" in a way that corresponds with the units of lexical ontologies, which are not necessarily individual orthographic words, but include non-compositional phrases, such as *single out*. Lemmas represent sets of word forms that are associated with certain morpho-syntactic properties (e.g., *cut*, *cuts*, and *cutting*). This is a computational definition, which makes no distinction between words that represent a single lexeme vs. homonyms such as *bank*. We do consider as distinct words that differ in part of speech (POS) or language.

Definition 1. *A* **word** *is a triple consisting of a lemma, POS, and language.*

Our focus is on content words, i.e. words that have semantic value (nouns, verbs, adjectives, and adverbs), as opposed to function words (e.g. determiners, conjunctions). Henceforth, we use the term "word" to refer to content words, unless noted otherwise.

It is important not to conflate words in lexical ontologies ("lemmas") with word instances in text corpora ("word tokens" or "words in context"). We refer to the text containing a word instance as its *context*, and to the word instance itself as the *focus*. Contexts that consist of the same discourse but differ in focus are considered distinct. Following Miller (1995), we refer to the limits of a linguistic context as a *sentence*, which need not necessarily correspond to an orthographic sentence.

2.3.2 The Sender Axiom

While sentences can be ambiguous, we make a simplifying assumption that any given sentence is intended by the sender to have a single specific meaning, even if it may appear ambiguous to the receiver. The intuition is that receivers should be able to clarify or confirm their interpretation of the sentence by responding with a paraphrase that the sender could verify as conveying the same meaning. We refer to this assumption as the *Sender Axiom*.[3]

The Sender Axiom implies that each word instance in a sentence has exactly one meaning, since a sentence containing an ambiguous word instance would necessarily also be ambiguous. The Sender Axiom excludes intentionally ambiguous expressions such as puns, as well as expressions where multiple word instances are compressed into a single orthographic instance, e.g., "Joan subscribes to the newspaper that Bill works for" where the second occurrence of *newspaper* is latent.

2.3.3 Concepts and Senses

A *lexical concept*, or simply *concept*, refers to a discrete unit of word meaning (Miller, 1995), which is unambiguously defined by a *concept gloss*. A gloss is a special type of a context, in which the entire definition is the focus (this will be elaborated on in later chapters). We assume that an expert lexicographer can derive a gloss from a set of contexts in which the concept is expressed.

A word that can express a given concept is said to *lexicalize* the concept. A single concept may be lexicalized by multiple words; for example, the nouns *path* and *route* express the same concept in some contexts.

Word *senses* correspond to distinct concepts that can be expressed by a given word.

Definition 2. *A word* **sense** *is a pairing of a word with a concept.*

[3] In support of this axioms generality, fewer than 0.3% of word tokens in SemCor are annotated with multiple WordNet senses.

	Language E				Language F		
	e_1	e_2	e_3	$\dots$	f_1	f_2	$\dots$
C_1	$s_{1,1}$				$t_{1,1}$		
C_2	$s_{1,2}$	$s_{2,2}$			$t_{1,2}$		
C_3		$s_{2,3}$	$s_{3,3}$			$t_{2,3}$	
$\dots$				$\dots$			$\dots$

Table 2.1: Word senses as the intersection of words (columns) and concepts (rows). Specifically, we could instantiate the variables with the English words e_1 = *earth*, e_2 = *ground*, e_3 = *reason*, and the Italian words f_1 = *terra*, f_2 = *motivo*.

For each word, there is a one-to-one mapping between its senses and the concepts that it can express. The number of senses of each word is equal to the number of concepts that it lexicalizes. A *monosemous* word has only one sense; a *polysemous* word has multiple senses (Miller, 1995). All homonymous words are also polysemous; the converse is not true.

Table 2.1, adapted from Miller et al. (1990), illustrates the relationship between words, concepts and senses: columns correspond to words, rows correspond to concepts, and each non-empty cell is a word sense.

2.3.4 Synonymy

Synonymy is the relation of sameness of meaning (Murphy and Koskela, 2010). Our focus is on word synonymy.

Definition 3. Synonyms *are words that express the same concept given some context.*

For example, the words e_1 and e_2 in Table 2.1 are synonyms, as both can express concept C_2.

Synonymy can be established by a *substitution test*: Two words are considered synonymous if they can be substituted for one another in some sentence without changing its meaning (Murphy and Koskela, 2010). For example, the words *gist* and *essence* are synonyms because the former can be substituted for the latter in the phrase "we understand the gist of the argument". Our definition of word synonymy above implies the correctness of the substitution test: if two words are interchangeable in a sentence, then they express the same concept given the context of that sentence.

In general, word synonymy is *not* an equivalence relation, because it is reflexive and symmetric, but not necessarily transitive. For example, *accusation* and *charge* are synonyms, as are *charge* and *cost*, but *accusation* and *cost* are not synonyms. Words that have the same meaning given *any* context are called absolute synonyms. (Edmonds and Hirst, 2002). Absolute synonymy is an equivalence relation.

2.3.5 Translational Equivalence

The cross-lingual analogue of monolingual synonymy is translational equivalence, the relation of sameness of meaning between expressions in distinct languages, which we refer to as cross-lingual synonymy (Urešová et al., 2018).

Definition 4. Translational equivalents *are cross-lingual synonyms.*

We postulate that the relations of monolingual and cross-lingual synonymy together constitute a single relation of *multilingual synonymy*, which is applicable to any pair of words in the same or different natural languages. For example, the words e_1 and f_1 in Table 2.1 are translational equivalents as both express concept C_2. Under our postulate, e_1, e_2, and f_1 are multi-lingual synonyms.

Cross-lingual synonymy can be established by a *translation test*, a cross-lingual analogue of a substitution test: two words are translational equivalents if one is a literal translation of the other in some sentence, such that the meaning of the sentence is preserved. More generally, the words are required to be mutual literal translations *given some context*. The argument for the correctness of the translation test is analogous to the argument for the substitution test in Section 2.3.4.

2.3.6 Wordnets and Synsets

A *wordnet* is a lexical ontology in which words are organized into sets of synonyms, or *synsets*. Our definition of a synset follows Miller (1995):

Definition 5. *A* **synset** *is the set of words that can express a given concept.*

For example, in the Princeton WordNet, the synset comprised of the words *plain*, *field*, and *champaign* represents a single concept which all three words can express, glossed as “extensive tract of level open land.” Each word instance corresponds to one concept, one sense, and one synset. That synset contains the word, as well as any other words that lexicalize the concept. It further follows that synsets can be equivalently defined as either sets of words or sets of unique word senses.

A different definition of a synset is provided in the Princeton WordNet documentation: “a set of words that are interchangeable in some context.”[4] Unfortunately, this definition fails to exclude the possibility of multiple “duplicate” synsets that correspond to the same concept but different contexts. Such duplicate synsets are considered highly undesirable (McCrae et al., 2020), as they result in duplicate word senses. To avoid this, *a wordnet should contain only one synset that corresponds to any given concept.*

If this uniqueness constraint is satisfied, the following five *synset properties* can be maintained in wordnets.

1. *A word is monosemous iff it is in a single synset. A word is polysemous iff it is in multiple synsets.*
2. *Synonyms share at least one synset. Absolute synonyms share all their synsets.*
3. *Words can express the same concept iff they are in the same synset.*
4. *Every word sense corresponds to exactly one synset.*
5. *Every sense of a polysemous word corresponds to a different synset.*

The above synset properties follow from the preceding definitions and assumptions. The only one that may require a proof is synset property #2. The *Wordnet Theorem* in the next section implies synset property #2, and establishes that the implication also holds in the other direction.

2.3.7 The Wordnet Theorem

Theorem 1. *Words share a synset* **if and only if** *they are synonyms.*

Proof. By the definitions in Sections 2.3.4 and 2.3.5, synonyms, in the same or different languages, can express the same concept, and thus must share the synset that corresponds to the concept. To prove the other direction, let w_x and w_y be

[4] `https://wordnet.princeton.edu/documentation/wngloss7wn`

words that share a synset. Then, both w_x and w_y express the synset's concept given the context of its gloss. Therefore, w_x and w_y are synonyms. $\square$

Why is this theorem important? Isn't the term "synset" short for "synonym set"? In fact, the basic units of wordnets are *sets of word senses*. Our theorem confirms the unproven intuition that they can be represented by sets of synonyms, which is implicitly assumed in the wordnet literature (Miller, 1995).

The Wordnet Theorem therefore establishes that the link between senses and synonymy need not be an axiom. Rather, it can be demonstrated to follow from our concept-based definitions. In addition, the proof of the theorem provides evidence that our theoretical framework is consistent.

Another consequence of Theorem 1 is that attempts to substantially reducing wordnet sense granularity, e.g. by clustering (Navigli, 2006; Hovy et al., 2006), fail to preserve the synset properties from Section 2.3.6. Thus, while the granularity of wordnets may occasionally be a practical inconvenience, it is a theoretical necessity.

2.3.8 Multilingual Wordnets

Under our unified treatment of monolingual and cross-lingual synonymy, the Wordnet Theorem also establishes the theoretical foundation of multilingual wordnets, such as BabelNet (Navigli and Ponzetto, 2012) and Open Multilingual WordNet (Bond and Foster, 2013). Like monolingual wordnets, they are comprised of interconnected synsets which contain words that can express a given concept. Since the Wordnet Theorem makes no assumptions about the languages of the words, its proof establishes that words from distinct languages share a synset *if and only if* they are cross-lingual synonyms.

The synset properties specified in Section 2.3.6 must be likewise maintained in multi-wordnets. In particular, words are translational equivalents if and only if they share a multilingual synset (Navigli and Ponzetto, 2010).

It follows that, given a multilingual wordnet which maintains the synset properties, monolingual synsets can be obtained from the multilingual synsets by restricting them to a given language. Multilingual wordnets typically link their synsets to the Princeton WordNet (Kafe, 2023), and so are expected to preserve its synset properties. However, care must be taken to maintain synonymy within synsets (Kwong, 2018).

2.3.9 The Bitext Theorem

In this section, we state and prove a theorem which establishes a connection between multilingual wordnets and word-aligned parallel corpora (bitexts). While bitexts can be mined for multilingual synonyms (i.e., translational equivalents), not all translation pairs are literal (i.e. meaning-preserving). Theorem 2 provides a way to distinguish between literal and non-literal translations. We posit that any literal translation pair which is identified in a bitext should also share a multilingual synset, which corresponds to the concept that both words express in their respective languages.

Theorem 2. *Let w_e and w_f be aligned words in the sentence S_e and its translation S_f, respectively. If w_e and w_f are synonyms then they express the same concept in S_e and S_f.*

Proof. For the purposes of this proof, we assume that S_f is produced from S_e by a translation agent TA, which could be a human or a computer program. We assume that in creating S_f, TA is guided by two priorities: (1) fidelity (preserving the meaning of S_e), and (2) brevity; a translation composed of fewer words is preferred to a longer translation.

In order to preserve meaning (the first priority), for every word w_e in S_e, TA must identify the concept s that is expressed by w_e, and attempt to find a word or phrase that expresses s in the target language. If such a cross-lingual synonym w_f can be found, TA will prefer it to a phrase, because of the second priority, conciseness. Thus, both w_e and w_f will express the same concept.

On the other hand, if TA cannot find a single word w_f that expresses s (for instance, due to a lexical gap), TA will prefer a phrase that preserves the meaning of s to a word that expresses a similar but different concept, as meaning fidelity is a higher priority than brevity. Another option available to TA is to forgo expressing the concept s directly, and instead literally translate a longer segment of S_e which includes w_e. In either case, S_f will not contain a single word w_f that aligns one-to-one with w_e, and so the theorem does not apply. Therefore, if the words w_e and w_f are present and correctly aligned, they must express the same concept. □

Theorem 2 establishes a theoretical foundation for algorithmic methods for synset construction from bitexts, which we explore in Section 2.4.2. Specifically, if aligned sentences are literally translated, and the aligned words are synonyms, then the two word instances express the same concept. If we can establish the sense of one of the two words, its translation can be immediately disambiguated as well. Furthermore, both words should be in the synset that corresponds to that concept.

2.3.10 Synonymy and Translation of Words

Yao et al. (2012) observe that prior work relating to word senses and translations, such as Gale et al. (1992a) and Diab and Resnik (2002), tend to make one of the two "alternate" assumptions, which have the same antecedent but different consequents:

Antecedent: *Two different words $e_x \in E$ and $e_y \in E$ are aligned to the same word f in language F.*

Consequents:

1. f is polysemous ("polysemy assumption")

2. e_x and e_y are synonyms ("synonymy assumption")

Yao et al. (2012) perform experiments on two bilingual corpora, using a lexical sample of 50 words from OntoNotes (Hovy et al., 2006), and conclude that neither assumption holds significantly more often than the other. However, they stop short of proposing a principled solution to the problem.

In our view, neither of the two assumptions need hold universally. For example, although both *time* and *weather* are translations of the Italian word *tempo*, it would be wrong to conclude that the two English words are synonyms. This is because synonymy of words is not transitive in either monolingual or multilingual setting. On the other hand, although both *bundle* and *package* are translations of the Italian *involto*, this does not imply that the Italian word is polysemous; indeed, both English words translate a single sense of *involto*. In fact, the two consequents are not exclusive; for example, *test* and *trial*, which are synonyms, are both translations of Italian *prova*, which is polysemous.

We postulate that the polysemy and synonymy assumptions can be integrated into a single theorem, which entails a *non-exclusive disjunction* of the two consequents, i.e., one or both of them may be true, but they cannot both be false.

Theorem 3. **If** *words e_x and e_y in language E are both literal translations of word f in language F* **then** *e_x and e_y are synonyms* **or** *f is polysemous.*

Proof. The antecedent implies that there exists a concept s_x which can be expressed by both e_x and f, and that there exists a concept s_y which can be expressed by both e_y and f. If s_x is different from s_y, f can expresses multiple concepts, so it is polysemous; otherwise, e_x and e_y can express the same concept, so they are synonyms. □

Since Theorem 3 is formulated at the level of lemmas, rather than word instances, it is applicable to word translations in bilingual dictionaries, under the assumption that such translations are literal. However, it can also be applied to word instances in bitexts, *provided that the aligned translations are literal,* i.e., they express the same concepts. This can be decided on the basis of Theorem 2 from Section 2.3.9.

In conclusion, our theory provides a theoretical explanation and resolution of the issue raised by Yao et al. (2012): Systems that are based exclusively on one of the two assumptions, such as Bannard and Callison-Burch (2005) or Lefever et al. (2011), fail to consider a substantial number of relevant instances, which adversely affects their effectiveness.

2.4 Experimental Evidence

In this section, we describe experiments that test the predictions of our theory. In particular, we demonstrate how our theory can be used to improve sense tags on parallel corpora and automatically expand wordnets with additional languages.

```
lex(s) - word of which s is a sense
M(s) - multi-synset that contains sense s
M(w) - set of multi-synsets that contain word w
for each aligned sense pair (s, t) do do
    if $CL - Syn(s,t)$ and $M(s) \neq M(t)$ then
        $C \leftarrow M(lex(s)) \cap M(lex(t))$
        if $M(s) \in C$ and $M(t) \notin C$ then
            CORRECT: $t \leftarrow (lex(t), M(s))$
        if $M(s) \notin C$ and $M(t) \in C$ then
            CORRECT: $s \leftarrow (lex(s), M(t))$
        if $M(s) \notin C$ and $M(t) \notin C$ then
            ADD: $lex(t)$ to $M(s)$
            CORRECT: $t \leftarrow (lex(t), M(s))$
```

Figure 2.1: Pseudocode for our sense tag correction method.

2.4.1 Automatic Sense Tag Correction

In this section, we empirically test our theory of sense, synonymy, translation, and multilingual wordnets. Specifically, we apply Theorem 2 (Section 2.3.9), which is based on said theory, to the task of correcting sense tags in a bitext.

Our task is, essentially, word sense disambiguation (WSD), the task of tagging a word in context with the correct entry in a given sense inventory. More specifically, we take as input a bitext that has been so disambiguated, and seek to improve the sense tags by detecting and correcting errors.

Method

Figure 2.1 shows the pseudocode for our method. Similar to Luan et al. (2020), Hauer et al. (2021c), and Mallik and Kondrak (2023), we use translation information to post-process and correct WSD output. Our approach differs in two ways: First, we test whether each pair of aligned words are synonyms. Under Theorem 2, the conjunction of synonymy and alignment provides strong evidence that the aligned words express the same concept – i.e. that the translation is literal – and therefore should have sense tags corresponding to the same multilingual synset. Second, we posit that our theory is sufficiently strong to provide evidence of the existence of senses not attested in the sense inventory. Our method therefore has the ability to tag a token with a synset that does not actually contain the word, effectively adding a new sense to the inventory.

The method works by examining each aligned word pair such that both words are tagged with a sense; call the senses s and t and the words $lex(s)$ and $lex(t)$. If $lex(s)$ and $lex(t)$ are synonyms (denoted by the predicate $CL - Syn(s, t)$), then, by Theorem 2, they must express the same concept. Therefore, if s and t do not refer to the same multilingual synset (denoted $M(s) \neq M(t)$), our theory predicts that either s or t must be an incorrect annotation. Following Mallik and Kondrak (2023), if $M(s)$, the synset to which s corresponds, contains $lex(t)$, but $M(t)$ does not contain $lex(s)$, we replace t with the sense of $lex(t)$ corresponding to $M(s)$. An analogous operation is performed if the roles of s and t are reversed. Different from prior work, if $lex(s)$ is not in $M(t)$ and $lex(t)$ is not in $M(s)$, our method adds a sense of $lex(t)$ corresponding to $M(s)$ – equivalently, it adds word $lex(t)$ to synset $M(s)$ – and replaces t with this new sense. In practice, the language of s will be English; since English WSD performance is typically higher than in other languages, we hypothesize that, in cases such as this, the English WSD output is more likely to be correct.

Resources

We use MultiSemCor (Bentivogli and Pianta, 2005) as our bitext. MultiSemCor, or MSC, was created by manually translating SemCor into Italian. It was then word-aligned using a knowledge-based aligner, KNOWA (Pianta and Bentivogli, 2004), and the gold sense tags from SemCor were propagated to the Italian side using MultiWordNet (Pianta et al., 2002). We use the sense annotations, on both the English and Italian sides, as a ground truth to evaluate against; they are not provided to our method.

We sense tag the corpus by applying AMuSE-WSD (Orlando et al., 2021), with the provided AMUSE-LARGE-MULTILINGUAL-CPU model. It is these semantic labels to which we apply our error correction method. The result is 116884 sense tagged tokens on the English side, and 114629 on the Italian side.

Language(s)	EN	IT	EN+IT
Improved	173	5324	5497
Broken	807	24	831
Neutral	121	1001	1122
Total	1101	6349	7450

Table 2.2: Results for our sense tag correction method.

We use BabelNet 5.1 as our multilingual wordnet, accessed via the Python API[5]. This is used to identify the synsets of a given word, as required by Algorithm 2.1.

We do not use BabelNet to implement the CL-Syn function, in an effort to keep the synonymy predicate independent of our multilingual wordnet. Instead, to evaluate the cross-lingual synonymy predicate, we use the freely available Wiktextract (Ylonen, 2022) and PanLex[6] dictionaries. Specifically, given an aligned sense pair (s, t), $CL - Syn(s, t)$ is true if and only if the words of which s and t are senses are translations according to either of those dictionaries.

Results

Table 2.2 shows the results of our experiment on MSC. Here, “improved” indicates an incorrect sense tag was changed to a correct tag, with the sense annotations provided with MSC serving as the gold standard. Contrariwise, “broken” indicates that a correct sense tag was changed to be incorrect. Finally, “neutral” indicates that the sense tag was incorrect before and after the change.

The results on Italian are particularly strong. Our algorithm proposes more than 6000 corrections to AMuSE-WSD’s sense tags, and approximately 84$\%$ improve the disambiguation with respect to the gold tags. Of the remaining 16$\%$, almost all are neutral, and so have no impact on the accuracy of the sense tags. Overall, our method, based on Theorem 2, substantially improves Italian WSD.

On English, our method proposes roughly one sixth as many corrections, of which the majority are erroneous. We attribute this to AMuSE-WSD’s relatively high reported accuracy on English, exceeding 80$\%$ on some datasets. If the automatic sense tags given to our method are already of relatively high quality, then a

[5] https://babelnet.org/guide

[6] https://dev.panlex.org/interface/

change in those tags is more likely to be incorrect. This empirically justifies our decision to "trust" the English annotation in the event that $M(s) \notin C$ and $M(t) \notin C$.

Overall, of the 7450 corrections proposed by our algorithm, 74% are improvements, while 15% are neutral. Only about 11% degrade the quality of the input sense tags, almost all of them on the English side. We therefore conclude that this experiment provides strong support for the soundness and empirical utility of our theory.

2.4.2 Automatic Wordnet Expansion

One of the central contributions of our theory is a formalization of the wordnet model, particularly the properties of synsets in multilingual wordnets. In this section, we test this theoretical framework by applying it to the task of automating the expand model of multilingual wordnet construction. In particular, we are given a wordnet $\mathcal{W}$ which covers language E, and are tasked with adding words from language F to the synsets of $\mathcal{W}$. We assume access to an E-F bilingual dictionary, and an E-F machine translation model.

Prior approaches to WordNet expansion have depended on large text corpora (Fišer and Sagot, 2015), manual input by lexicographers (Pianta et al., 2002), word sense disambiguation systems (Diab, 2004), other wordnets (De Melo and Weikum, 2009), or other resources (Navigli and Ponzetto, 2010). Our approach therefore serves to demonstrate how a sound theory of multilingual lexical semantics can be used to devise methods which are easier to apply in practice, due to making fewer assumptions about available resources. Moreover, since our method is based on clearly articulated definitions, axioms, and results, the created resources are more readily interpretable.

Method

Given a wordnet $\mathcal{W}$ containing words from a single language E, we first select a subset of its synsets, $\mathcal{S}$ (we will describe how we choose $\mathcal{S}$ in the next section). For each synset $S \in \mathcal{S}$, we retrieve the set L of lemmas it contains, its part of speech p (noun, verb, adjective, or adverb), and its gloss g. We then create a sentence using the template 'In this context, the p "$lemma$" means "g".' Here, $lemma$ is a randomly selected element of L. Thus, we create a template sentence corresponding to each synset $S \in \mathcal{S}$.

We then translate these sentences from language E to language F. For each such sentence, we identify the translation of $lemma$, call it t, by extracting the first quoted string. This heuristic assumes that the translation system preserves the meaning and ordering of quoted strings; this assumption holds reliably in practice.

We then consult the E-F bilingual dictionary, to verify that t is indeed a translation of $lemma$. If so, Theorem 2 predicts that the translation is literal, that is, the translation t expresses the same concept as $lemma$ in the given context, and is not e.g a hypernym or otherwise related word. If this condition holds, we add t to the synset S corresponding to that sentence; equivalently, we create a sense consisting of the lemma t and the concept corresponding to synset S. These new senses comprise the output of our method. We refer to this method as BIDICTNET.

Resources

We use PWN as our wordnet $\mathcal{W}$, making English our source language E. Our language of translation F is French. The set $\mathcal{S}$ of synsets is comprised of 1000 PWN synsets chosen at random. (This was done to reduce the running time of our method and the evaluation.) We take this approach in order to avoid biasing our evaluation toward more frequent senses, or otherwise artificially skewing the distribution of concepts.

For translating the template sentences, we use a commercial translation service provided by DeepL[7]. We lemmatize the translations using SpaCy[8], specifically the FR_CORE_NEWS_MD model. We derive an English-French dictionary from the union of Wiktextract and PanLex, analogous to Section 2.4.1.

Results

Of the 1000 PWN synsets in our random sample, our method creates a total of 426 French senses. That is, for 426 of the template sentences, the English lemma and

[7] https://www.deepl.com/translator

[8] https://spacy.io/

its French translation are found in our dictionary. Note that, since we generate one template sentence per synset, and translate each such sentence once, our method is limited to creating at most one sense per synset. So, in short, our method expands 426 PWN synsets by adding a French lemma.

To automatically evaluate these French senses, we compare them to three freely-available multilingual wordnets: BabelNet (BN), Open Multilingual WordNet (OMW), and Universal Wordnet (UWN). Each of these resources provides links between its synsets and those of PWN. We access BabelNet through the Python API, as before, OMW through the Python NLTK library, and UWN through the provided database[9]. We consider a sense correct if it is found in the corresponding synset of at least one of BN, OMW, or UWN.

We found that 345 of the 426 senses proposed by our method are correct, indicating a precision of 81.0%. We also found that 147 of the 1000 synsets do not contain any French senses in any of BN, OMW, and UWN; under our criteria, it is not correct to propose any French sense for these synsets (they may, for example, represent lexical gaps). Since our method proposes correct French senses for 345 out of 853 synsets with plausible French senses, we interpret this as a recall of 40.4%. These values yield an F1 score of 53.9%.

Comparison to prior work is complicated by the lack of a generally accepted framework for evaluating wordnets. Metric definitions and gold standards vary widely across the literature, and often involve highly subjective manual evaluation. Further complicating analysis is the wide variety of resources used, assumptions made, and manual effort involved in wordnet construction. The most directly comparable example is Sagot and Fišer (2008), who report 80% accuracy upon manual evaluation of their French wordnet WOLF, which we interpret as being comparable to our 81% precision. We perform a more controlled comparison in the next section.

Comparison to UWN

To further evaluate BIDICTNET, and in an effort to develop an evaluation framework for multilingual wordnets, we directly compare the French senses it extracts to

[9] http://wordnets.org/

MLWN	Evaluation	P	R	F
UWN	Senses	50.6	14.1	22.0
BIDICTNET	Senses	77.9	12.2	21.1
UWN	Synset	59.6	27.6	37.7
BIDICTNET	Synset	77.9	33.2	46.6

Table 2.3: Comparison of UWN and our BIDICTNET.

those found in UWN. We chose UWN due to its dependence on automatic methods of identifying synset translations. As our gold standard, we use BN and OMW; that is, a French sense found in UWN or BIDICTNET is correct if it is in BN or OMW.

We consider two approaches to calculating precision and recall: *sense-level evaluation* and *synset-level evaluation*. They vary in how the metrics precision and recall depend on – true positives, false positives, and false negatives – are defined. In sense-level evaluation, a sense is counted as a true positive (i.e. correct) if it is in BN or OWM, otherwise it is a false positive (i.e. incorrect). A sense which is in BN or OMW, but which is not in the wordnet to be evaluated, is a false negative (i.e. an omission).

Synset-level evaluation instead computes these metrics at the level of synsets. A synset is counted as a true positive if a correct sense is proposed for that synset, a false positive if an incorrect sense is proposed for that synset, and a false negative if there is a correct sense that is not proposed. A single synset may, under these conditions, be counted as a true positive, a false positive, and a false negative (or any subset of the three); one can view this as a variant of sense-level evaluation in which each synset contributes at most one to the number of true positives, false positives, and false negatives. In this way, synset-level evaluation avoids giving greater influence to synsets with more target language senses.

In both cases, we sum the true positive, false positive, and false negative values across all synsets, and compute precision, recall, and F1 score using the standard formulae. The results are shown in Table 2.3. Under both evaluation strategies, our BIDICTNET method has substantially higher precision compared to UWN. Using sense-level evaluation has a disproportionately negative impact on BIDICTNET due to its inability to propose more than one sense per synset: for a synset with k senses, BiDictNet will incur at least $k - 1$ false negatives. Despite this, BiDictNet achieves a recall within 2% of UWN, and an F1 result within 1%.

Using synset-level evaluation, BiDictNet gains the advantage in terms of recall, outperforming UWN by 5.6%, while the gap in precision narrows only slightly. BiDictNet achieves an F1 score 8.9% higher than that of UWN.

We interpret these results as strong evidence that our theory-driven, dictionary-based method for expanding a monolingual wordnet is effective, yielding senses which are more precise, and achieving comparable or greater coverage on a random sample of synsets. This finding provides further evidence for the soundness and utility of our theory.

2.5 Discussion

In this section, we discuss the universality of concepts, and the implications of our theory for wordnets.

2.5.1 Universality of Concepts

Lexical concepts are the semantic equivalence classes of word senses, and, equivalently, of words in context. Since wordnet senses are induced by concepts, they are discrete and well-defined, unlike dictionary senses which are designed by lexicographers independently for each word (Kilgarriff, 1997). In a monolingual wordnet, the set of concepts is grounded in word synonymy; one way of verifying synonymy is the substitution test (Section 2.3.4).

While different languages lexicalize different sets of concepts, we posit that *all lexical concepts are universal*. That is, *any lexical concept from any language can be expressed in any other language* (not necessarily by a single word). For example, the concept expressed by the Spanish adverb *anteayer* corresponds to a lexical gap in English, but it can be expressed as *day before yesterday*. We refer to this thesis as *concept universality*.

Concept universality implies that concepts are not language specific, but rather they are drawn by different languages from a single, shared pool of concepts.

If an English word and a Chinese word share meaning, then they lexicalize the same language-independent concept, rather than distinct, language-specific concepts. Once a new concept is lexicalized in any natural language, it is immediately available for other languages to adopt, which is often accomplished by "borrowing" the form of the word.

Concept universality can be equivalently formulated in terms of translation: *any word in context can be literally translated into any other language* (either by a word or a phrase). This thesis can be viewed as a word-level analogue of the *translatability thesis* of Katz (1976: 39), which states that any sentence can be literally translated. For brevity, we will employ the phrase *translate a concept* to mean *literally translate a word instance that expresses that concept.* Thus, we can succinctly express the *concept translatability* thesis as *every concept can be literally translated into any language.*

2.5.2 Glossability of Concepts

Concept universality underlies the idea of multilingual wordnets because every synset *gloss* (i.e., concept definition) is an expression of the corresponding concept. We posit that *a gloss of any lexical concept in any language can be expressed in any language.* We refer to this thesis as *concept glossability*.

We further postulate that concept glossability implies concept universality. That is, if a concept can be glossed in a given language, then it can be expressed in that language. This is because a concept gloss expresses the concept in any context (Section 2.3.3), so it can always serve as a translation of the concept.

Conversely, we postulate that concept universality implies concept glossability. A concept gloss can be derived from a set of contexts in which the concept is expressed (Section 2.3.3). Such a set of contexts can be obtained by *translating* the contexts in which the concept is lexicalized in some source language. This is always possible as concept universality guarantees that any lexical concept can be literally translated into any other language.

We conclude that concept universality, translatability, and glossability are all mutually equivalent. This proposition can be viewed as a lexical semantics analogue to *Turing Completeness*. Just as Turing Completeness establishes a universal set of computable functions, the three concept theses establish a universal set of expressible concepts across natural languages.

2.5.3 Universal Wordnet

Concept universality implies a matching between lexicalized concepts across languages, in which no two matches share a concept. The matching between monolingual concepts is grounded in translational equivalence, which is demonstrable by the translation test (Section 2.3.5). If a lexicalization of concept s_x can literally translate a lexicalization of concept s_y, then $s_x = s_y$. We can refer to a hypothetical wordnet that encompasses all concepts lexicalized in at least one natural language as *universal wordnet*.[10]

Each concept in the universal wordnet corresponds to exactly one universal synset. Universal synsets are sets of intra-lingual and cross-lingual synonyms. The meaning of any word in context corresponds to exactly one universal synset. We posit that other semantic relations, such as hypernymy and meronymy, are also universal. That is, they can be uniquely defined on the set of universal concepts, such that they are consistent with the set of relations in any individual language.

In practice, wordnets should avoid conflating distinct concepts in a single synset. This is particularly important to avoid if a multilingual wordnet is created according to the *expand* paradigm. If any language makes a lexical distinction between concepts then those concepts need to be represented by distinct synsets.[11] This is necessary to ensure that synsets contain all and only translational equivalents, as stipulated by Theorem 1 in Section 2.3.7.

On the other hand, if two distinct synsets in a wordnet correspond to the same universal concept, this should be regarded as an error and corrected. This is particularly important to avoid if a multilingual wordnet is created according to the *merge* paradigm. We can view a pair of monolingual wordnets as a bipartite graph

[10] This theoretical term should not be confused with Universal Wordnet (UWN) (De Melo et al., 2012), which includes only a small fraction of natural languages.

[11] For example, a multilingual wordnet covering English and Chinese should separate lexicalizations of the concepts of “sister”, “elder sister”, and “younger sister” into their respective synsets, of which the last two would contain no English words.

in which nodes are synsets, and edges represent the relation of translational equivalence. Every node in the graph should have a degree of at most one: a synset in one language should not correspond to more than one synset in another language. This is necessary to maintain the synset properties enumerated in Section 2.3.6.

2.5.4 English Hegemony

Our theory entails several practical guidelines for constructing multilingual wordnets on the basis of lexical translation, In particular, it provides a principled solution to the problem of bias towards the set of concepts lexicalized by the base language, which is inherent in the expand model. Multi-wordnets that are founded on the synset structure of the original Princeton WordNet often lack synsets that correspond to lexical gaps in English.[12] This can be resolved by first creating new synsets for all concepts in the target language that are not represented in the base wordnet, and then expanding those synsets to include lexicalizations from all languages under consideration.

Another major source of errors in the expand model is the unconstrained use of contextual translations for populating synsets, which may results in adding non-synonymous target lexicalizations.[13] Our work suggests a principled solution to this problem, which is to only admit translations that can express the same concept. In practice, the literalness of a given translation in context can be verified by a word synonymy check, as stipulated by Theorem 2 (Section 2.3.9).

Our theory also provides guidance for the implementation of the merge model. The merge model avoids the issue of lexical gaps by starting from two complete wordnets. Multilingual synsets can be effectively constructed by adding synonymy links between monolingual synsets. However, as explained above, no synset should be "merged" with more than one other synset, in order to guarantee that the linked synsets correspond to a single universal concept. In addition, our theory supplies a practical method for identifying matching synsets, which is to apply a translation

[12] For example, OMW and MWN have no synset for the concept of "female cousin." As a result, words that lexicalize this concept in other languages may be altogether missing from these resources.

[13] For example, the BabelNet synset for "cousin" includes both *prima* and *primo*, even though the two Spanish words lexicalize the antonymous concepts of female and male cousins, respectively. This is analogous to including *mother* and *father* in the same synset.

test (Section 2.3.5) to the corresponding lexicalizations given the context of a synset gloss.

We hope that our work will lead to more accurate representation of conceptual distinctions in multilingual wordnets, which would facilitate the evolution of these resources away from the hegemony of English, and toward greater linguistic diversity.

2.6 Conclusion

We have proposed a unifying treatment of the notions of sense, synonymy and translational equivalence. The resulting theory formalizes the relationship between words and senses in both monolingual and multilingual settings. In the future, we plan to expand on the application of our theory to the automatic construction of interpretable semantic resources, such as wordnets. We also expect that sound theoretical foundations will also lead to improvements in important semantic tasks.

Chapter 3

One Homonym per Translation

The study of homonymy is vital to resolving fundamental problems in lexical semantics. In this chapter, we propose four hypotheses that characterize the unique behavior of homonyms in the context of translations, discourses, collocations, and sense clusters. We present a new annotated homonym resource that allows us to test our hypotheses on existing WSD resources. The results of the experiments provide strong empirical evidence for the hypotheses. This study represents a step towards a computational method for distinguishing between homonymy and polysemy, and constructing a definitive inventory of coarse-grained senses.[1]

3.1 Introduction

Many words are semantically ambiguous, in that they have multiple senses. The relationship between two senses of a word is called *polysemy* if they are semantically related, and *homonymy* otherwise (Jurafsky and Martin, 2009). Senses that belong to the same homonym are polysemous (e.g. #2 and #5 in Table 3.1), while senses of distinct homonyms are homonymous (e.g. #2 and #1 in Table 3.1).

The differentiation of homonymous and polysemous word senses is one of the central problems of lexicography (Mel'čuk, 2013). A textbook on theoretical semantics devotes an entire chapter to the problem, concluding that it may be insoluble, as the intuitions of native speakers cannot be relied upon (Lyons, 1995).

[1] This chapter is based on Hauer and Kondrak (2020a). See the preface for details. We thank Genna Cockburn, Amy Hua, and Jacob Skitsko for the assistance in preparing the homonym resource. We thank Yixing Luan and Haozhou Pang for performing additional experiments and analysis.

BANK$_n^1$		BANK$_n^2$	
#2	financial institution	#1	sloping land
#5	stock held in reserve	#3	long ridge or pile
#6	funds held by a house	#4	arrangement of objects
#8	container for money	#7	slope in a road
#9	building	#10	flight maneuver

Table 3.1: The senses of the noun "bank" from WordNet 3.0, grouped by its two homonyms.

Psycho-linguistics furnishes evidence for a common representation of closely related senses in the mental lexicon (Brown, 2008), which suggests that NLP applications would benefit from the ability to distinguish homonym-level meaning differences (Utt and Padó, 2011). In fact, standard neural machine translation systems make a substantial number of errors on homonyms (Liu et al., 2018).

The study of homonymy is also of utmost importance to the problem of establishing the set of senses for a given word. In word sense disambiguation (WSD), which is the task of selecting the intended sense of an ambiguous word token, the quality and granularity of the sense inventory greatly influences the design, evaluation, and utility of any system. The standard sense inventory, WordNet (Miller, 1998), makes no distinction between homonymy and polysemy, and is widely considered to be excessively fine-grained for many practical applications (Navigli, 2018), as evidenced by a low inter-annotator agreement (Snyder and Palmer, 2004). This has inspired substantial prior work on clustering fine-grained senses to create more coarse-grained sense inventories (Hovy et al., 2006; Navigli, 2006; Snow et al., 2007; Dandala et al., 2013; McCarthy et al., 2016).

Following the observation that different senses of a word often correspond to distinct words in another language (Resnik and Yarowsky, 1997), another branch of prior work has sought to use translations to define sense inventories (Resnik and Yarowsky, 1999; Diab and Resnik, 2002; Ng et al., 2003; Chan et al., 2007; Apidianaki, 2008; Bansal et al., 2012; Taghipour and Ng, 2015). In order to be successful, such an approach would have to resolve the challenging issues of mapping senses to translations in a set of diverse target languages, as well as projecting them onto a standard sense inventory, such as WordNet.

In summary, clustering fine-grained senses and defining sense distinctions using translations are two competing methodologies for creating coarse-grained sense inventories. Regardless of which one is adopted, an understanding of the nature and characteristics of homonymous senses is a necessary step toward a principled method of defining senses and sense distinctions. In particular, distinctions between homonymous senses must be preserved in any sense inventory. This motivates our study, which contributes to such an understanding by directly linking homonymy to the concepts of translation and sense clustering, and thus bridging the gap between the two approaches.

The contributions of this work are both theoretical and empirical. The main goal is to create theoretical foundations for the study of homonymy, which could pave the way for developing a computational method for distinguishing between homonymy and polysemy, and facilitate the task of constructing a definitive inventory of coarse-grained senses. We propose four hypotheses about the unique behavior of homonyms in the context of translations, discourses, collocations, and sense clusters. The hypotheses are formulated using established semantic concepts, and formalized in mathematical notation. Our principal hypothesis, as stated in the title, implies a sufficient condition for polysemy which is observable and replicable.

Apart from introducing the hypotheses, we perform experiments to provide empirical evidence for them. It is clear from prior work that what is true at one level of semantic granularity may not be true at another. For example, the well-known hypotheses *one sense per discourse* and *one sense per collocation* have been found not to hold consistently for WordNet senses. It is critical that all claims be formally stated and experimentally tested, regardless of whether the results are considered surprising; we have found no prior work that fulfills this requirement with respect to the four hypotheses presented in this chapter. To facilitate our experiments, we create a new annotated resource, by identifying nearly three thousand English homonyms, and mapping them onto WordNet senses. The results of our experiments on multiple annotated corpora and language pairs strongly support our hypotheses.

3.2 Homonym Hypotheses

In this section, we formally define the notion of a homonym, and formulate our hypotheses using set notation. We attempt to keep the notational complexity to a minimum, while at the same time striving to avoid ambiguity.

3.2.1 Preliminaries

Lexemes are units of language that are represented in the lexicon (Murphy and Koskela, 2010). *Words* are sets of word-forms that represent lexemes, and are associated with certain morpho-syntactic properties. This definition of words includes compounds, such as 'single out', as is the case in WordNet. We consider both lexemes and words that differ in part of speech as distinct. We write lexemes in capital letters, abstract words in single quotes, actual word-forms in italics, and sense meanings in double quotes. For example, the lexeme CUT$_v$ is represented by the verb 'cut', with the word-forms *cut, cuts,* and *cutting*. A lexeme is called polysemous if it contains multiple senses, and monosemous if it has only a single sense. Senses that belong to the same lexeme are semantically related, and therefore polysemous (Jurafsky and Martin, 2009).

A *homonymous word* (e.g., the noun 'bank' in Table 3.1) represents more than one lexeme, and those lexemes are called *homonyms*. Senses associated with distinct homonyms are unrelated and therefore homonymous (Murphy and Koskela, 2010). Consequently, the problem of deciding whether two senses of a homonymous word are polysemous is equivalent to deciding whether they belong to the same lexeme. Furthermore, since a non-homonymous word represents only a single lexeme, all of its senses are polysemous.

We are now ready to formally define homonyms. Let $\mathcal{L}$ and $\mathcal{W}$ denote the sets of lexemes and words of a given language, respectively, and let w: $\mathcal{L} \mapsto \mathcal{W}$ be a function that maps each lexeme to the word that represents it. In later sections, we will use w^{-1}: $\mathcal{W} \mapsto \mathcal{P}(\mathcal{L})$ to denote the function which maps each word to the set of lexemes it represents. We define the set of homonymous words $\mathcal{H}$ as the set of all words that represent multiple lexemes:

$$\mathcal{H} \overset{\text{def}}{=} \{W \in \mathcal{W} \mid \exists L, L' \in \mathcal{L} : \quad (L \neq L') \wedge (w(L) = w(L') = W)\}$$

For example, w(BANK$_n^1$) = w(BANK$_n^2$) = ‘bank’ $\in \mathcal{H}$.

3.2.2 One Homonym per Translation

In general, there is no simple correspondence between word senses and their translations: a single sense may be translated by any of several synonyms, and different senses of the same word may have the same translation. Ide and Wilks (2007) observe that cross-lingual distinctions often correspond to homonym-level disambiguation. We posit a direct relationship between translations and homonyms. Intuitively, if we randomly selected two different words from a bilingual dictionary, we would not expect them to have translations in common. The same reasoning applies to homonyms, since they are semantically unrelated lexemes that coincidentally share the same form. We formalize this insight as our principal hypothesis.

Put simply, the one homonym per translation hypothesis (OHPT) states that homonyms have disjoint translation sets. Formally, let $T(L)$ be a set of translations of a lexeme L, and let w^{-1} be as defined as in Section 3.2.1. Then,

$$\forall H \in \mathcal{H} : \forall L, L' \in w^{-1}(H) : \quad (L \neq L') \Rightarrow T(L) \cap T(L') = \emptyset$$

For example, the Italian translations of the noun ‘yard’ can be partitioned into two disjoint sets $T($YARD$_n^1)$ = {‘iarda’,‘yard’} and $T($YARD$_n^2)$ = {‘cortile’, ‘giardino’}, which correspond to two English homonyms, with the meanings of “unit” and “garden”, respectively.

This hypothesis implies an important generalization: *the existence of a shared translation is a sufficient condition for polysemy*. Indeed, for homonymous words, senses that can be translated by the same word must belong to the same lexeme, and so are polysemous. As all other words represent only single lexemes, all their senses are polysemous by definition (Section 3.2.1). Therefore, we consider the OHPT hypothesis as a major step towards solving the problem of distinguishing between homonymy and polysemy.

3.2.3 One Homonym per Discourse

The *one sense per discourse* (OSPD) hypothesis was introduced in the seminal paper of Gale et al. (1992a). They observe that "well-written discourses tend to avoid multiple senses of a polysemous word", and confirm that the property holds with high probability on a set of 82 instance pairs involving 9 ambiguous words. However, Krovetz (1998) reports that OSPD holds for only 67% of ambiguous words in SemCor, and conjectures that the hypothesis may only apply to homonymous senses.

We formulate Krovetz's conjecture as the *one homonym per discourse* hypothesis (OHPD), which can be viewed as a specialization of OSPD to homonyms. The hypothesis states that *all occurrences of a homonymous word in a discourse represent the same homonym*. A possible explanation of this phenomenon is that writers avoid the use of homonyms by employing their synonyms in order to reduce ambiguity in a discourse. Another explanation is that most discourses cover topics within a single domain, and therefore are unlikely to contain lexemes that are completely unrelated to each other.

Our formulation of the OHPD hypothesis states that no more than one lexeme of a homonymous word occurs in any given discourse. Formally, let D be the set of lexemes that occur in a discourse, and let w be again the function that maps lexemes to words. Then,

$$\forall L, L' \in D : \quad (w(L) = w(L')) \Rightarrow (L = L')$$

We close this section by considering the relationship between OHPD and the *one translation per discourse* (OTPD) hypothesis of Carpuat (2009). They report that approximately 80% of French words have a single English translation per document, which they interpret as strong support for their hypothesis. We note that the conjunction of our OHPT and OHPD hypotheses does not imply OTPD. Indeed, consider the example in Figure 3.1, which shows how the occurrence of three Spanish translations of the homonymous noun 'span' in two different documents leads to a violation of OTPD, but not of OHPD or OHPT.

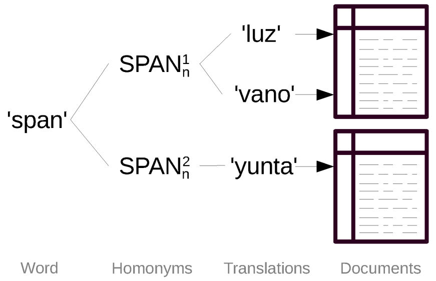

Figure 3.1: An example of an exception to the *one translation per discourse* hypothesis.The top two Spanish translations of ’span’ are synonymous.

3.2.4 One Homonym per Collocation

Yarowsky (1993) proposes the *one sense per collocation* (OSPC) hypothesis, broadly defining a collocation as “the co-occurrence of two words in some defined relationship”. Yarowsky reports that the hypothesis holds with the average 95% precision on a sample of words of an unreported size. However, Martinez and Agirre (2000) find much weaker evidence for OSPC on WordNet senses, with precision values rarely exceeding 70%.

The explicit focus of Yarowsky (1993) is on the most coarse-grained sense distinctions. Their word sample includes pseudo-words, words with different French translations, words spelled the same but pronounced differently (*homographs*), words pronounced the same but spelled differently (*homophones*), and words that are visually confusable in optical character recognition. All these types of words can be viewed as approximations of homonymy, as they involve pairs of distinct lexemes. We formalize this notion with the *one homonym per collocation* (OHPC) hypothesis, which states that only one homonym of a word should appear in any given collocation.

Formally, given a corpus of text, let $\mathcal{R}$ be the set of all collocations. For lexeme $L \in \mathcal{L}$, and collocation $r \in \mathcal{R}$, let $C_r(L)$ be a proposition which is true if and only if $w(L)$ occurs in collocation r in the corpus. Then,

$$\forall H \in \mathcal{H} : \forall L, L' \in w^{-1}(H) : \forall r \in \mathcal{R} : \quad (C_r(L) \land C_r(L')) \Rightarrow (L = L')$$

For example, if BANK^1_n (“repository”) is found to occur in the collocation [word-to-right = *hired*] then BANK$_n^2$ ("ridge") is unlikely to occur in this collocation.

3.2.5 One Homonym per Sense Cluster

Sense clustering is the task of grouping together senses that are closely related (Dandala et al., 2013). Although the criteria for eliminating sense distinctions vary depending on the purpose of the sense inventory, a common motivation is to reduce the excessive granularity of WordNet (Snow et al., 2007). In particular, a manual clustering of WordNet senses was created as part of the OntoNotes project, with the objective of increasing the inter-annotator agreement on WSD to 90% (Hovy et al., 2006). Sense clustering has been shown to improve performance on a number of NLP tasks (Pilehvar et al., 2017), and can serve as an extrinsic evaluation for learned representations of senses (Mancini et al., 2017).

Since homonyms are distinct lexemes, we posit that any well-grounded clustering approach must avoid merging homonymous senses. Formally, let $\mathcal{C}$ be a sense clustering, a set of disjoint sets of senses, and let $S(L)$ be the set of senses of lexeme L. Then,

$$\forall C \in \mathcal{C} : \exists L \in \mathcal{L} : C \subseteq S(L)$$

In plain words, while the senses of a homonym may be divided between multiple clusters, no cluster should contain senses from different homonyms.

3.3 Homonym Data

In order to provide experimental evidence for our homonym hypotheses, we need a large set of "gold" homonyms, as well as a mapping between those homonyms and the sense annotations in existing corpora. Since no such resource is publicly available, we create our own collection of English homonyms (see Table 3.2). In this section, we present a binary typology of homonyms, our methodology for creating a list of homonyms, and the method for mapping those homonyms onto the WordNet sense inventory.

3.3.1 Typology of Homonyms

There are generally two ways of defining homonyms. In linguistics (and in this chapter), homonyms are considered to be distinct lexemes that happen to share the same form (Murphy and Koskela, 2010). In lexicography, homonymy is sometimes defined more narrowly, by additionally requiring the etymological origins of the lexemes to be different (Stevenson, 2010). Homonyms can therefore be divided into two types: those that satisfy the requirement of different origins, and those that do not. Due to the lack of commonly-accepted terminology, we refer to these two types of homonyms simply as Type-A and Type-B, respectively.

The two types of homonyms, which are schematically illustrated in Figure 3.2, stem from different diachronic phenomena. Type-A homonyms arise from a convergence of distinct words into a single form. This can occur through the process of sound change or inter-lingual borrowing. For example, both the Old English word *cæg* "locking implement" and the 17th-century Spanish borrowing *cayo* "island" evolved into the modern English *key*. Type-B homonyms, on the other hand, arise when a single lexeme splits into two lexemes due to the process of semantic drift. For example, the two meanings of *staff*, "pole" and "people", have developed from a single etymon, which is attested in Old English as *stæf*. Importantly, as native speakers are generally unaware of the etymological history of words, these two types of homonyms are indistinguishable in the synchronic analysis of languages (Lyons, 1995).

The crucial methodological advantage of Type-A homonyms is that they can be objectively identified by consulting existing etymological dictionaries. Even though the process of compiling an exhaustive list of Type-A homonyms for any language is time-consuming, it is still much easier and less controversial than conducting psychological experiments with human subjects Brown (2008), or obtaining consensus within teams of linguistic experts (Weischedel et al., 2013). We have accomplished this task for English by creating a homonym resource that we describe next.

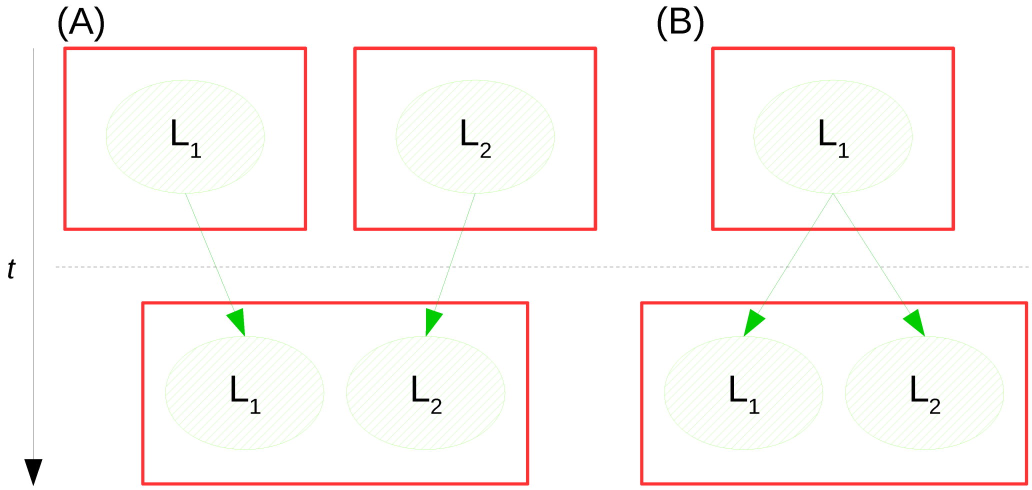

Figure 3.2: A schematic illustration of the diachronic distinction between two types of homonyms. Circles represent lexemes; boxes represent words.

POS	Origin	Gloss	French
N,V	Old French *espan*	distance	*portée*
N,V	Low German *spannen*	rope	*filin*
Adj	Old Norse *spán-nýr*	clean	*impeccable*
V	Old English *spinnan*	rotate	*tourné*

Table 3.2: Sample entries of the homonym resource, which correspond to six homonyms of the English lemma *span*.

3.3.2 List of Type-A Homonyms

The new homonym resource[2], which enables us to empirically test our homonym hypotheses, contains words that represent multiple lexemes with distinct etymological origins. We compiled the list by collecting all homonyms that we could find in dictionaries, including the English Oxford Living Dictionary[3] and the Concise Oxford Dictionary of English Etymology[4]. We include all homonyms that at some point during language evolution existed as separate words, even those that can be traced to a single proto-word. For example, we include the homonyms of the noun *sole* (“undersurface” vs. “fish”) because of their distinct histories, even though both ultimately come from Latin *solea* “sandal”.

[2] *https://webdocs.cs.ualberta.ca/˜kondrak*
[3] *https://en.oxforddictionaries.com*
[4] *http://www.oxfordreference.com*

Table 3.2 shows sample entries from our resource. The list contains 2759 Type-A homonyms that correspond to 804 lemmas, 1601 unique lemma/POS pairs, and 1967 distinct etymologies. The number of distinct etymologies per lemma ranges from two to six. Each entry includes etymological information (the form and the language of origin), and a list of possible parts of speech (noun, verb, adjective, adverb). For the purpose of disambiguation in subsequent stages of annotation, each entry was manually assigned a brief English gloss, as well as a single French translation. We excluded from our list all proper nouns and abbreviations.

About two dozen of the homonymous words in our resource represent homographs, which are homonyms that differ in pronunciation. For example, the noun 'bass' is pronounced [bæs] or [bes] depending on whether it refers to a fish or a musical instrument, respectively. Although most of the dictionary words with alternative pronunciations appear to involve Type-A homonyms, we found a number of exceptions. They include Type-B homonyms (e.g. 'pension'), polysemous words (e.g. 'undertaking'), common vs. proper nouns (e.g. 'job'), matching word-forms of distinct lemmas (e.g. 'putter'), as well as pronunciation variants (e.g. 'puissance'). Since our focus is on written language, our resource excludes homophones, such as 'cellar' vs. 'seller'.

Although we make no claim about the completeness of our homonym resource, we consider it to be representative of English homonyms in general. This is based on the fact that Type-A and Type-B homonyms cannot be distinguished without access to etymological expertise.

3.3.3 Mapping WordNet Senses to Homonyms

In order to test our homonym hypotheses, we must be able to convert the existing word sense annotations into homonym annotations. For example, we need to know which homonym from our list is represented by a word token *spans* which is sense-annotated as "two items of the same kind" in some corpus. The standard sense inventory for WSD is WordNet. In this section, we describe our method of mapping the homonyms in our new resource to WordNet senses.

Because of the large number of fine-grained senses in WordNet, it was not practical to directly map each WordNet sense of each homonymous word to the corresponding homonym. Instead, we made use of the existing clustering (Navigli, 2006), which was created by automatically mapping WordNet 2.1 senses to more coarse-grained senses defined by the Oxford Dictionary of English (ODE). Our 804 homonymous lemmas correspond to 2644 sense clusters, which contain 5361 senses. We manually mapped each cluster of senses to a single homonym on the basis of their WordNet sense glosses.

The resulting mapping is imperfect for two reasons. First, the ODE clustering itself is not always correct, which sometimes results in homonymous senses being placed in the same cluster. Second, our human annotator made some errors in mapping clusters to homonyms. We performed the following validation experiment in order to estimate the accuracy of the overall mapping. A second annotator performed a direct mapping of 268 WordNet senses corresponding to a random sample of 77 homonymous words, without any reference to the ODE clustering. We found that the two independent mappings of the 268 senses differed in only 17 instances, which implies that the overall error rate has an upper bound of 6%.

The errors in the sense-to-homonym mapping are a source of "false alarms" in the experiments described in Section 3.4. We are confident in our ability to determine which of the apparent exceptions are actual exceptions to our hypotheses by careful analysis of the available data. While the distinction between homonymy and polysemy can be highly subjective, the mapping of WordNet senses to known homonyms is much easier, as confirmed by our validation experiment described above.

3.4 Homonym Evidence

In this section, we describe the experiments that test the four hypotheses formulated in Section 3.2 using the full set of homonyms in our new homonym resource from Section 3.3.

3.4.1 SemCor and Translations

For testing the OHPD and OHPC hypotheses, we use SemCor (Miller et al., 1993), a large sense-annotated English corpus which was created as part of the WordNet project (Petrolito and Bond, 2014). In particular, we adapt the version of SemCor from Raganato et al. (2017).[5] The number of word tokens, types, and senses are in Table 3.3 (words are defined as lemma/POS pairs)

For testing the OHPT hypothesis, we require not only sense annotations, but also the corresponding translations. At the minimum, we need a large word-aligned bitext that has both sense and part-of-speech annotations on the source side, and lemma annotations on both sides. In addition, the sense inventory has to be the same as the one in our homonym resource. Although such resources are rare, we managed to adapt two bitexts to meet these requirements: MultiSemCor (Bentivogli and Pianta, 2005), and JSemCor[6] (Bond et al., 2012). These corpora, which we refer to as MSC and JSC, contain partial word-aligned translations of SemCor into Italian and Japanese, respectively.

3.4.2 WordNet

The use of WordNet presents a number of technical challenges. For the purpose of replicability, we describe here two major issues.

The first issue concerns two distinct conventions for referring to individual WordNet senses: *sense keys* (used in SemCor, JSC, and the ODE clustering) and *sense numbers* (used in MSC and OntoNotes). We converted the former into the latter using the WordNet::SenseKey package.[7] Because the mapping is not always one-to-one, 16 out of 60,655 WordNet senses in the ODE clustering had to be excluded; however, none of the affected words are in our homonym resource.

The second issue is the mapping between different WordNet versions. We converted the sense keys from WordNet 2.1 – the version of WordNet used in the clustering described in Navigli (2006) – to WordNet 3.0 – the version used by all other

[5] *http://lcl.uniroma1.it/wsdeval*

[6] Experiments on JSemCor were performed by Yixing Luan, a native Japanese speaker.

[7] *https://metacpan.org/release/LINAS/WordNet-SenseKey-1.03*

	SemCor	MSC	JSC
Word tokens	226,034	92,992	58,257
Word types	20,399	11,451	8,445
WordNet senses	33,308	17,875	12,516

Table 3.3: The size of the English side of each corpus.

Hypothesis	Focus	Corpus	Instances	Exceptions		Support
				Apparent	Actual	(in %)
OHPT	translations	MSC	1093	7	1	99.9
OHPT	translations	JSC	1093	3	2	99.8
OHPD	documents	SemCor	2126	14	9	99.6
OHPC	collocations	SemCor	522	16	11	97.9[9]
OHPSC	sense clusters	OntoNotes	1578	23	2	99.9

Table 3.4: Summary of the evidence for the homonym hypotheses from our five experiments.

resources in this chapter – using WordNetMapper.[8] The package failed to map 551 out of 60,655 senses in the ODE clustering, which resulted in 22 WordNet senses being excluded from our homonym resource. Due to these issues, we decided not to further map all WordNet senses in our resources to WordNet 3.1.

3.4.3 One Homonym per Translation

The OHPT hypothesis characterizes the relationship between homonymous words and their translations in another language. We validate the hypothesis on two language pairs using the annotated bitexts described in Section 3.4.1.

In the experimental evaluation, we compute the percentage of type-level instances that are consistent with the OHPT hypothesis. For each English word (i.e. lemma/POS pair) that appears in our homonym resource, we identify the set of its translations on the target side of the bitext. Each unique word/translation pair constitutes a single instance. An instance is consistent with the OHPT hypothesis if and only if all of its occurrences in the bitext represent the same homonym. For example, the Italian translation ‘gioco’ corresponds to three different senses of the noun ‘game’ in MSC, but since all of them belong to the same homonym, this instance is consistent with OHPT.

[8] *https://github.com/cltl/WordNetMapper*

The results of the evaluation on the MSC and JSC bitexts are shown in Rows 1 and 2 of Table 3.4. Coincidentally, MSC and JSC have the same number of unique word/translation pairs (1093). The two corpora contain only 3 actual exceptions to OHPT. The single actual exception in MSC involves the homonyms represented by the noun 'band' which is often translated in Italian as 'banda'. In this case, the homonymy in English ("ring" vs. "group") is mirrored by an analogous case of homonymy in Italian. The two actual exceptions in JSC involve the English lexical loans 'case' and 'club', which have the same Katakana written form regardless of the homonym they represent. We attribute these exceptions to the phenomenon of *parallel homonymy*, which may arise in the process of lexical borrowing.

In addition to the 3 actual exceptions, the experiment identified 7 exceptions that are caused by data errors in the two corpora. The data errors can be divided into four categories: (1) incorrect sense annotations in SemCor, e.g. *"the* <u>*case*</u> *of Jupiter"* annotated with the sense of "container"; (2) an incorrect sense translation in MSC: *flag* in the sense of "flower" translated as *bandiera* instead of *iride*; (3) errors in the ODE clustering, e.g. two homonymous senses of 'club' ("team" and "playing card") in the same cluster; (4) an error in our manual mapping between the ODE clustering and the homonyms: 'light' in the sense of "free from troubles" being mapped to the homonym "not dark". We conclude that the OHPT hypothesis is supported in over 99.8% of instances in either bitext.

In order to verify that partitioning of translations is a property of homonyms, and not simply of any sense clusters, we perform an additional experiment on MSC. We randomly select two sets of 20 words (i.e. lemma/POS pairs) from our homonym resource and the OntoNotes clusters, respectively. We consider only words that are represented in MSC by senses from exactly two homonyms or two OntoNotes sense clusters. None of the OntoNotes words occur in our homonym resource. This yields 40 words with a similar number of sense-annotated tokens: 6.80 per homonym, and 7.25 per OntoNotes cluster, on average. We find that 16 of the 20 homonym pairs, and 6 of the 20 OntoNotes cluster pairs exhibit strict translation partitioning in MSC. In total, there are 4 instances of overlapping translations between 4 homonym pairs (a subset of the 7 apparent exceptions in Table 3.4), and 17 such instances between 14 OntoNotes cluster pairs (3 cluster pairs share multiple translations). This result is statistically significant ($p < 0.005$) according to the χ^2 test. We conclude that homonyms are significantly more likely to exhibit translation partitioning than OntoNotes sense clusters.

3.4.4 One Homonym per Discourse

The OHPD hypothesis predicts that all tokens of a given homonymous word in a discourse correspond to the same homonym. We validate the hypothesis on English SemCor (Section 3.4.1), taking each of its documents as a single discourse.

In the experimental evaluation, we compute the percentage of type-level instances that are consistent with the OHPD hypothesis. For each English word (i.e. lemma/POS pair) that appears in our homonym resource, we identify all its occurrences in the corpus. Each unique word/document pair constitutes a single instance. An instance is consistent with the OHPD hypothesis if and only if all of the occurrences of the word in the document represent the same homonym.

When a homonymous word occurs only once in a document, there is of course no possibility of an actual OHPD violation. However, we consider those instances to support the hypothesis as well, because the writer may have chosen to replace a homonym with one of its synonyms in order to avoid potential ambiguity.

The results of the evaluation are shown in Row 3 of Table 3.4. SemCor is divided into 352 documents, with an average of 642 sense-annotated open-class words per document. A careful analysis of the 14 apparent exceptions reveals that four of them are caused by sense annotation errors in SemCor (e.g., *sharp* <u>*bow*</u> *of a skiff* is annotated as "weapon for shooting arrows"), and one results from an error in the ODE clustering. The 9 actual exceptions involve the homonymous nouns 'bank', 'lead', 'list', 'port', 'rest', and 'yard', as well as the verb 'lie'. We conclude that fewer than 0.5% of instances in SemCor contradict the OHPD hypothesis.

[9] This number is a lower bound estimate.

3.4.5 One Homonym per Collocation

The OHPC hypothesis predicts that only one homonym of a word appears in any given collocation. Due to the broad definition, wide variety, and large number of possible collocations, it is difficult to definitively establish the extent to which the OHPC hypothesis holds for a given corpus. Instead, we follow the methodology of Yarowsky (1993) and Martinez and Agirre (2000), who test the OSPC hypothesis by analyzing the performance of a supervised WSD system in which each feature corresponds to a distinct type of a collocation. The rationale is that the accuracy of the WSD system indicates the level of support for the hypothesis in the training corpus.

For the experimental evaluation, we adopt the IMS system of Zhong and Ng (2010). IMS learns a separate classification model for each ambiguous word in the training data, with each class corresponding to one sense of the word. The system employs three types of features, which broadly correspond to different kinds of collocations: (1) the presence of specific content words in specific positions relative to the focus word; (2) the set of POS tags in the context of the focus word; (3) the presence of specific content words in the *bag-of-words* context of the focus word. We train IMS on English SemCor, and test on the concatenation of five benchmark datasets of Raganato et al. (2017).

The results of the experiment strongly support the OHPC hypothesis. The test set contains 528 occurrences of words from our homonym resource. Six of those words, each appearing in one instance, are not attested at all in SemCor. IMS selects a sense of the correct homonym in 506 out of the remaining 522 instances. Of the 16 classification mistakes, three are attributable to errors in the ODE clustering, and two are due to the WordNet mapping issues described in Section 3.4.2. Thus, the effective accuracy of IMS on the homonymous words in the test set is 97.9%.

Analysis of the remaining 11 errors made by IMS shows that their principal cause is insufficient training data. For example, the noun ‘match’ in the sense of “piece of wood” occurs only once in the entire SemCor corpus, which prevents IMS from reliably recognizing this sense. Other obvious mistakes, such as *“follow the lead”* misclassified as “metal,” are explained by the lack of training examples involving the collocations that occur in the test set. We conclude that the IMS accuracy on the test set should be interpreted as a lower bound for the applicability of OHPC.

3.4.6 One Homonym per Sense Cluster

We test our fourth hypothesis, OHPSC, by searching an existing resource for clusters that contain senses from distinct homonyms. We cannot perform this experiment on the ODE clustering because we use it to derive our mapping from WordNet senses to homonyms (Section 3.3.3). Instead, we run it on the high-quality, handcrafted OntoNotes clustering[10], which previously used as a gold standard by Snow et al. (2007). The clustering includes 439 of the 1601 lemma/POS pairs that are listed in our homonym resource. Those words involve 2467 WordNet senses that are grouped into 1578 clusters, of which 1555 (98.5%) are found to contain no homonymous senses, as our hypothesis predicts.

We manually analyze the 23 clusters that appear to combine senses from distinct homonyms. The vast majority (21) of these apparent exceptions are artifacts of errors in the ODE clustering. The errors are easy to spot by native speakers because senses within a single cluster clearly correspond to distinct coarse-grained senses in ODE. In the remaining two cases, OntoNotes clusters two pairs of homonymous senses: (1) the noun ‘tap’ as “the sound made by a gentle blow” and “a faucet for drawing water,” and (2) the verb ‘pose’ as “introduce” and “be a mystery to.” Even though we find these two clustering decisions somewhat debatable, we treat them as actual exceptions to our hypothesis. We conclude that the OHPSC hypothesis is corroborated in over 99.8% of the OntoNotes clusters.

3.5 Conclusion

We have investigated the concept of homonymy, formulating four hypotheses that follow a common pattern. Taken together, our hypotheses suggest that, figuratively speaking, homonyms seem to repel each other, like particles with the same electric

[10] *https://catalog.ldc.upenn.edu/LDC2013T19*

charge. The experiments performed using our new resource confirm that distinct homonyms are rarely observed in connection with a single translation, discourse, collocation, or sense cluster. In addition, they demonstrate that contraventions of the empirical predictions made by our theory more often than not identify errors in existing resources.

We envisage several directions for building upon the theoretical basis established in this chapter. In order to extend our homonym resource, we plan to develop an operational method for identifying Type-B homonyms on the basis of translation sets involving multiple languages. We anticipate that translations extracted from parallel corpora will facilitate the creation of high-quality coarse-grained sense inventories via sense clustering. As a step towards this goal, we will investigate the problem of automated mapping between senses and translations.

3.6 Addendum: Distinguishing Between Homonymy and Polysemy

While the preceding chapter discusses the theoretical properties of homonymous senses, it leaves for future work the task of automatically distinguishing between homonymy and polysemy. Following the publication of the material in this chapter, such work was published by Habibi, Hauer, and Kondrak (2021). They developed and tested methods for classifying a given word as homonymous – i.e. having two semantically unrelated senses, as defined in Section 3.2.1 – or not. I was a major contributor to, and second author of, this publication, and so I consider it pertinent to summarize the findings of that research here[11].

This work seeks to improve upon prior methods for homonymy classification and related tasks. In particular, Dyvik (2004) leverage parallel corpora to evaluate the relatedness of word senses. Utt and Padó (2011) model homonymy as a continuous phenomenon, and grade the degree of homonymy a word exhibits using a statistical model. van den Beukel and Aroyo (2018) apply WordNet-based similarity metrics, while Beekhuizen et al. (2018) use vector representations of words and

[11] The first author, Amir Ahmad Habibi, performed the experiments in this paper.

contexts.

All six of the methods proposed by Habibi et al. (2021) utilize a graph model, with senses as vertices and edges representing semantic relatedness between senses. They differ in the criteria for considering two vertices to be semantically related. Of particular interest here is their "Two Senses, One Translation" method, which is explicitly described as applying OHPT to homonymy detection. This approach considers two senses to be semantically related if they have a translation in common.

Significantly, in support of this method, they state and prove a theorem which "generalizes the OHPT hypothesis to account for the few exceptions found [in Section 3.4]". The theorem states that, if senses x_1 and x_2 of word x can both be translated by word y, then exactly one of the following holds: (1) x_1 and x_2 are semantically related; (2) x and y exhibit parallel homonymy. Their proof of this theorem uses the theoretical properties of synsets formulated in Chapter 2. Given the results in Section 3.4.3 which show that parallel homonymy is very rare in practice, this theorem and its proof imply, as Habibi et al point out, that the high reliability of OHPT follows from our theory of sense, synonymy, and translation set out in Chapter 2.

The authors evaluate their homonym classification methods, including the OHPT-based method, on a dataset of English words. This dataset was constructed using a manually expanded and corrected version of the homonym database discussed in Section 3.3. They find that OHPT sets a new state of the art for homonym detection, outperforming all other methods tested, including that of van den Beukel and Aroyo (2018), and therefore claim a new state-of-the-art result.

In sum, this paper not only provides a theoretical justification for OHPT, but also demonstrates the utility of our work for an interesting lexico-semantic task. They also continue the work undertaken earlier in this chapter, working to complete and correct the homonym resource. Taken together, the work described in this chapter represents a strong demonstration of how linguistic phenomena, theoretical arguments, and empirical hypothesis testing can be combined to achieve new results in lexical semantics.

Chapter 4

One Sense per Translation

Word sense disambiguation (WSD) is the task of determining the sense of a word in context. Translations have been used in WSD as a source of knowledge, and even as a means of delimiting word senses. In this chapter, we define three theoretical properties of the relationship between senses and translations, and argue that they constitute necessary conditions for using translations as sense inventories. The key property of One Sense per Translation (OSPT) provides a foundation for a translation-based WSD method. The results of an intrinsic evaluation experiment indicate that our method achieves a precision of approximately 93% compared to manual corpus annotations. Our extrinsic evaluation experiments demonstrate WSD improvements of up to 4.6% F1-score on difficult WSD datasets.[1]

4.1 Introduction

Word sense disambiguation (WSD) is the task of classifying a word in context according to its sense. For example, given the context "the field was covered in green grass," a WSD system would need to classify *field* as having its "flat open land" sense, rather than its "area of study" sense. Throughout its history, WSD has been associated with translation (Weaver, 1949), as it is understood that different senses of a word may translate differently. For instance, in the above example, *field* could be translated into French as *champ*, but not as *domaine* (the latter could, however, translate the "area of study" sense of *field*). In this chapter, we address the open

[1] This chapter is based on Hauer and Kondrak (2021). See the preface for details.

question, *to what extent can a translation-based method improve modern WSD?*

This question is surely an important one: WSD remains an active area of research (Blevins and Zettlemoyer, 2020; Barba et al., 2021a; Barba et al., 2021c), but despite the rapid improvements brought on by transformer-based (Vaswani et al., 2017) language models such as BERT (Devlin et al., 2019), substantial room for improvement remains (Maru et al., 2022). WSD has been used as a benchmark to compare and analyze transformer-based language models (Loureiro et al., 2021). It has also been shown to have applications to tasks such as translation (Liu et al., 2018), semantic parsing (Martínez Lorenzo et al., 2022), and metaphor detection (Maudslay and Teufel, 2022). New variants of the task are still being proposed, such as visual WSD, in which candidate senses are represented by images (Raganato et al., 2023). Clearly, the ability to map a word in context to an entry in a discrete lexical knowledge base remains relevant in natural language processing, for both human end users and downstream tasks.

Incorporation of translation information has been shown to be useful for both classic (Dagan et al., 1991) and modern (Luan et al., 2020) WSD methods. Despite such proof-of-concept works, current state-of-the-art WSD methods do not explicitly leverage translation, leaving a potential source of knowledge untapped. It is therefore of interest to the lexical semantics community to investigate the extent to which senses and translations correspond, and how this correspondence can be leveraged in practice.

Our investigation has the following structure: (1) We begin by clearly defining the theoretically "ideal" mapping between senses and translations. (2) We show that such mappings are rare in practice, even between unrelated languages, offering an explanation as to why translation-based WSD methods became less common as the field developed. (3) We posit that it is possible to improve supervised WSD performance by leveraging instances where the translation of a word *does* determine its sense. (4) We propose and evaluate a translation-based disambiguation method to test this hypothesis. (5) We discuss the relationship between various theoretical properties and synonymy and polysemy.

Our empirical results strongly support our hypothesis. A large-scale intrinsic evaluation of our method using existing lexical knowledge bases shows that it achieves very high precision. Our extrinsic evaluation shows that synthetic training data produced by our method, when used to train a supervised model, can yield improvements in F1-score of up to 4.6% on difficult WSD benchmark datasets. We conclude that the explicit incorporation of contextual translations has great potential to improve WSD research, and lexical semantics research in general.

The principal goal of this chapter is the examination of the sense-translation connection from both theoretical and empirical perspectives in a modern context. Thus our contributions are twofold: a theoretical analysis of the relationship between senses and translation, supported by empirical analysis; and a method for efficient, unsupervised, large-scale semantic annotation via translations, which yields substantial WSD improvements.

4.2 Related Work

The use of translations as a source of information about word senses rose to prominence in the 1990s, supported by the increasing availability of machine-readable multilingual resources. Brown et al. (1991) and Dagan et al. (1991) developed statistical approaches to WSD, with the former presenting a direct application to statistical machine translation. Gale et al. (1992b) were the first to explicitly define WSD in terms of identifying the correct translation: they identify a set of six English words, each with two senses, with a one-to-one mapping between those senses and their French translations. This paradigm of translation-informed WSD influenced the landmark WSD works of Yarowsky (1995) and Schütze (1998), among others. By the late 1990s, translation was so prevalent in the WSD literature that Resnik and Yarowsky (1997) explicitly proposed "to restrict a word sense inventory to those distinctions that are typically lexicalized cross-linguistically."

Interest in translation in the WSD literature continued throughout the 2000s (Ide, 2000; Chan et al., 2007; Apidianaki, 2008), culminating in two SemEval-2010 shared tasks: cross-lingual lexical substitution (Mihalcea et al., 2010), and cross-lingual WSD (Lefever and Hoste, 2010). The former can be viewed as the task of finding translations for a word in a given context. In the latter, translations from word-aligned parallel corpora were used to create a "multilingual sense inventory". The dataset was limited to small lexical samples, and involved substantial manual-annotation effort for each tested language pair. Neither the exact annotation criteria nor the datasets themselves are available. The successes and difficulties of this task motivate further research into the use of translations as sense inventories.

Yao et al. (2012) observed that prior work made conflicting assumptions about the correspondence between senses and translations. They consider the case where a single word e in a parallel corpus is aligned, in different contexts, with two different words, f_1 and f_2, in another language. They point out that some prior works, such as Lefever et al. (2011), assume that e is polysemous, with f_1 and f_2 translating distinct senses of e, while others, such as Bannard and Callison-Burch (2005), instead assume that f_1 and f_2 translate a single sense of e, and so are synonymous. Our work builds upon this observation, analyzing the various possible relations between senses and translations in greater detail, and leveraging them them to improve WSD.

Despite the early successes of translation-based WSD, methods based on monolingual resources, namely WordNet (Miller et al., 1990) and SemCor (Miller et al., 1993), became prominent in the 2010s. *It Makes Sense* (Zhong and Ng, 2010), a supervised WSD system based entirely on monolingual contextual features, remained state-of-the-art for most of the decade (Papandrea et al., 2017) before being replaced by methods based on contextual embeddings (Hadiwinoto et al., 2019). In the early 2020s, WSD systems leveraging increasingly sophisticated pre-trained language models approached and finally exceeded 80% accuracy on standard WSD datasets (Blevins and Zettlemoyer, 2020; Barba et al., 2021a; Barba et al., 2021c). In response to these advances, Maru et al. (2022) proposed to focus on more difficult WSD instances, such as those involving rare senses, or on which modern WSD systems tend to make errors. We support this proposal, and make use of their "challenge" datasets in our experiments.

4.3 Mapping Senses and Translations

While the use of translation information to identify or even *define* word senses was frequent in early WSD research, today it primarily serves as supplementary data, rather than as the core of the method (Luan et al., 2020). In this section, we lay the theoretical groundwork for explaining this paradigm shift; an empirical analysis follows in the next section.

Given an ideal one-to-one mapping between senses of a word and its lexical translations, each sense could be unambiguously defined by a distinct translation, and each translation would indicate a different sense. Figure 4.1 shows a graphical representation of a sense-translation mapping which does *not* conform to this ideal, with three Italian translations of the English noun *wood*. An edge between a sense and a translation indicates that the former can be translated by the latter. As the sense-translation mapping is not bijective, we cannot use translation knowledge alone to determine the sense of an instance of *wood*.

We can analyze the theoretical properties of such a mapping in terms of three word-level binary predicates, which are defined on a given source word e and language of translation F. Each of these predicates is a necessary condition for such an ideal mapping to exist. Moreover, in conjunction, they represent a sufficient condition for using a word's translations as a sense inventory. The three sense-translation mapping predicates are discussed in the following subsections.

4.3.1 One Sense per Translation (OSPT)

One Sense per Translation (OSPT) is the key predicate for translation-based WSD, as it facilitates the inference of a word's sense from its translation. OSPT underlies the method that we propose in Section 4.5.

OSPT(e, F) := "all senses of the word e have disjoint sets of lexical translations in language F"

If OSPT holds, each translation of e corresponds to exactly one sense, and so we can use the sense-translation mapping to perform WSD. Exceptions to OSPT occur when words from different languages share multiple senses, a phenomenon which

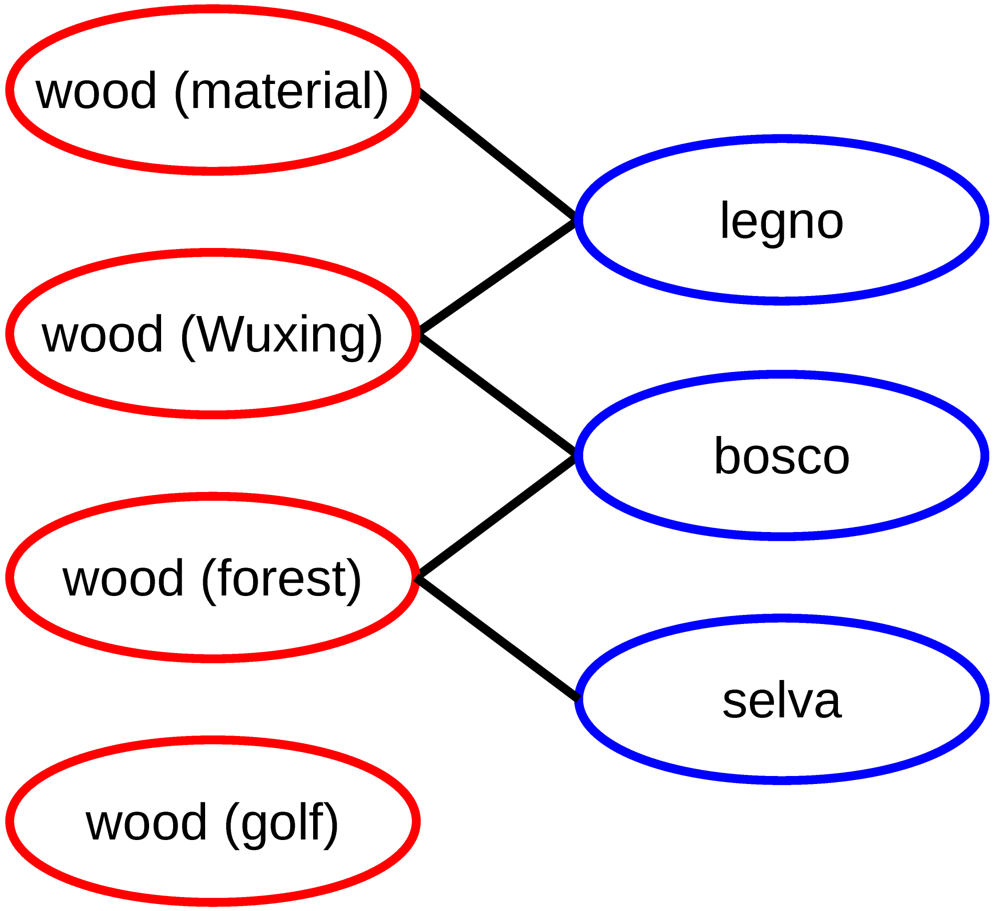

Figure 4.1: An example mapping between senses and translations. Each translation corresponds to at least one sense.

we refer to as *parallel polysemy*. For OSPT to hold, the source word cannot exhibit parallel polysemy with any of its translations. For example, Figure 4.1 shows a violation of OSPT, as the Italian word *legno* maps to two distinct senses of the English word *wood*. Therefore, the sense of *wood* in a given sentence cannot be inferred solely from the fact that it is translated as *legno*. On the other hand, if an instance of *wood* is translated into Italian as *selva*, we can infer that it is used in its “forest” sense.

4.3.2 One Translation per Sense (OTPS)

The One Translation per Sense (OTPS) predicate can be viewed as a dual of OSPT, reversing the roles of senses and translations.

OTPS(e, F) := "each sense of the word e has at most one lexical translation in language F"

In other words, no pair of translations translate the same sense. Exceptions to OTPS are instances of synonymy between translations of a given source word.[2] For example, the "forest" sense in Figure 4.1 maps to two distinct translations, *bosco* and *selva*, violating OTPS. This presents a challenge to the proposal to use translations as sense inventories (Resnik and Yarowsky, 1997) by creating cases where instances of a word need not be distinguished by their translations. Moreover, this also poses a problem for aligning sense distinctions with translation distinctions (Lefever and Hoste, 2010), as *bosco* and *selva* must somehow be "clustered" to avoid identifying instances of *wood* with these translations as being semantically distinct. Note, however, that unlike violations of OSPT, translations that cause OTPS violations can still be used to disambiguate the translated word in some cases.

4.3.3 No Lexical Gaps (NoLG)

The No Lexical Gaps (NoLG) predicate reflects the importance of *lexical gaps* (Bentivogli and Pianta, 2000) in multilingual semantics.

NoLG(e, F) := "each sense of the word e has at least one translation in language F"

Since it is not practical to enumerate all possible phrasal translations of each sense, such lexical gaps generally preclude translation-based WSD: we cannot identify a sense based on its lexical translation if it *doesn't have* a lexical translation. For example, the "golf" sense of *wood* in Figure 4.1 corresponds to a lexical gap, and so would need to be translated into Italian by a compositional phrase, such as "legno da golf".

[2] Interestingly, the WSD algorithm of Diab and Resnik (2002), which disambiguates English words based on their French translations, is based on the assumption that all target-language words are monosemous.

In summary, an ideal one-to-one sense-translation mapping seems to be a very brittle structure. Any exception to OSPT, OTPS, or NoLG would complicate the use of translations to define sense inventories. Moreover, any exception to OSPT or NoLG will outright preclude the use of translations alone for WSD. The viability of translation-informed WSD therefore rests on the extent to which these properties hold in practice, which we investigate in the next section.

4.4 Empirical Analysis

We focus on English, with three languages of translation which represent various degrees of relatedness to English: Italian, Polish, and Chinese, For each language, we compute the proportion of English words for which OSPT, OTPS, and NoLG hold in BabelNet (Navigli and Ponzetto, 2012), a large lexical knowledge base frequently used as a sense inventory for multilingual WSD (Pasini et al., 2021). We consider only English words with at least two senses in WordNet 3.0 (BabelNet inherits senses from WordNet), and at least one translation in the target language in BabelNet 4.0. There are 20,426 such words with Italian as the target language, 17,404 for Polish, and 19,973 for Chinese.

Table 4.1 summarizes the results. The NoLG values indicate that the majority of English words involve at least one lexical gap in any of the three languages of translation. The OTPS row shows that even fewer words have no more than one translation per sense. The OSPT property is more reliable, covering almost 60% words with Italian as the language of translation, and approaching 80% with less related languages such as Polish or Chinese. However, the last row in the table demonstrates that only a very small percentage of English words satisfy all three properties at the same time.

Since we have argued that the conjunction of the three properties is a necessary condition for an ideal one-to-one sense-translation mapping, these empirical results provide an explanation why using translations as sense inventories is infeasible in practice. Furthermore, even if we had a sense inventory with a complete mapping between senses and translations (something BabelNet and comparable resources

	Italian	Polish	Chinese
OSPT	59.5	77.4	75.7
OTPS	16.3	22.8	10.3
NoLG	47.4	38.3	40.3
ALL	1.9	2.6	1.5

Table 4.1: The percentage of English polysemous words in BabelNet which exhibit each of the three sense-translation mapping properties with respect to three languages of translation.

aspire to provide), the OSPT values in our results table indicate that a substantial portion of words cannot be disambiguated on the basis of their translations alone. We conclude that this was a key factor in the abandonment of the use of translations to induce sense inventories, or perform WSD on all words. Nevertheless, we posit that translations *can* be leveraged to improve WSD, specifically be exploiting those cases where a translation of a word in context uniquely determines its sense. In the next section, we present and apply a method for using translations to tag a subset of the tokens in a parallel corpus.

4.5 Corpus Tagging with OSPT

Although the results in Section 4.4 demonstrate that translations alone are not sufficient for all-words WSD, prior work such as Gale et al. (1992b) and Lefever and Hoste (2010) have shown that they can still be applicable to lexical samples. In this section, we explore the idea of using translations to improve WSD on modern standard datasets. Specifically, we leverage those cases where the translation of a word corresponds to exactly one of its senses in order to create supplementary training data for a supervised WSD system.

4.5.1 Corpus Tagging

The generation of "silver datasets" for WSD is a way to address the knowledge acquisition bottleneck (Pasini, 2021), the difficulty of obtaining training data for supervised WSD. To this end, the goal of semantic corpus tagging is not to disambiguate all word tokens, or any particular subset of lemmas; rather, the goal is to partially sense-annotate a corpus to produce supplementary training data for a supervised WSD system.

Automatic sense tagging has been a popular area of research in lexical semantics. Taghipour and Ng (2015) used a mapping of Chinese translations to English senses to annotate the English side of an English-Chinese parallel corpus; however, this mapping is not available. Pasini and Navigli (2017) sense-tag Wikipedia articles using a variant of the personalized page-rank algorithm (PPR), while Delli Bovi et al. (2017) applies a similar approach to the EuroParl parallel corpus. Barba et al. (2020) use a pre-trained language model to identify semantically-equivalent translations of manually sense-annotated tokens. Most recently, Hauer et al. (2021b) propose a family of pipeline approaches employing WSD methods, machine translation, lexical resources, and various filtering techniques.

Our work differs from prior work on using translations for WSD in that (a) we show that our method can achieve good results with only one language of translation, (b) our method is independent of statistical information such as relative sense frequencies, and (c) our method does not explicitly require any contextual information. In contrast, the method of Apidianaki and Gong (2015) backs off to the BabelNet first sense (BFS), a frequency-based baseline, if it is unable to narrow down the sense of the target word. This back-off strategy is particularly undesirable for tagging tokens that correspond to rare word senses. Moreover, their method is tested only with multiple languages of translation, and is applied directly to all-words WSD on a parallel corpus, rather than to generation of high-precision training data. The method of Bonansinga and Bond (2016) similarly depends on sense frequency information, and is evaluated only intrinsically, with multiple languages of translation. The method of Luan et al. (2020) depends on an existing disambiguation of the text, in addition to translations. Thus, our method is unique in that it can produce supplementary WSD training data with minimal assumptions about the available resources.

```
for each token $e$ on the $S$ side of $C$ do
    if $\exists$ token $f$ aligned with $e$ then
        $M_e \leftarrow$ the set of synsets containing $e$
        $M_f \leftarrow$ the set of synsets containing $f$
        if $|M_e \cap M_f| = 1$ then
            Let $s$ be the sole synset in $M_e \cap M_f$
            Tag $e$ with sense $(e, s)$
```

Figure 4.2: Pseudo-code for the sense tagging algorithm.

4.5.2 Method

Our method is inspired by Loureiro and Camacho-Collados (2020). They sense annotate only tokens that correspond to monosemous words, i.e., those that have only one sense, which is a trivial task in itself. However, they also show that a WSD method which propagates information between senses of different words can benefit from these annotations. For example, the monosemous word *airplane* is a synonym of the word *plane*, which is polysemous. Therefore, an annotated instance of *airplane* can inform a model about the context in which the corresponding sense of *plane* may appear.

In our approach, instead of monosemous words, we sense tag tokens which can be disambiguated based on their translations. For example, the English noun *vault* has four senses, corresponding to a burial vault, a bank vault, an arched ceiling, or a jump over an obstacle. The Polish word *wolta* can translate only the “jump” sense. Therefore, if we find an instance of *vault* translated as *wolta*, we can annotate *vault* with its “jump” sense, as no other sense could have been so translated. The absence of parallel polysemy between *vault* and *wolta* is a sufficient condition for the correctness of this annotation, regardless of whether OSPT holds for all Polish translations of *vault*. Our method uses this approach to partially annotate a parallel corpus, creating new sense-annotated WSD training data. Our hypothesis is that adding our translation-based annotations to a standard training corpus will improve the results of a supervised WSD system.

We follow the theoretical framework established in Chapter 2. The sense inventories, as well as the mapping between senses and translations, can be obtained from a multilingual wordnet, such as BabelNet. Multilingual wordnets consist of synonym sets, or *synsets*, each corresponding to a concept, and containing the words which can express that concept. The synsets that contain a word correspond to its senses; a sense can be viewed as a pair of a word and a synset that contains it. The target-language words in that synset are the words which can translate that sense. For example, in Figure 4.1, a multilingual wordnet should have a synset corresponding to the concept of "wood (material)" which contains *wood* and *legno*, but not *selva*.

The pseudo-code of the algorithm is shown in Figure 4.2. It takes as input a sentence-aligned parallel corpus C, involving the source language S (in our experiments, English) and the target language T, which has been tokenized, lemmatized, POS-tagged, and word-aligned. The algorithm generates sense tags for a subset of the tokens on the source side of C. The algorithm consults a wordnet that covers languages S and T. For each content word token e on the source side aligned with a single target-language token f, we determine the number of synsets which contain both e and f. Since each sense of a word uniquely corresponds to a synset containing that word, this is equivalent to determining how many senses of e can be translated by f. If the result is exactly one, we annotate e with its sense corresponding to the synset s that it shares with f. For example, if an instance of *wood* is aligned with *selva*, it is tagged with its "forest" sense, given that it is the only sense of *wood* which *selva* can translate.

Our method is unsupervised, efficient, scalable, and fully explainable. Its running time scales linearly with the size of the corpus. The resources upon which it depends are freely available for a wide variety of languages. These include the parallel corpora our method annotates, a multilingual wordnet, as well as tools for tokenization, POS-tagging, and alignment. It operates purely on the basis of contextual translation, without the need for additional tools such as knowledge-based WSD systems or contextual embeddings.

4.5.3 Intrinsic Evaluation

We test our translation-based corpus-tagging method on the manual sense annotations in MultiSemCor, or MSC (Bentivogli and Pianta, 2005), a word-aligned sense-annotated bitext, which was created by manually translating SemCor (Miller et al., 1993). It is tokenized, POS-tagged, and word-aligned with a knowledge-based aligner. There are 91,937 English word tokens in MSC annotated with exactly one WordNet 1.6 sense, and aligned with a single Italian word. We randomly select 10,000 of these tokens, and strip them of their sense annotations to form our test set.

As our multilingual wordnet, we use MultiWordNet, or MWN (Pianta et al., 2002) version 1.5.0. MWN was created by expanding Princeton WordNet 1.6 by adding Italian translations, as well as new synsets to cover English lexical gaps. Each English word in each multilingual synset is associated with a corresponding WordNet 1.6 sense, and each such English sense is associated with a (possibly empty) set of Italian translations. This provides the basis for our mapping of English senses to Italian translations. To mitigate the sense omission errors in MWN, we enrich it with 81,937 sense-translation pairs from MSC, excluding those which are in our 10k-token test set.

The results of the application or our method to the 10,000 annotated tokens in the test set yield a coverage of 33.3% and a precision of 92.6%, with the majority of errors caused by missing translations in MWN. Thus, our unsupervised method achieves higher precision than contemporary supervised WSD systems on standard English WSD datasets (Barba et al., 2021c). While these results are not directly comparable due to the different test sets, we interpret this as strong evidence for the efficacy and utility of our method for generating high-quality WSD training data.

4.5.4 Extrinsic Evaluation

Having demonstrated that our method can accurately disambiguate a subset of the tokens in a corpus, in this section we test whether sense-annotated data produced in this way can be used to improve the performance of a supervised WSD system.

Corpus	Tokens	Senses	Lemmas
SemCor	226,036	33,316	22,899
F10	219,793	28,589	23,033
FFSC	117,646	16,818	15,329
FFLC	90,616	13,147	12,406

Table 4.2: Statistics on the sets of sense annotations generated using the three filtering procedures.

This is achieved by appending the data that our translation-based method produces to SemCor, a standard training corpus for English WSD. Note that no manual sense annotations exist for the corpus that we annotate in these experiments; we are creating novel sense-annotated data.

Experimental Setup

Our parallel corpus is the English-Italian part of the OpenSubtitles corpus (Lison and Tiedemann, 2016), which contains approximately 35M sentence pairs. We tokenize, lemmatize, and POS-tag both sides of the corpus with TreeTagger (Schmid, 2013) using pre-trained models.[3] We perform word alignment with BabAlign (Luan et al., 2020), which refines the output of FastAlign (Dyer et al., 2013) by leveraging BabelNet as a source of lexical knowledge.

We again derive a sense-translation mapping from MultiWordNet, but this time without adding information from MultiSemCor. Since MultiWordNet is based on WordNet 1.6, we map each sense annotation to its most probable WordNet 3.0 equivalent, using a publicly available probabilistic mapping.[4]

As our supervised WSD system, we adopt the latest version of LMMS (Loureiro et al., 2022), which exploits relations between senses derived from WordNet in order to share information across related senses.

Filtering Annotations

Supervised WSD systems tend to exhibit a bias toward senses which are more frequent in the training data (Loureiro et al., 2020). Therefore, even a set of perfectly

[3] https://cis.uni-muenchen.de/~schmid

[4] http://www.lsi.upc.es/~nlp

Dataset	Full	MFS	LFS	ZSS	ZSL
SE2	2,282	1,486	796	385	255
SE3	1,850	1,213	637	198	112
S07	455	250	205	53	20
S13	1,644	1,031	613	341	202
S15	1,022	623	399	204	103
ALL	7,253	4,603	2,650	1,181	692

Table 4.3: Number of instances in each of the subsets of each dataset and the concatenation of all five datasets.

correct sense annotations may degrade the model's performance if the sense frequency distribution in the newly produced data diverges from that of the test data, which is not known in advance. We therefore filter the generated annotations to avoid greatly altering the sense frequency distribution of SemCor.

Following the example of Loureiro and Camacho-Collados (2020), we limit the number of annotated instances of each individual sense to 10, selected at random. This not only helps to prevent highly unbalanced sense frequency distributions, but also reduces the training time on the generated corpora. We refer to this set of instances as `F10`. In order to focus on gaps in the coverage on SemCor, we also test two additional filtering strategies that are applied to the annotations in `F10`. The first filters for lemma coverage (`FFLC`), by removing all annotations for *lemmas* which appear in SemCor. The second filters for sense coverage (`FFSC`), by removing all annotations for *senses* which appear in SemCor. Therefore, the `FFLC` annotations are a subset of the `FFSC` annotations, which in turn are a subset of the `F10` annotations.

Datasets

We obtain baseline results by training LMMS on SemCor, specifically the version provided by Raganato et al. (2017). To test our method, we train three additional LMMS models which augment SemCor annotations with `F10`, `FFSC`, and `FFLC`, respectively. The sizes of these generated supplementary datasets, and of SemCor itself, are shown in Table 4.2.

We evaluate our models on the standard WSD benchmark of Raganato et al. (2017), henceforth "R17". In addition to providing the baseline SemCor training corpus, R17 also contains five English WSD test sets created for five shared tasks: Senseval-2, or “SE2” (Edmonds and Cotton, 2001), Senseval-3, or “SE3” (Snyder and Palmer, 2004), SemEval-2007, or “S07” (Pradhan et al., 2007), SemEval-2013, or “S13” (Navigli et al., 2013), and SemEval-2015, or “S15” (Moro and Navigli, 2015). Following prior work, we use the S07 dataset to develop our method. We also evaluate our models on the concatenation of all five datasets, referred to as ALL[5], using the provided evaluation program; since LMMS disambiguates all words, the metrics precision, recall, F1, and accuracy are all equal throughout these experiments.

Following Blevins and Zettlemoyer (2020), we also test on the following subsets of ALL:

1. MFS (most frequent sense): Instances for which the correct sense is the WordNet first sense (i.e. the most frequent in SemCor).
2. LFS (less frequent sense): Instances for which the correct sense is *not* the WordNet first sense.
3. ZSS (zero-shot senses): Instances for which the sense is not in SemCor.
4. ZSL (zero-shot lemmas): Instances for which the lemma is not in SemCor.

MFS and LFS are disjoint, and their union is the complete dataset; ZSL is a subset of ZSS. Table 4.3 shows the size of each such subset.

We also test our models on five new benchmark datasets of Maru et al. (2022):

1. 42D: A newly created set, designed to be challenging by ensuring that all correct sense tags are neither the most frequent sense of their word, nor observed in SemCor. In other words, by design, all instances in this set satisfy the conditions for LFS and ZSS.
2. ALLamended (ALLa): A revised version of ALL from R17.
3. S10amended (S10a): A revised version of the dataset from SemEval-2010 Task 17 (Agirre et al., 2010).
4. hardEN (hEN): Those instances from 42D, ALLa, and S10a which were found to be answered incorrectly by all of a selection of WSD systems.
5. softEN (sEN): Those instances from 42D, ALLa, and S10a which are not in

[5] This includes S07, as is standard in the WSD literature.

Training Data	R17						M22				
	SE2	SE3	S07	S13	S15	ALL	42D	ALLa	S10a	hEN	sEN
SemCor (Baseline)	76.1	**73.9**	67.0	**75.2**	77.4	75.0	35.9	74.9	**77.3**	12.6	**78.0**
SemCor + F10	74.9	72.6	65.9	72.8	78.2	73.7	**40.5**	73.3	77.1	**15.1**	76.6
SemCor + FFLC	**76.7**	**73.9**	**67.5**	75.0	77.5	**75.1**	34.9	**75.1**	76.6	13.4	77.9
SemCor + FFSC	76.2	72.3	66.8	73.5	**78.3**	74.3	38.4	74.1	76.3	14.7	77.1

Table 4.4: F1-scores (in %) on the 10 WSD test sets. SE07 is the development set. The best results are in bold.

hardEN.

We henceforth refer to these datasets collectively as M22.

Results

The results in Tables 4.4 and 4.5 show that adding supplementary training data created by our method generally increases WSD accuracy, especially on rare and unseen senses. On the recently proposed 42D and hardEN challenge sets, we observe accuracy improvements of 4.6% and 2.5% respectively, using the F10 filtering strategy. This same approach yields improvements on LFS, ZSS, and ZSL partitions of the R17 ALL set, demonstrating that our method makes models more robust against such instances. We interpret these results as evidence for the efficacy and utility of our translation-based corpus tagging method.

The results further suggest that filtering generated annotations has a substantial impact on the resulting model. The frequency with which a word can be tagged with a particular sense by leveraging lexical translation need not correlate with the frequency of that sense in practice. Therefore, when using such generated corpora, care should be taken to select an appropriate filtering strategy. For instance, in a corpus where unseen senses or words are expected (e.g., in an unusual genre or domain), the FFSC filtering strategy may be the best option, as shown by its accuracy yields on ZSS and ZSL instances.

We conclude that our method for translation-based sense tagging offers substantial benefits, especially on difficult instances (Blevins et al., 2021). These improvements are obtained using a recent WSD method which is based on pre-trained transformer-based language models. This demonstrates that lexical translation can be a useful source of information even for modern WSD systems.

Training Data	R17 - ALL			
	MFS	LFS	ZSS	ZSL
SemCor (Baseline)	85.4	51.2	58.9	88.9
SemCor + `F10`	83.1	**52.1**	61.7	89.5
SemCor + `FFLC`	**85.5**	51.3	60.1	89.6
SemCor + `FFSC`	83.9	51.9	**62.7**	**89.7**

Table 4.5: F1-scores (in %) on subsets of the concatenation of all R17 datasets. The best results are in bold.

As a final note, we note that since the phenomenon of *parallel polysemy* is closely related to that of *parallel homonymy*, our approach is well-suited to homonym-level disambiguation. In Chapter 3, we argued that homonym distinctions are the coarsest possible sense inventory, and that almost all homonyms have disjoint sets of translations. Therefore, unlike OSPT, One Homonym per Translation (OHPT) *does* hold in general. Our translation-based approach could therefore be applied with near-perfect accuracy to disambiguate words at the homonym level.

4.6 Discussion

Our theoretical analysis in Section 4.3 established that OSPT is a sufficient condition for the ability to determine the sense of a word given its translation in context. However, the subsequent empirical analysis in Section 4.4 showed that OSPT does not hold in general. Nevertheless, our experiments in Section 4.5 provide clear evidence that we can leverage translations to produce high-precision sense annotations on the subset of word instances for which OSPT holds. These results demonstrate the importance of investigating the relations between senses, synonymy, polysemy, and translation. In this section, we further explore these ideas, taking the assumptions examined by Yao et al. (2012) (c.f., Section 4.2) to their logical extremes.

4.6.1 One Concept per Word: No Polysemy

First, let us consider an extreme scenario: a hypothetical language in which polysemy does not exist; that is, every content word has exactly one sense. In such a language, there could be no semantic ambiguity, and so WSD would be trivial: any given word could only express a single concept, regardless of its context. OSPT would always hold in such a language, no matter the language of translation, since each translation of a word could only translate its single sense.

To the best of our knowledge, no natural language contains only monosemous words. For example, 77.8% of English words in BabelNet occur in only one synset, with many of those being rare or technical terms. Similarly, Loureiro and Camacho-Collados (2020) observe that nearly 80% of lemmas in WordNet have only one sense, which allows them to generate useful resources for WSD. Only some constructed languages, such as Lojban/Loglan, strive to enforce complete monosemy on the lexicon (Cowan, 1997).

The untenable position that rejects any partitioning of word meanings into senses ("one sense per word") relates to various approaches to both theoretical and computational linguistics. In theoretical linguistics, the *monosemist* approach holds that different observed senses of a polysemous word result from a combination of its unique core meaning with the pragmatics of each specific context (François, 2008). In computational linguistics, methods that rely on exclusive use of static word embeddings, such as those learned by word2vec (Mikolov et al., 2013) make no allowance for discrete senses or sense embeddings.

4.6.2 One Word per Concept: No Synonymy

Now, let us consider the opposite extreme: a hypothetical language without synonymy. If a wordnet were constructed for such a language, every synset would contain exactly one word. For any given concept, there would be at most one word that could be used to express it. One Translation per Sense (OTPS) would always hold if such a language was used as the language of translation.

Again, it is unlikely that the entire lexicon of any natural language could satisfy this requirement. A language could perhaps be *constructed* according to this principle: for example, in Esperanto, synonymy and homonymy are considered undesirable (Puškar, 2015). Moreover, there will be a subset of any language which *does* satisfy this property. Indeed, approximately 56% of WordNet 3.0 synsets contain only one word (e.g., *proton*).

A similar position in computational linguistics ("one sense per context") is diametrically opposite to the monosemist approach described above. For example, Martelli et al. (2021) propose "dropping the requirement of a fixed sense inventory" and instead using representations which assign each word token a unique contextualized embedding. Such a position can be interpreted as an assignment of a unique sense to every occurrence of a given word in a distinct context. In view of our theoretical investigation, such an approach is effectively incompatible with our definition of synonymy. Nevertheless, the existence of synonymy in any human language is widely accepted in linguistics. In addition, computational linguistics tasks, such as machine translation, need to account for synonymy, given that the goal is to produce fully fluent, rather than just semantically correct texts and utterances.

4.6.3 One Word $\equiv$ One Concept

If the two constraints described above are combined, it would result in a language that has neither polysemy nor synonymy. We refer to this hypothetical language as *Interlingua*. In Interlingua, every concept could be expressed by exactly one word, which could express only that concept; every synset would have a size of one, and every word would be in one synset. Assuming a sense-translation mapping is available, e.g. via a multilingual wordnet which includes Interlingua, lexical translation *into* Interlingua could be reduced to identifying the sense of the source word. The converse also holds: the sense of a word could always be identified, given its translation into Interlingua. Working in the other direction, given a perfect multilingual wordnet, finding a translation for an Interlingua word would only require selecting a word from the corresponding synset in the target language.

Perhaps the most direct application for Interlingua is language-independent semantic parsing. Martínez Lorenzo et al. (2022) propose the *BabelNet Meaning Representation* (BMR), a semantic parsing formalism which converts an input sentence into a language-independent representation. Each content word is mapped to the unique identifier of the BabelNet synset corresponding to the concept it refers to. This creates a formal meta-language in which every concept is unambiguously expressed in exactly one way: by the corresponding BabelNet synset ID. Hence, the BMR satisfies one "word" per concept *and* one concept per "word", with BabelNet IDs taking the place of words. There is no synonymy, as each ID is by design unique in representing its particular concept, nor is there polysemy, as each ID is unambiguous in its reference to some lexicalized concept. Thus, what may appear as a completely hypothetical and abstract construct can in fact be viewed as a theoretical model of a modern semantic approach.

4.7 Conclusion

In this chapter, we formulated several propositions related to senses, translations, synonymy, and polysemy. We show empirically that the assumptions that would allow translations to serve as a sense inventory hold simultaneously only for a small fraction of words. Nevertheless, we also demonstrate that the link between word senses and translations is not merely of theoretical interest. In particular, we present a method for leveraging translations to perform high-precision unsupervised sense annotation. We observe substantial WSD improvements especially on senses or lemmas that are less frequent or not found at all in existing training data.

Considering the above applications to constructed languages, contextual embeddings, and semantic parsing, we intend to continue our theoretical investigations into open issues in multilingual lexical semantics, and guide empirical research toward more explainable models and results.

Chapter 5

WiC = TSV = WSD: On the Equivalence of Three Semantic Tasks

The word-in-context (WiC) task has attracted considerable attention in the NLP community, as demonstrated by the popularity of the recent MCL-WiC SemEval shared task. Systems and lexical resources from word sense disambiguation (WSD) are often used for the WiC task and WiC dataset construction. In this chapter, we establish the exact relationship between WiC and WSD, as well as the related task of target sense verification (TSV). Building upon a novel hypothesis on the equivalence of sense and meaning distinctions, we demonstrate through the application of tools from theoretical computer science that these three semantic classification problems can be pairwise reduced to each other, and therefore are equivalent. The results of experiments that involve systems and datasets for both WiC and WSD provide strong empirical evidence that our problem reductions work in practice.[1]

5.1 Introduction

This chapter answers an open question about the the relation between two important tasks in lexical semantics. Word sense disambiguation (WSD) is the task of tagging a word in context with its sense (Navigli, 2009). The word-in-context (WiC) problem is the task of deciding whether a word has the same meaning in two different contexts (Pilehvar and Camacho-Collados, 2019). A crucial difference between the

[1] This chapter is based on Hauer and Kondrak (2022). See the preface for details.

two tasks is that WSD depends on a pre-defined sense inventory[2] while WiC does not involve any identification or description of word meanings. Despite ongoing interest in both tasks, there is substantial disagreement in the literature as to whether WiC is a re-formulation of WSD (e.g. Levine et al. (2020)) or an entirely distinct task (e.g. Martelli et al. (2021)).

By establishing that WSD and WiC are equivalent, we construct a theoretical foundation for the transfer of resources and methods between the two tasks. WSD has been intensively studied for decades, while WiC has recently attracted considerable attention from the research community. For example, the MCL-WiC SemEval shared task (Martelli et al., 2021) attracted 48 teams, and WiC instances have been integrated into the SuperGLUE benchmark (Wang et al., 2019). Understanding how the two tasks relate to each other allows us to correctly interpret and confidently build upon those results, including prior work on using WSD systems for WiC (e.g. Loureiro and Jorge (2019)).

We establish the theoretical equivalence of WiC and WSD by specifying reduction algorithms which produce a solution for one problem by applying an algorithm for another. In particular, we employ the target sense verification (TSV) task (Breit et al., 2021) as an intermediate step between WSD and WiC, and specify three reductions: WiC to WSD, WSD to TSV, and TSV to WiC. We formalize the three problems using a common notation, and provide both theoretical and empirical evidence for the correctness of our reductions. While we focus on English in this chapter, we make no language-specific assumptions.[3]

The soundness of all three tasks hinges on the consistency of judgments of sameness of word meaning, whether with respect to discrete sense inventories as in WSD, a representation of a single sense in TSV, or two occurrences of a word in WiC. We posit that *different instances of a word have the same meaning if and only if they have the same sense.* This empirically falsifiable proposition, which we refer to as the *sense-meaning hypothesis*, implies that WiC judgements induce

[2] For the purposes of this chapter, we assume that the WSD sense inventory, the discrete enumeration of the senses of each content word, is the WordNet sense inventory (Miller et al., 1990), which is a standard practice in WSD (Raganato et al., 2017).

[3] Hauer et al. (2021a) leverage translations from multiple languages for the WiC task by applying the substitution test for the synonymy of senses (see Chapter 2).

sense inventories that correspond to word senses. This counter-intuitive finding has intriguing implications for the task of word sense induction (WSI), as well as algorithmic wordnet construction.

We empirically validate our hypothesis by conducting multiple experiments and analyzing the results. In particular, we test our WSD-to-WiC and WiC-to-WSD reductions on standard benchmark datasets using state-of-the-art systems. We find that our reductions perform remarkably well, revealing no clear counter-examples to our hypothesis in the process.

Our contributions are as follows: (1) We answer the open question of the relation between WiC and WSD by constructing a theoretical argument for their equivalence, which is based on the novel sense-meaning hypothesis. (2) We carry out a series of validation experiments that strongly support the correctness of our reductions. (3) We release the details of our manual analysis and annotations of the instances identified in the validation experiments.

5.2 Theoretical Formalization

In this section, we formally define the three problems, present a theoretical argument for their equivalence, and specify the reductions.

5.2.1 Problem Definitions

Senses in our problem definitions refer to *wordnet senses*. A *wordnet* is a theoretical construct which is composed of synonym sets, or *synsets*, such that each synset corresponds to a unique concept, and each sense of a given word corresponds to a different synset. Actual wordnets, such as Princeton WordNet (Miller et al., 1990; Fellbaum, 1998), are considered to be imperfect implementations of the theoretical construct.

In the problem definitions below, C, C_1, C_2 represent contexts, each of which contains a single focus word w used in the sense s. We assume that every content word token is used in exactly one sense.[4]

[4] This is empirically supported by the fact that 99.7% of annotated tokens in SemCor are assigned a single sense.

- **WSD**(C, w): Given a context C which contains a single focus word w, return the sense s of w in C.
- **TSV**(C, w, s): Given a context C which contains a single focus word w, and a sense s, return TRUE if s is the sense of w in C, and FALSE otherwise.
- **WiC**(C_1, C_2, w): Given two contexts C_1 and C_2 which contain the same focus word w, return TRUE if w has the same meaning in both C_1 and C_2, and FALSE otherwise.

5.2.2 Problem Equivalence

The theoretical argument for the sense-meaning hypothesis is based on the assumption that the relation of sameness of word meaning is shared between the three problems. This is supported by the lack of distinction between *meanings* and *senses* in the original WiC task proposal.[5] On the other hand, WordNet exhibits a strict one-to-one correspondence between distinct meanings, synsets, and concepts (see Chapter 2), with each word sense corresponding to a specific synset. This implies that senses are ultimately grounded in sameness of meaning as well.[6] Therefore, every word meaning distinction should correspond to a pairwise sense distinction. Contrariwise, if two tokens of the same word express different concepts, their meaning must be different. This equivalence also includes the TSV problem, provided that the given sense of the focus word corresponds to a single synset.

5.2.3 Problem Reductions

We now present the three problem reductions. For our purposes, a **P-to-Q** reduction is an algorithm that, given an algorithm for a problem **Q**, solves an instance of a problem **P** by combining the solutions of one or more instances of **Q**.

Proposition 1. *WiC is reducible to WSD.*

To reduce **WiC to WSD**, we directly apply the sense-meaning hypothesis from Section 5.1 by assuming that the focus word has the same meaning in two contexts

[5] "The proposed dataset, WiC, is based on lexicographic examples, which constitute a reliable basis to [. . .] discern different **meanings of words**." (Pilehvar and Camacho-Collados, 2019).

[6] "[Each] synonym set represents one underlying lexical concept. [. . .] **Word meaning** [refers] to the lexicalized concept that a [word] form can be used to express." (Miller, 1995).

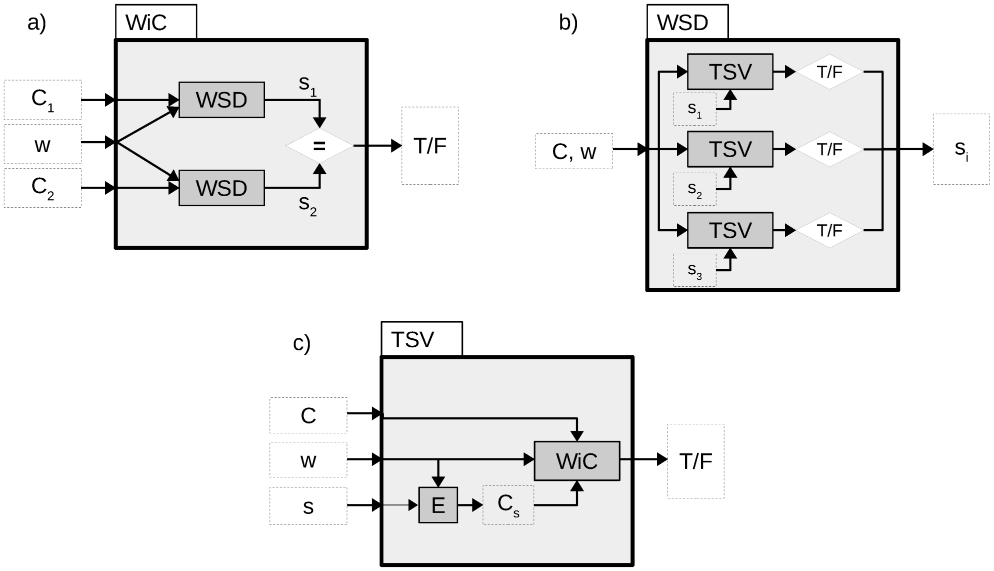

Figure 5.1: Three problem reductions: a) WiC to WSD, b) WSD to TSV, and c) TSV to WiC.

if and only if it can be independently tagged with the same sense in both contexts. Formally:

$$\text{WiC}(C_1, C_2, w) \Leftrightarrow \text{WSD}(C_1, w) = \text{WSD}(C_2, w)$$

Thus, given a method for solving WSD, we can solve any given WiC instance by solving the two WSD instances which consist of the focus word in the first and second context, respectively. We return TRUE if the returned senses are equal, FALSE otherwise (Figure 5.1a).

Proposition 2. *WSD is reducible to TSV.*

To reduce **WSD to TSV**, we take advantage of the fact that TSV can be applied to a variety of different sense representations, without any explicit dependence on a specific sense inventory. We can therefore query a TSV system with various senses of the focus word, using the same sense inventory as the WSD task:

$$\text{WSD}(C, w) = s \Leftrightarrow \text{TSV}(C, w, s)$$

Thus, given a TSV solver, for any WSD instance we can construct a list of k TSV instances, one for each sense of the focus word in the corresponding WSD sense inventory. We return the sense for which the TSV instance returns TRUE (Figure 5.1b). The correctness of this reduction hinges on the assumption that every content word in context is used in exactly one sense.

Proposition 3. *TSV is reducible to WiC.*

To reduce **TSV to WiC**, we again leverage our sense-meaning hypothesis by assuming that a content word used in a particular sense will be judged to have the same meaning as in an example sentence for that sense. Formally:

$$\text{TSV}(C, w, s) \Leftrightarrow \text{WiC}(C, C_s, w)$$

where C_s is a context in which w is unambiguously used in sense s. So, given a method for solving WiC, we can solve a TSV instance by replacing the given sense representation with an example, yielding a WiC instance (Figure 5.1c). This reduction depends on the existence of an algorithm E that, given a sense s of a word w, can generate an example sentence C_s that contains w used in sense s.[7]

These three reductions are sufficient to establish the equivalence of **WSD**, **TSV**, and **WiC**. A method which solves any of these problems can be used to construct methods which solve the other two, using a sequence of at most two of the above reductions.

In particular, we can reduce **WSD to WiC**:

Corollary 1. *WSD is reducible to WiC.*

To reduce **WSD to WiC**, first reduce the WSD instance to TSV, producing one TSV instance for each sense s of w. Then, reduce each of these TSV instances to a WiC instance, by pairing the context of the WSD instance with an example context for each sense. Succinctly:

$$\text{WSD}(C, w) = s \Leftrightarrow \text{WiC}(C, C_s, w)$$

Thus, solving the original WSD instance can be achieved by identifying the single positive instance in the list of k WiC instances.

[7] This is related to a well-defined and actively researched task known as exemplification modelling (Barba et al., 2021b).

5.3 WiC Datasets

In this section, we discuss and analyze the existing WiC datasets with the aim of finding a dataset suitable for validating our equivalence hypothesis. An instance that contradicts one of the reduction equivalences in Section 5.2.3 would be an exception to the hypothesis. Since natural language is not pure logic, falsifying the hypothesis would require finding that such exceptions constitute a substantial fraction of instances, excluding apparent exceptions caused by errors and omissions in lexical resources.

5.3.1 WiC

WiC was originally proposed as a dataset for the evaluation of contextualized embeddings, including neural language models (Pilehvar and Camacho-Collados, 2019). The original WiC dataset consists of pairs of sentences drawn mostly from WordNet, which were further filtered to remove fine-grained sense distinctions. The reported inter-annotator agreement was 80% for the final pruned set, and only 57% for the pruned-out instances.

Since, regardless of the source, all instances were annotated automatically by checking the sense identity in WordNet, the WiC dataset cannot, *by its construction*, contain any exceptions to the equivalence hypothesis. Therefore, we do not use the original WiC dataset in our experiments. Nevertheless, it is possible to automatically identify both senses in about half the instances in the dataset by matching them to the sense usage example sentences in WordNet 3.0. It is interesting to note that combining such a WordNet lookup with a random back-off on the remaining instances results in correctly solving 76.1% of the WiC instances in the test set, which exceeds the current state-of-the-art results of 72.1% (Levine et al., 2020).

5.3.2 WiC-TSV

Breit et al. (2021) propose *target sense verification* (TSV), the task of deciding whether a given word in a given context is used in a given sense. TSV is similar to WiC in that it is also a binary classification task, but only one context is provided.

TSV is also similar to WSD in that there is an explicit representation of senses, but there is only one sense to consider. Three sub-tasks are defined depending on the method of representing a sense: (a) definition, (b) hypernyms, and (c) both definition and hypernyms.

Approximately 85% of the instances in the WiC-TSV dataset are derived directly from the original WiC dataset, and so are ultimately based on WordNet senses.[8] Specifically, the sense of the focus word was established by reversing the process by which the WiC instances were created, as in the WordNet lookup procedure applied to the WiC dataset in Section 5.3.1. Because of this construction method, no exceptions to the equivalence hypothesis can be found in the WiC-TSV dataset.

5.3.3 MCL-WiC

Martelli et al. (2021) introduce the Multilingual and Cross-lingual Word-in-Context dataset. The English portion of the dataset consists of 10k WiC instances, divided into a training set (8k instances), as well as development and test sets (1k instances each). The task is exactly the same as the original WiC task, and matches our WiC problem formalization in Section 5.2.1. In particular, while the dataset covers multiple languages, the task itself remains monolingual, in the sense that the system need only consider one language at a time; that is, all input and output for a given instance is in a single language.

In contrast with the original WiC dataset, which was largely derived from WordNet, the sentence pairs in MCL-WiC were manually selected and annotated. Annotators consulted “multiple reputable dictionaries” to minimize the subjectivity of their decisions on the identity of meaning. As a result, both the inter-annotator agreement ($\kappa = 0.968$), and the best system accuracy (93.3% on English (Gupta et al., 2021)) are much higher than on the original WiC dataset.

The MCL-WiC dataset (Section 5.3.3) is especially valuable for testing our sense-meaning equivalence hypothesis because it does not rely on pre-existing Word-

[8] Three smaller sets are devoted to cocktail, medical, and computer terms, respectively, and appear more related to named entity recognition than to WSD.

Net sense annotations, and is agnostic toward WordNet sense distinctions. For this reason, we make the MCL-WiC dataset the focus of our empirical validation experiments in the next section.

5.4 Empirical Validation

In this section, we aim to quantify and analyze any apparent counter-examples to the sense-meaning hypothesis which are identified in the process of testing the WSD-to-WiC and WiC-to-WSD reductions. We are particularly interested in the exceptions that cannot be attributed to errors in the resources that are used to implement the reductions, because such exceptions represent potential evidence against our hypothesis.

5.4.1 Systems

In order to implement the WSD-to-WiC and WiC-to-WSD reductions, we adopt two recent systems designed for the WiC and WSD tasks, respectively.

Our WiC system of choice is LIORI (Davletov et al., 2021). In the MCL-WiC shared task, LIORI obtained an accuracy of 91.1% on the English *test set*, which was within 2% of the best performing system. LIORI works by concatenating each sentence pair into a single string, and fine-tuning a neural language model for binary classification. We use the code made available by the authors[9], and derive our model from the MCL-WiC English training set.

As our WSD system, we adopt ESCHER (Barba et al., 2021a). ESCHER reformulates WSD as a span extraction task: For a given WSD instance, the context is concatenated with all glosses of the focus word into a single string, from which the gloss of the correct sense is extracted. We derive our model using the implementation and training procedure provided by the authors[10]. The training data includes SemCor (Miller et al., 1993). In our replication experiments, this model achieves 80.1% F1 on the standard WSD benchmark datasets of Raganato et al. (2017).

[9] https://github.com/davletov-aa/mcl-wic

[10] https://github.com/SapienzaNLP/esc

5.4.2 Solving WSD with WiC

Our first experiment involves an implementation of the reduction of WSD to WiC. For each WSD instance, we construct a set of WiC instances that correspond to its possible senses, solve them with LIORI, and return a single sense, in accordance with the reduction specified in Corollary 1 from Section 5.2.3. We then present and analyze the results on a standard WSD dataset.

Implementation of the Reduction

Given a WSD instance consisting of a focus word w in a context C, we create a set of k WiC instances, where k is the number of senses of w. In WordNet 3.0, each sense s has a gloss g_s, and sometimes also a usage example of w being used in sense s. Since not all synsets are accompanied by usage examples, we instead generate a new synthetic usage example C_s for each sense of w using the following pattern: $C_s :=$ " 'w' *in this context means* g_s". Thus C_s represents an unambiguous example of w being used in sense s. The resulting WiC instance for s is then composed of contexts C and C_s, both of which include the focus word w.

Our LIORI model returns a binary classification and a score for each of the constructed WiC instances. While LIORI may classify zero, one, or more instances as true, LIORI also produces a score for each instance, and our implementation returns only the sense with the highest score. This is in accordance with the definition of the WSD task as identifying a single correct sense for a word in context (Section 5.2.1).

Results and Discussion

To estimate the expected accuracy of the above implementation, we first apply LIORI to the 1000 instances in the MCL-WiC English development set. LIORI achieves an accuracy of 88.0%, which we use as an estimate of the probability that LIORI correctly classifies any given WiC instance. The average number of senses per instance in our WSD dataset is approximately 8.5. Since any error by LIORI can cause the WSD-to-WiC reduction to output the wrong sense, we estimate the expected probability that LIORI correctly classifies a single WSD instance as $0.880^{8.5} \approx 0.34$.

We test the reduction on the SemEval 2007 dataset, as provided by Raganato et al. (2017). This test set contains 455 WSD instances, all but four of which (over 99%) are annotated with exactly one sense. Our reduction implementation obtains an accuracy of 47.9% by returning a single predicted sense for every WSD instance in the test set. As this result is substantially higher than the expected accuracy of 34%, we interpret it as evidence in favor of our hypothesis.

In theory, for each WSD instance, LIORI should classify as true exactly one of the constructed WiC instances, which represents the single correct sense. In practice, this is the case in only 48 out of 455 cases. Our reduction implementation predicts the correct sense for 38 out of 48, yielding a precision of 79.2%. We verified that ESCHER, trained on over 226k sense annotations in SemCor, correctly annotates 39 of these 48 instances. On this subset of instances, our WSD-to-WiC reduction based on LIORI is therefore competitive with state-of-the-art supervised WSD systems, despite not depending on any sense-annotated training data. This constitutes further evidence for the correctness of our reduction, and our hypothesis.

5.4.3 Solving WiC with WSD

In this experiment, we apply a state-of-the-art supervised WSD system to solve, via our WiC-to-WSD reduction, all WiC instances in an independently-annotated test set. We then manually analyze a sample of the errors to assess whether the experiment supports our hypothesis and the correctness of our reduction.

Implementation of the Reduction

The implementation of the WiC-to-WSD reduction is conceptually simpler that the previously described WSD-to-WiC reduction.[11] Given a WiC instance consisting of contexts C_1 and C_2 for a word w, we create two corresponding WSD instances: (C_1, w) and (C_2, w). Both WSD instances are passed to ESCHER, which independently assigns senses s_1 and s_2 to w in each of the two contexts. We classify the WiC instance as positive if and only if $s_1 = s_2$.

[11] In fact, Loureiro and Jorge (2019) implicitly apply this reduction on a WiC dataset with their WSD system LMMS.

There are two types of possible counter-examples to our hypothesis: (1) a WiC instance which is annotated as positive (i.e., the same meaning) in which both focus tokens have different senses; and (2) a WiC instance which is annotated as negative (i.e., different meanings) in which both focus tokens have the same sense. These two types could arise from WSD sense distinctions that are too fine-grained or too coarse-grained, respectively.

Expected Accuracy

The expected accuracy of the WiC-to-WSD reduction is more complex to calculate than that of the WSD-to-WiC reduction. Our calculation is based on the simplifying assumption that all WSD errors are independent and equally likely. For the probability that ESCHER disambiguates any WSD instance correctly, we use the value of $p = 0.801$, based on our replication result in Section 5.4.1. The average number of senses per focus token in the dataset used in our experiment is $k = 4.73$. Since there are $k-1$ incorrect senses for each WSD instance, we approximate the probability of predicting a given incorrect sense in either WiC sentence as $q = (1-p)/(k-1) = 0.053$.

In order to estimate the probability of a correct classification, we consider two main cases.

1. A *positive* WiC instance is *correctly* classified as positive if either (1.1) both corresponding WSD instances are disambiguated correctly, or (1.2) both instances are tagged with the same incorrect sense: $P_1 = p^2 + (k-1)q^2 = 0.642 + 0.011$.
2. A *negative* WiC instance is *incorrectly* classified as positive if either (2.1) one of the corresponding WSD instances is disambiguated correctly and the other is incorrectly tagged with the same sense, or (2.2) both instances are tagged with the same incorrect sense: $P_2 = 2pq + (k-2)q^2 = 0.085 + 0.008$.

Assuming that the dataset is balanced, the expected probability of classifying a WiC instance correctly is therefore: $P_1/2 + (1-P_2)/2 =$ **0.779**.

Results and Discussion

We test the reduction on the MCL-WiC English development set, which consists of 500 positive and 500 negative WiC instances. We tokenize, lemmatize, and POS-tag all 2000 sentences with TreeTagger[12] (Schmid, 1999) as a pre-processing step. ESCHER is then applied to predict the sense of the focus word in each sentence. In 25 cases, ESCHER failed to make a sense prediction, that is, one or both focus words were not disambiguated, due to TreeTagger tokenization or lemmatization errors. The accuracy on the remaining 975 instances is 78.5%, which is within 1% of our theoretical estimate in Section 5.4.3. We conclude that this experiment provides strong empirical support for our hypothesis and the correctness of our reductions.

Analysis

To further evaluate our WiC-to-WSD reduction, we manually analyzed a sample of 10 false positives and 10 false negatives from this experiment. The sample was *not* random; instead, we attempted to automatically select the instances that were most likely to represent exceptions to our equivalence hypothesis. Specifically, we restricted the analysis to WiC instances that were correctly classified by LIORI, in order to reduce the impact of erroneous annotations, which are unavoidable in any gold dataset. As a result, the accuracy of ESCHER on the WSD instances in this sample is expected to be lower than in the entire dataset. In fact, in 13 of the 20 instances (six false positives, seven false negatives), the misclassification was due to an error made by ESCHER.

In three of the seven remaining cases (all false positives), the WiC misclassification was caused by the WordNet sense inventory not including the correct sense of one of the focus tokens. Since we require ESCHER to produce a WordNet sense as output, such omissions preclude the correct disambiguation of the focus word. In all such cases, we were able to find the omitted sense in one of the dictionaries that we consulted (Oxford or Merriam-Webster). For example, the correct sense of the verb *partake* in the WiC sentence “he has **partaken** in many management

[12] `https://cis.uni-muenchen.de/~schmid/tools/TreeTagger`

	Lemma	Gloss	Dict
1	partake (v)	join in (an activity)	OED
2	instant (adj)	prepared quickly and with little effort	OED
3	familiar (adj)	of or relating to a family	MW
4	breach (v)	to leap out of water	MW
5	spotter (n)	a member of a motor racing team	OED
6	campaign (n)	an organized course of action to achieve a goal	OED
7	campaign (n)	a set of organized actions that a political candidate undertakes in an election	OED
8	drive (n)	determination and ambition to achieve something	OED
9	drive (n)	an organized effort by a number of people	OED
10	wedding (n)	a marriage ceremony with accompanying festivities	MW
11	wedding (n)	an act, process, or instance of joining in close association	MW
12	analyst (n)	someone who analyzes	Wik
13	analyst (n)	a financial analyst; a business analyst	Wik

Table 5.1: Examples of senses that are not in WordNet (Rows 1-5), and sense distinctions found in external dictionaries (Rows 6-13): OED (Oxford English Dictionary), MW (Merriam-Webster), Wik (Wiktionary).

courses" is "join in (an activity)" which is in the Oxford English Dictionary, but not in WordNet 3.0. The missing WordNet senses for each of these instances are shown in rows 1-3 of Table 5.1.

Among the remaining four instances, in one anomalous case we were unable to reach a consensus on the WordNet sense of the adverb *richly* in the phrase *richly rewarding*. However, in the other three cases, ESCHER's annotations were unquestionably correct. We defer the discussion of those three interesting instances to the next section.

5.4.4 Manual Annotation Experiment

To further expand our analysis, we manually analyzed 60 additional randomly selected instances from the English MCL-WiC training set. The size of the sample was limited because WSD instances are difficult and time-consuming to analyze, especially when multiple annotators are involved and an effort is made to avoid any unconscious bias.

For each such instance, we assigned WordNet senses to each of the two focus tokens, without accessing the gold MCL-WiC labels. Our judgments were based on the glosses and usage examples of the available senses, as well as the contents of the corresponding synsets and their hypernym synsets. Subsequently, we analyzed each instance where the WiC prediction obtained by applying the WiC-to-WSD reduction did not match the WiC classification in the official gold data.[13]

We found that 55 out of 60 instances (91.7%) unquestionably conform to the equivalence hypothesis. The remaining five instances can be divided into three categories: (1) tokenization errors in MCL-WiC, (2) missing senses in WordNet, and (3) possible annotation errors in MCL-WiC. We discuss these three types of errors below.

In two instances, word tokenization errors interfere with the MCL-WiC annotations: (1) *together* in "the final **coming together**" is annotated as an adverb instead of a particle of a phrasal verb, and (2) *shiner* in "**shoes shiners** met the inspector" is annotated as a stand-alone noun instead of a part of a compound noun. These tokenization errors prevent the proper assignment of WordNet senses.

In two instances (rows 4 and 5 in Table 5.1), one of the senses of the focus word is missing in WordNet: (1) *breach* referring to an animal breaking through the surface of the water, and (2) *spotter* referring to a member of a motor racing team who communicates by radio with the driver. Neither of these senses is subsumed

[13] We publish the annotated set of 60 WiC instances at `https://webdocs.cs.ualberta.ca/\~kondrak`

by another sense in WordNet, and both of them are present in one of the consulted dictionaries.

In the final problematic instance, MCL-WiC classifies the noun *campaign* as having the same meaning in the contexts "during the election **campaign**" and "the **campaign** had a positive impact on behavior." Since the distinction between these two senses of *campaign* is found in the Oxford English Dictionary, which was among the ones consulted by the MCL-WiC annotators (Martelli et al., 2021), we classify it as an MCL-WiC annotation error (rows 6 and 7 in Table 5.1).

Similarly, we posit an MCL-WiC annotation error in each of the three outstanding false negatives from Section 5.4.3, which could not be attributed to ESCHER, based on the verification in external dictionaries. For example, unlike WordNet, Oxford and Merriam-Webster both distinguish the emotional and organizational meanings of *drive*. Similar analysis applies in instances involving the words *wedding* and *analyst* (rows 8-13 in Table 5.1). Since the meanings of the focus words in these contexts are distinguished in a dictionary, they should be considered distinct meanings according to the annotation procedure of Martelli et al. (2021). We conclude that in these cases, the MCL-WiC label is incorrect, and so they do not constitute exceptions to our hypothesis.

In summary, a careful analysis of 25 apparent exceptions made by our reduction across 80 instances, using both automatic and manual WSD, reveals no clear evidence against the correctness of our reduction. We therefore conclude that the results of these experiments strongly support our hypothesis.

5.5 Discussion

Having presented theoretical and empirical evidence for the equivalence of WiC, WSD, and TSV, we devote this section to the discussion of the relationship between WordNet and WiC.

Most English WiC and TSV datasets are based, in whole or in part, on WordNet. If no sense inventory is used for grounding decisions about meaning, the inter-annotator agreement is reported to be only about 80% (Pilehvar and Camacho-

Collados, 2019; Breit et al., 2021). For the MCL-WiC dataset, however, annotators consulted other dictionaries, and obtained “almost perfect agreement” (Martelli et al., 2021). This suggests that sense inventories, and semantic resources in general, are crucial to reliable annotation for semantic tasks. However, because the exact MCL-WiC procedure for resolving differences between dictionaries is not fully specified, and because such dictionaries vary in their availability, the correctness of the annotations cannot be readily verified (c.f. Section 5.4.4).

Our experiments provide evidence that, even when the WordNet sense inventory is not explicitly used in constructing WiC datasets, WiC annotations nevertheless tend to agree with WordNet sense distinctions, as our hypothesis predicts. Namely, the MCL-WiC instances in which both focus tokens have the same sense are almost always annotated as positive by the MCL-WiC annotators. The converse also holds, with any exceptions being explainable by errors in the resources. Thus, empirical validation confirms our sense-meaning hypothesis, which implies that the meaning distinctions induced by WiC judgements closely match WordNet sense inventories. This is a remarkable finding given the high granularity of WordNet.

We postulate that the adoption of WordNet as the standard sense inventory for WiC would have several practical benefits: (1) it has been adopted as the standard inventory for WSD, and so would simplify multi-task evaluation; (2) it allows seamless application of systems across datasets; (3) it facilitates rapid creation of new WiC datasets based on existing sense-annotated corpora; (4) it is freely available; (5) it can be modified and extended to correct errors and omissions (McCrae et al., 2020); and finally (6) it can be extended to facilitate work with other languages, as in the XL-WiC dataset (Raganato et al., 2020).

In addition, WordNet has strong theoretical advantages. Its fine granularity is a consequence of its grounding in synonymy and lexical concepts. Therefore, the sense distinctions found in other dictionaries either already correspond to different WordNet concepts, or should lead to adding new concepts to WordNet. Furthermore, unlike in dictionaries, senses of different words in WordNet are linked via semantic relations such as synonymy and hypernymy, which facilitate an objective assignment of every word usage to a single WordNet concept. This property of

WordNet may be the reason that the WSD methods based on sense relation information have surpassed the inter-annotator agreement ceiling of around 70% (Navigli, 2006).

5.6 Conclusion

We formulated a novel sense-meaning hypothesis, which allowed us to demonstrate the equivalence of three semantic tasks by mutual reductions. We corroborated our conclusions by performing a series of experiments involving both WSD and WiC tools and resources. We have argued that these relationships originate from the WordNet properties, which are highly desirable in semantics research. We expect that our findings will stimulate future work on system development, resource creation, and joint model optimization for these tasks.

Chapter 6

Taxonomy of Problems in Lexical Semantics

Semantic tasks are rarely formally defined, and the exact relationship between them is an open question. We introduce a taxonomy that elucidates the connection between several problems in lexical semantics, including monolingual and cross-lingual variants. Our theoretical framework is based on the hypothesis of the equivalence of concept and meaning distinctions. Using algorithmic problem reductions, we demonstrate that all problems in the taxonomy can be reduced to word sense disambiguation (WSD), and that WSD itself can be reduced to some problems, making them theoretically equivalent. In addition, we carry out experiments that strongly support the soundness of the concept-meaning hypothesis, and the correctness of our reductions.[1]

6.1 Introduction

This chapter proposes a taxonomy of several problems in lexical semantics, consisting of a clear definition of each task, and a theory-driven analysis establishing the relationships between them (Figure 6.1). The taxonomy includes word sense disambiguation (WSD), word-in-context (WiC), lexical substitution (LexSub), and word synonymy (Syn). We consider their monolingual, cross-lingual, and multilingual variants. With the exception of WSD, they are all defined as binary decision problems.

[1] This chapter is based on Hauer and Kondrak (2023). See the preface for details.

Our theoretical problem formulations correspond to well-studied semantic tasks. In practice, these tasks are rarely precisely defined, and instead depend on annotated datasets. For example, the definitions of lexical substitution differ between publications, and involve imprecise terms, such as "the overall meaning of the context" or "suitable substitute." The exact relationships between these tasks have not been rigorously demonstrated. Altogether, the recent literature suggests that a more detailed taxonomy is very much needed.

We start by formally defining the problems in terms of concepts and contexts, and proceed to determine their relative hardness by specifying reduction algorithms which produce a solution for one problem by applying an algorithm for another. In particular, we demonstrate that all problems in the taxonomy can be reduced to WSD, which confirms the principal role of this problem in lexical semantics. Furthermore, we show by mutual reductions that WSD and multilingual variants of WiC and LexSub are theoretically equivalent. Finally, we shed light on how they relate to lexical translation and wordnets.

The soundness of the problems in the taxonomy hinges on the consistency of judgments of sameness of word meaning. In Chapter 5, we demonstrated the theoretical equivalence of the monolingual WiC and WSD via mutual reduction. We posit the following generalization of their sense-meaning hypothesis to multilingual concepts: *different word instances have the same* **meaning** *if and only if they express the same* **concept**. This empirically falsifiable proposition, which we refer to as the *concept-meaning hypothesis*, allows us to incorporate multilingual tasks, including lexical synonymy and substitution, into our theoretical framework.

In addition to showing that our theoretical propositions follow directly from our definitions and assumptions, we perform a series of experiments for the purpose of testing their empirical applicability and soundness. In particular, we test three problem reductions on standard benchmark datasets using independently developed systems based on pre-trained language models. Manual error analysis reveals no counter-examples to our concept-meaning hypothesis.

Our main contribution is a novel taxonomy of formally-defined problems, which establishes the reducibility or equivalence relations between the principal tasks in

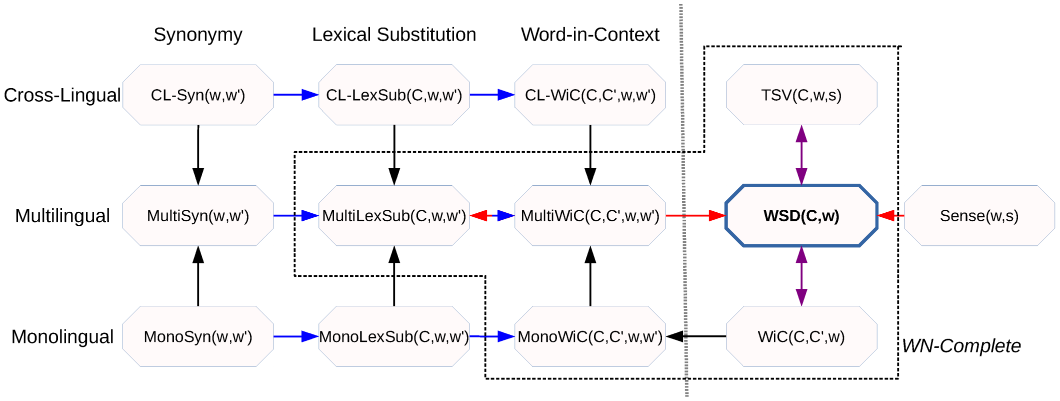

Figure 6.1: Taxonomy of problems in lexical semantics. Arrows indicate reducibility. The six wordnet-complete problems within the dotted area are equivalent, and all other problems in the taxonomy are reducible to them.

lexical semantics. In addition, we carry out a series of experiments that support the correctness of our theoretical findings.

6.2 Theoretical Formalization

In this section, we formally define the problems in our proposed taxonomy, and discuss the relationship between these theoretical problems and the computational tasks addressed in prior work.

6.2.1 Words

All semantic problems in Figure 6.1 take at least one word as a parameter. In our definitions, a *word* is not necessarily an orthographic word, but rather a triple consisting of a lemma, a part of speech, and a language. The problems are divided into three categories based solely on the language of the words (rather than contexts): *monolingual* (same language), *cross-lingual* (different languages), and *multilingual* (same or different languages). Thus, a multilingual problem can be seen as the union of the corresponding monolingual and cross-lingual problems. While this categorization theoretically admits “monolingual” problem instances consisting of a word in one language and a context in a different language, such instances are rare in practice.

6.2.2 Contexts and Concepts

Alternatively, we can categorize semantic problems according to the number of contexts which must be considered in each instance: zero, one, or two, respectively, in the leftmost three columns of Figure 6.1. Contexts are denoted by the variable names starting with C. We broadly define a *context* as a discourse (not necessarily a sentence) with a *focus*, which is a word or sequence of words that express a specific concept. Contexts that consist of the same discourse but differ in focus are considered distinct. The expression “a word expresses a concept given a context” signifies that the word can be used to refer to the concept that corresponds to the focus of that context. Note that the word itself is not required to occur in the context, or even match the language of the context.

For example, consider the context “bats live in caves” which disambiguates the word *bat* to its animal sense. The underlined word represents the focus of the context, which can be expressed by the words *bat* or its synonym *chiropteran*. The languages of the word and the context need not be the same. For example, the Spanish context “un murciélago entro en mi casa” disambiguates the English word *bat* as an animal rather than an instrument.

A lexical concept, or simply *concept*, refers to a discrete word meaning. A *concept gloss*, such as “flying nocturnal rodent,” is a special type of a context, in which the entire definition is the focus, and which uniquely determines the concept. We assume that the concept gloss C_s which defines the meaning of the concept s can be expressed in any language.

We assume the availability of complete sets of words (i.e., *lexicons*) and lexical concepts. The methods for creating such resources are beyond the scope of this work.

6.2.3 Monolexical Problems

We first define three problems that take a single word argument. We refer to these theoretical problems by the same acronyms as their corresponding computational tasks: WSD, TSV, and WiC.

Word sense disambiguation (WSD) is the task of classifying a word in context according to its sense, given an inventory of possible senses for each word. For each word, there is a one-to-one mapping between its senses and the concepts that it can express. We can therefore define the WSD problem more generally, to return a concept rather than a sense. This avoids the need for a predefined sense inventory for each word.

WSD(C, w) := "the concept which is expressed by the word w given the context C"

Note that this formulation does not require the word to occur in the context. By convention, the return value of the WSD predicate is *undefined* if the word is not meaningful given the context; for example, the English word *metre* does not express any concept given the Italian context "la metro di Roma è efficiente" ("the Rome metre is efficient"). In contrast, any binary predicate is assumed to return FALSE in such cases.

Target sense verification, or TSV (Breit et al., 2021), is the binary classification task of deciding whether a given word in a given context expresses a given sense. As with WSD, we define the TSV problem on concepts rather than senses. We assume that the concept s is represented by its gloss C_s.

TSV(C, w, s) := "the word w expresses the concept s given the context C"

The TSV problem can be viewed as a binary analogue of the WSD problem, such that the following equivalence holds:

$$\text{TSV}(C, w, s) \Leftrightarrow \text{WSD}(C, w) = s$$

The word-in-context task (WiC) is a binary classification task proposed by Pilehvar and Camacho-Collados (2019): given a pair of sentences, decide whether or not a word has the same meaning in both sentences. We define the corresponding WiC problem using concepts, on the basis of the concept-meaning hypothesis:

WiC(C_x, C_y, w) := "the word w expresses the same concept given the contexts C_x and C_y"

In Chapter 5, we demonstrated the equivalence of WiC, TSV, and WSD by pairwise reductions, which are denoted by purple arrows in Figure 6.1. In particular, the following formula specifies the reduction of WiC to WSD:

$$\text{WiC}(C_x, C_y, w) \Leftrightarrow \text{WSD}(C_x, w) = \text{WSD}(C_y, w)$$

6.2.4 Word-in-Context Problems

We now introduce a set of binary predicates which include WiC and its variants. We start with the most general problem of the set, MultiWiC, and then define MonoWiC, and CL-WiC as its special cases, in which the two words w_x and w_y are constrained to be either in the same or different languages, respectively.

> **MultiWiC**$(C_x, C_y, w_x, w_y) :=$ "the words w_x and w_y express the same concept given the contexts C_x and C_y, respectively"

The WiC problem defined in Section 6.2.3 is a special case of MonoWiC, in which $w_x = w_y$.

> **MonoWiC**(C_x, C_y, w_x, w_y):= "the words w_x and w_y from the same language express the same concept given the contexts C_x and C_y, respectively"

Martelli et al. (2021) extend the WiC task to include cross-lingual instances, which consist of a pair of contexts in different languages, in which the two focus words have the same meaning.[2] Our definition of the corresponding theoretical problem is similar:

> **CL-WiC**(C_x, C_y, w_x, w_y): "the words w_x and w_y from different languages express the same concept given the contexts C_x and C_y, respectively"

Clearly, any instance of MultiWiC is either an instance of MonoWiC or CL-WiC.

6.2.5 Lexical Substitution Problems

The next set of problems each involve a pair of words in a single context. These problems formalize the semantic task of lexical substitution (McCarthy and Nav-

[2] An instance was annotated as positive "if and only if the two target word occurrences were used with exactly the same **meaning** or, in other words, if, using a dictionary, the **definition** of the two target words was the same" (Martelli et al., 2021).

igli, 2007), and its different variants and settings, such as cross-lingual substitution (Mihalcea et al., 2010). Our definitions are more precise than conventional ones, as we define substitutes on the basis of identity of expressed concepts. By virtue of our concept-meaning hypothesis, the definitions formalize the notions of "meaning-preserving substitutions" and "correct translations" present in previous work. However, they are restricted to lexical substitutions, excluding compositional compounds and phrases.

> **MonoLexSub**(C, w_x, w_y) := "the words w_x and w_y from the same language express the same concept given the context C"

In other words, w_x and w_y are mutually substitutable given the context C. For example, MonoLexSub returns TRUE given $C =$ "the gist of the prosecutor's argument", $w_x =$ *core*, and $w_y =$ *heart*.

The CL-LexSub problem is a cross-lingual counterpart of MonoLexSub. The definition of CL-LexSub is the same as that of MonoLexSub, except that the two words are required to be in different languages. For example, MonoLexSub("she batted the ball", *bat*, *murciélago*) returns FALSE.

> **CL-LexSub**(C, w_x, w_y) := "the words w_x and w_y from different languages express the same concept given the context C"

Finally, we define a multilingual lexical substitution problem which generalizes MonoLexSub and CL-LexSub by removing their respective language constraints:

> **MultiLexSub**(C, w_x, w_y) := "the words w_x and w_y from any language(s) express the same concept given the context C"

While the goal of many conventional lexical substitution datasets is to produce sets of substitutes, these generative problems are reducible to the corresponding binary classification problems by iterating over the set of substitution candidates. More formally, the problem of generating lexical substitutes reduces to MultiLexSub by returning the set: $\{w \mid \text{MultiLexSub}(C, w_x, w)\}$.

6.2.6 Word Synonymy Problems

Our final set of semantic problems are defined on a pair of word lemmas, without any context parameters.

The MonoSyn predicate formalizes the relation of word synonymy in the monolingual setting. Given two words in the same language, it returns TRUE iff they are mutually substitutable in some context.

MonoSyn(w_x, w_y) := "the words w_x and w_y from the same language express the same concept in some context"

For example, MonoSyn(*core*, *heart*) is TRUE because there exist a contexts in which the two words express the same concept (c.f., Section 6.2.5). The MonoSyn problem formalizes the linguistic Substitution Test for synonymy: *w_x and w_y are synonyms if the meaning of a sentence that contains w_x does not change when w_y is substituted for w_x* (Murphy and Koskela, 2010).

We define the cross-lingual synonymy problem CL-Syn in a similar manner. The only difference with MonoSyn is that the two words are required to be from different languages.

CL-Syn(w_x, w_y) := "the words w_x and w_y from different languages express the same concept in some context"

The CL-Syn predicate corresponds to the relation of translational equivalence between words. Two words in different languages are translationally equivalent if there exists a context in which they are literal translations. For example, CL-Syn(*heart/EN*, *cœur/FR*) is TRUE because the two words are mutual translations given the context "the heart of the matter."

As with the other problem families, we unify MonoSyn and CL-Syn into a single predicate, MultiSyn, which places no constraints on the language of the given words:

MultiSyn(w_x, w_y) := "the words w_x and w_y from any language(s) express the same concept in some context"

MultiSyn is not only a generalization but also the union of the relations of synonymy and translational equivalence, which are represented by MonoLexSub and CL-LexSub, as postulated in Chapter 2.

6.3 Problem Reductions

Given an algorithm for a problem **Q**, a **P**-to-**Q** reduction solves an instance of a problem **P** by combining the solutions of one or more instances of **Q**. The reducibility of **P** to **Q** is denoted **P** $\leq$ **Q**. Mutual reductions of two problems to one another, i.e. **P** $\leq$ **Q** and **Q** $\leq$ **P**, demonstrate their equivalence.

In this section, we present several problem reductions, which constitute the main contribution of this chapter. The reductions are shown in Figure 6.1 by the directed arrows from **P** to **Q**. The black arrows denote the special cases, which immediately reduce to the more general problems. Taken together, the reductions establish the equivalence of six problems: WSD, TSV, WiC, MonoWiC, MultiWiC, and MultiLexSub. A method which solves any of these problems can be used to construct methods which solve the other problems by applying a sequence of reductions. As well, a method for one of those six problems can be used to solve any of the other problems in Figure 6.1, again via reductions.

6.3.1 *Syn $\leq$ *LexSub $\leq$ *WiC

We first present a set of six reductions, which are denoted by blue arrows in Figure 6.1. Each of the corresponding nine problems involves comparing the meanings of a pair of words, given some contexts.

The three lexical substitution problems defined in Section 6.2.5 can be viewed as special cases of the corresponding word-in-context problems, in which both contexts are identical. Succinctly:

$$\text{*LexSub}(C, w_x, w_y) \Leftrightarrow \text{*WiC}(C, C, w_x, w_y)$$

The asterisk in these and the following reductions can be replaced on both sides by “Mono”, “CL-”, or “Multi”. To reiterate, a cross-lingual problem explicitly assumes that the input words are in different languages, while a multilingual problem can accept inputs in the same or different languages.

The three word synonymy problems defined in Section 6.2.6 are reducible to the corresponding lexical substitution problems. In particular, to reduce MultiSyn to MultiLexSub, we search for a concept gloss C_s in which both words express the same concept. Succinctly:

$$\text{*Syn}(w_x, w_y) \Leftrightarrow \exists s : \text{*LexSub}(C_s, w_x, w_y)$$

The correctness of these six reductions follows from the fact that the (infinite) set of all contexts is partitioned into equivalence classes, each of which corresponds to a single concept.

6.3.2 Reductions to WSD

The reductions in the preceding section demonstrates that all theoretical problems defined in Section 6.2 can be reduced to MultiWiC. We next demonstrate that all those problems, including MultiWiC itself, can also be reduced to WSD. Thus, an algorithm that solves WSD would be sufficient to solve all other problems. For clarity, the nine reductions in this section are not shown explicitly in Figure 6.1, with the exception of the crucial MultiWiC-to-WSD reduction, denoted by a red arrow.

Given a method for solving WSD, we can solve any *WiC instance by checking whether the concepts expressed by the two words in the corresponding contexts are the same. This set of reductions generalize the WiC-to-WSD reduction (Chapter 5) to MonoWiC, CL-WiC, and MultiWiC:

$$\text{*WiC}(C_x, C_y, w_x, w_y) \Leftrightarrow \text{WSD}(C_x, w_x) = \text{WSD}(C_y, w_y)$$

Similarly, to solve any *LexSub instance, it is sufficient to check the identity of the concepts expressed by the two words in the given context:

$$\text{*LexSub}(C, w_x, w_y) \Leftrightarrow \text{WSD}(C, w_x) = \text{WSD}(C, w_y)$$

Finally, the word synonymy problems can be solved by searching for a concept which can be expressed by both words:

$$\text{*Syn}(w_x, w_y) \Leftrightarrow \exists s : \text{WSD}(C_s, w_x) = \text{WSD}(C_s, w_y)$$

The correctness of the reductions in this section follows directly from the concept-meaning hypothesis which underlies our theory.

6.3.3 MultiWiC $\leq$ MultiLexSub

We close this section by demonstrating that MultiWiC is reducible to MultiLexSub, which is denoted by a red arrow in Figure 6.1. This reduction, along with the reverse reduction presented in Section 6.3.1, establishes the equivalence between the two problems. Formally:

$$\text{MultiWiC}(C_x, C_y, w_x, w_y) \Leftrightarrow \text{MultiLexSub}(C_x, w_x, w_y) \wedge \text{MultiLexSub}(C_y, w_y, w_x) \wedge$$
$$\forall w : \text{MultiLexSub}(C_x, w_x, w) \Leftrightarrow \text{MultiLexSub}(C_y, w_y, w)$$

The first two terms on the right-hand side of the reduction formula test whether the two words are mutually substitutable in their respective contexts. The universal quantifier ensures that every substitute in one of the contexts is also an appropriate substitute in the other context, and vice versa.

The correctness of this reduction hinges on the assumption that there are no universal colexifications (stated and empirically supported by Bao et al. (2021)), i.e. that *for any pair of concepts, there exists some language which lexifies but does not colexify them.* In other words, there exists a language in which no word can express both concepts. Therefore, if the sets of contextual synonyms of w_x in C_x and w_y in C_y are identical, the concept expressed by the two word tokens must be the same.

In theory, the universal quantifier in the reduction formula is defined over all words in all languages. In practice, only the synonyms and translations of the two words need to be checked, and a smaller set of diverse languages may be sufficient to obtain good accuracy.

6.4 Relationship to Synsets

A wordnet is a theoretical construct which is composed of synonym sets, or *synsets*, such that each synset corresponds to a unique concept, and each sense of a given word corresponds to a different synset. Actual wordnets, such as Princeton WordNet (Miller, 1995), are considered to be imperfect implementations of the theoretical construct.

We define the following monolexical problem, which decides whether a given word can express a given concept:

Sense(w, s) := "the word w expresses the concept s in some context"

An algorithm for the Sense problem could be used to decide whether a given word belongs to the synset that corresponds to a given concept.

6.4.1 *Syn $\leq$ Sense $\leq$ WSD

The word synonymy problems defined in Section 6.2.6 are reducible to the Sense problem. Two words are synonyms if they both express the same concept in some context. In particular, to reduce MultiSyn to Sense, we search for a concept which can be expressed by both words.

$$\text{MultiSyn}(w_x, w_y) \Leftrightarrow \exists s : \text{Sense}(w_x, s) \wedge \text{Sense}(w_y, s)$$

A monolingual wordnet can be converted into a thesaurus, in which the entry for a given word consists of all of its synonyms. A bilingual wordnet can be converted into a translation dictionary, in which the entry for a given word consists of all its cross-lingual synonyms possibly grouped by sense, and accompanied by glosses.

Given a method for solving WSD, we can solve a Sense instance by checking whether the word expresses the concept given the context of its gloss. Formally:

$$\text{Sense}(w, s) \Leftrightarrow \text{WSD}(C_s, w) = s$$

The correctness of this reduction follows from the assumption that a concept gloss uniquely determines the concept. Under our definitions, given a concept gloss, the WSD predicate can only return the corresponding concept, and does so if and only if the given word can express that concept; otherwise the return value is undefined.

The reducibility of Sense to WSD implies that implementing the WSD predicate as it is defined in Section 6.2.3 would make it possible to construct synsets from nothing more than a list of concept glosses, as well as correct and expand existing wordnets to new domains and languages. In fact, any of the set of six

WSD-equivalent problems (Figure 6.1) could be used for these tasks; we therefore refer to them as *wordnet-complete* or *WN-complete*.

6.4.2 Substitution Lemma

The final proposition formalizes the relationship between synsets, senses, and lexical translations. It follows directly from the previously stated definitions, reductions, and assumptions.

$$\text{MultiLexSub}(C_x, w_x, w_y) \Leftrightarrow \text{Sense}(w_y, \text{WSD}(C_x, w_x))$$

The lemma provides a theoretical justification for methods that associate contextual lexical translations and synonyms with the synset identified by a WSD model. For example, BabelNet synsets are populated by translations of word instances that correspond to a given concept (Navigli and Ponzetto, 2010). Specifically, the existence of a translation pair (w_x, w_y) in a context C_x implies that w_y lexicalizes the concept expressed by w_x in C_x. Another example is the method of Luan et al. (2020), which leverages contextual translations to improve the accuracy of WSD.

6.5 Empirical Validation

In this section, we implement and test three principal reductions: MultiWiC to WiC, MultiWiC to WSD, and MultiLexSub to WSD. For each reduction, we reiterate its theoretical basis, describe our implementation, and discuss the results. We emphasize that the goal of our experiments is not challenging the state of the art, but rather empirically testing the reductions, and, by extension, the hypothesis they are based on. Since the resources used for the implementations are necessarily imperfect, and the systems are each designed and optimized for a different target task, the reductions are expected to produce much less accurate predictions on the existing benchmark datasets compared to state-of-the-art methods.

Our primary interest is in identifying any possible counter-examples to our concept-meaning hypothesis. However, it must be noted that the presence of a small number of such exceptions in the existing datasets does not invalidate the theory. On the other hand, the scarcity of counter-examples should not be interpreted as a *proof,* but rather as supporting evidence for the correctness of our theoretical claims.

6.5.1 Solving MultiWiC with WiC

We first empirically test the counter-intuitive proposition that a multilingual semantic task can be reduced to a set of monolingual instances. In particular, given a method for solving WiC, we can solve any MultiWiC instance by deciding whether there exists a concept such that both given words express the concept given their corresponding contexts and the concept gloss. Formally:

$$\text{MultiWiC}(C_x, C_y, w_x, w_y) \Leftrightarrow \exists s : \text{WiC}(C_x, C_s, w_x) \wedge \text{WiC}(C_y, C_s, w_y)$$

The correctness of this reduction follows from the assumption that a concept gloss uniquely disambiguates every word that can express the concept.

Implementation of the Reduction

In practice, instead of checking all possible concepts, we limit our search to concepts that can be expressed by either of the two words. For each such concept, we create two WiC instances, one in each language, using a gloss retrieved from a lexical resource, and translated, as needed, into the language of each instance. We then solve each of the created WiC instances using a model trained exclusively on WiC data in that language. The reduction returns TRUE iff both WiC instances are classified as positive.

We test the reduction on the English-French test set of the MCL-WiC shared task (Martelli et al., 2021), which contains 1000 English-French MultiWiC instances. The dataset is agnostic toward WordNet sense distinctions and annotations. We train the English WiC model on the English training and development sets (8k and 1k instances, respectively), and the French WiC model on the French development set (1k instances). The latter set is quite small, but we are not aware of any larger dedicated French WiC training data.

We create each WiC instance by prepending the input word, followed by a separator token, to each input context, including concept glosses. We retrieve concept glosses from BabelNet (Navigli and Ponzetto, 2010), using the Python API.[3] While English lemmas are provided in the dataset, French lemmas are not. We therefore lemmatize French words using the SpaCy FR_CORE_NEWS_MD model. Since BabelNet does not contain French glosses for all concepts, we generate them by translating the first English gloss in BabelNet using the OPUS-MT-EN-FR model from Helsinki NLP.[4]

We train our English and French WiC models using LIORI (Davletov et al., 2021). All training was completed in under eight hours on two NVIDIA GeForce RTX 3090 GPUs. With the default hyper-parameter settings, the models obtain the accuracy of 87.0% and 73.3% on the English and French monolingual test sets, respectively. This is lower than the 91.1% and 86.4% results reported by Davletov et al. (2021). We attribute this to our use of smaller, purely monolingual training data, which is in line with our theoretical reduction. Based on these numbers, we estimate the probability of a pair of WiC instances being both correctly classified as $0.870 * 0.733 = 0.638$.

Results and Discussion

Our implementation correctly classifies 631 out of the 1000 instances in the test set. This is very close to the estimate computed in the previous section, which suggests that our reduction is approximately as reliable as our imperfect resources and systems allow.

We manually analyzed a random sample of 50 MultiWiC classification errors. For each of the 25 false negatives, LIORI returned FALSE for *all* sentence pairs in either English (12 instances), French (8 instances), or both languages (5 instances). Each instance could be explained by either a LIORI error, or a missing sense in BabelNet. For the 25 false positives, we identified one or more incorrect positive WiC classifications. The final false positive was caused by an incorrect tokenization of the target word in the MCL-WiC dataset: The target word is *disordered*, however in the given context this token is actually part of the compound adjective *mentally*

[3] https://babelnet.org/guide\#python

[4] https://huggingface.co/Helsinki-NLP

disordered. As a result, LIORI's classifications pertaining to *disordered* were not reliable, leading to a spurious false positive classification.

Since all errors can be attributed to the systems and resources, they constitute no evidence against the correctness of our reduction. On the other hand, these results support our theoretical finding that multilingual problems can be reduced to monolingual problems. This in turn supports our methodology of grounding lexical semantics in the expression of language-independent concepts.

6.5.2 Solving MultiWiC with WSD

In this section, we test our MultiWiC-to-WSD reduction. In doing so, we generalize the WiC-to-WSD reduction in Chapter 5 to multiple words and languages. Given a MultiWiC instance, we apply a WSD system to each context-word pair, and classify it as positive iff both words are tagged with the same BabelNet synset:

$$\text{MultiWiC}(C, C', w, w') \Leftrightarrow \text{WSD}(C, w) = \text{WSD}(C', w')$$

Implementation of the Reduction

Our system of choice is AMuSE-WSD (Orlando et al., 2021). It provides pre-trained WSD models for a diverse set of languages, and handles all stages of the WSD pipeline, including tokenization, lemmatization, and part-of-speech tagging. We apply the provided AMUSE-LARGE-MULTILINGUAL-CPU model, with all other parameters left at their default values.

As in Chapter 5, we estimate an upper-bound on the performance of our reduction, using analogous notation and formulae. For the expected accuracy of English and non-English WSD, we use the English-ALL and XL-WSD accuracy results reported by the AMuSE-WSD authors, 0.739 and 0.673. This estimation method also depends on the average number of senses per target word. Per the BabelNet API[5], an average MCL-WiC target word has 14 senses. The resulting overall accuracy estimate is 0.752, which is the average of 0.539 and 0.965 for the positive and negative MultiWiC instances, respectively.

[5] https://babelnet.org/guide\#python

Results and Discussion

The results on the MCL-WiC test sets range from 51.8% on English-Arabic to 55.1% on English-French. While the estimate in the previous section is substantially higher, it does not take into account tokenization errors and missing senses in BabelNet. On the English-French dataset, we found that false negatives outnumber false positives by a factor of six; the accuracy is 22.8% and 87.4% on the positive and negative MultiWiC instances, respectively.

For our manual analysis, we randomly selected 25 false positives and 25 false negatives produced by our implementation on the English-French test set. In 41 of the 50 cases, we determined the cause of the incorrect MultiWiC classification to be an incorrect sense returned by AMuSE-WSD for one or both target words. In addition, 7 of the 50 cases represent tokenization errors. One MultiWiC instance, which involves English *reflected* and French *consignée*, is most likely a MCL-WiC annotation error. The final error is attributable to a sense missing from BabelNet, which prevents AMuSE-WSD from considering it as a candidate. Specifically, it is the "administer" sense of the verb *dispense* (as in "dispense justice"), which can be found in the Merriam-Webster Online Dictionary.[6]

Since manual analysis yields no counter-examples to our theory, we interpret the results as empirical support for this reduction, and, by extension, our taxonomy of semantic tasks, and the hypothesis on which it is based.

6.5.3 Solving MultiLexSub with WSD

In the final experiment, we test the MultiLexSub-to-WSD reduction derived in Section 6.3.2:

$$\text{MultiLexSub}(C, w, w') \Leftrightarrow \text{WSD}(C, w) = \text{WSD}(C, w')$$

The overall method is similar to that of Guo and Diab (2010), but using our precise binary formulation of lexical substitution.

[6] https://www.merriam-webster.com/dictionary/dispense

Implementation of the Reduction

We use the dataset from the SemEval 2010 shared task on cross-lingual lexical substitution (Mihalcea et al., 2010), which consists of a trial set of 300 instances, and a test set of 1000 instances. Each instance consists of an English sentence which includes a single target word and a list of Spanish gold substitutes provided by annotators.

Since our formulation of lexical substitution is binary rather than generative or ranking-based, we convert each of the SemEval instances into a pair of binary instances: one positive and one negative. For the positive instance, we take the first Spanish substitute, the one that was most frequently suggested by the annotators. For the negative instance, we randomly select a Spanish word from the set of all substitutes in the dataset for that English target word, provided that it is not among the gold substitutes for that specific instance. If there is no such substitute, we instead choose a random Spanish word from the dataset.

For each binary instance created in this way, we create two WSD instances using a simple template: *'w' as in 'C'*, where w is the target word, and C is the context. We obtain the context for the Spanish word by translating the English context via Helsinki NLP's OPUS-MT-EN-ES model. We return a positive MultiLexSub classification iff AMuSE-WSD assigns the same BabelNet synset ID to both English and Spanish target words.

Our procedure for estimating the expected accuracy of our reduction is the same as in Section 6.5.2. The only difference is the average number of senses per word, which in this case is 23, yielding an estimated accuracy of 75.8%.

Results and Discussion

The binary classification accuracy of our implementation on 2000 MultiLexSub instances created from the SemEval test set is 63.2%, which is substantially below the estimate in the previous section. This can be partially explained by a relatively high number of tokenization errors in the test set. We again observe a strong bias toward negative classification: the results on the positive and negative instances are 27.1% and 99.3% accuracy, respectively. Because of this, we selected only positive instances for our error analysis.

We manually analyzed a sample of 50 randomly-selected false negatives from the test set. In 44 of the 50 cases, the cause of the misclassification was an AMuSE-WSD error (on English in 30 cases, on Spanish in at least 14). Some of those errors may be caused by an imperfect translation of the English context, or a missing BabelNet sense of the Spanish gold substitute. In 5 cases, the English input was incorrectly tokenized; for example, the compound noun *key ring* was split into two word tokens, with one instance having *ring* as its focus. The final case likely involves an annotation error in the SemEval dataset: *campo* as a translation of *field* given the context of "effective law enforcement in the field."

We conclude that all incorrect classifications can be attributed to a resource or system used by our implementation, and thus none of them represents a counterexample to our hypothesis.

6.6 Conclusion

Starting from basic assumptions about the expression of concepts by words in context, we have developed consistent formulations of thirteen different problems in lexical semantics. We have shown that a "wordnet-complete" subset of these tasks can each be used to solve any of the others via reduction. These problems can be used to construct, correct, or expand multilingual synonym sets, the building blocks of important linguistic resources such as WordNet and BabelNet. We believe that this work will lead to a greater understanding of lexical semantics and its underlying linguistic phenomena, as well as new applications and better interpretation of empirical results. Based on our theory, we intend to develop methods for constructing fully explainable and interpretable linguistic resources.

Chapter 7

Conclusion

This document began with the following thesis statement: **An empirically-validated theory of sense, synonymy, translation, and lexical concepts yields an improved understanding of lexical resources, methods and tasks, including novel evidence for linguistic hypotheses, and a taxonomy of semantic problems.** Each of the five preceding chapters presented novel research contributions which demonstrate this statement.

In Chapter 2, we formulated a first-of-its-kind theory of lexical semantics. The theory relates senses, synonymy, translation, and wordnet synsets to the unifying notion of lexical concepts. We showed that this theory can be used to construct a formal model of wordnets, and argued for the soundness of this theory on the basis of two experiments, one of which demonstrated that our theory can be applied to the construction of multilingual wordnets. We also argued for concept universality, a linguistic hypothesis postulating a universal set of concepts which may be expressed, by words or phrases, in any language.

In Chapters 3 and 4, we examined the relation between word senses and translations in the context of modern natural language processing. Chapter 3 presented a novel, formal treatment of homonymy and polysemy. By considering the semantic relatedness between the concepts expressed by word senses, we provided novel evidence for four hypotheses relating to lexical translation, discourses, collocations, and sense clusters. Chapter 4 demonstrated that the connection between word senses and translations rarely conforms to an ideal one-to-one mapping, but that nevertheless, this relation can be exploited to improve the performance of even a contemporary WSD system. Taken together, these two chapters show how a theory-driven approach to lexical semantics can allow us to test hypotheses and discover useful new methods.

In Chapters 5 and 6, we extended our theoretical analysis from semantic phenomena – senses, synonymy, translations, etc. – to semantic tasks themselves. Taken together, we argued in these chapters that the space of word meaning can effectively be discretized on the basis of lexical concepts. This novel concept-meaning hypothesis implies that human judgements of word meaning will tend to align with discrete word senses, as defined by our theory in terms of concepts. On the basis of this hypothesis, we formulated a series of problem reductions involving thirteen semantic tasks, leading to a first-of-its-kind taxonomy of semantic tasks, culminating in an argument for the class of wordnet-complete problems. A series of experiments yielded no substantial evidence against our hypothesis, with apparent exceptions being generally attributable to errors in the methods and resources we used in our implementations. This demonstrates that our concept-based theory of lexical semantics can indeed be used to study semantic tasks, further supporting the thesis statement.

A brief note on the limitations of this thesis: While we made an effort to include multilingual datasets in our experiments, our error analysis was often limited to languages of the Indo-European family (e.g. English, French, Spanish, etc.). In addition, it is possible to question some of the assumptions made in our theory, which should be kept in mind when considering our work. For example, we assume that, for each content word token in a discourse, there exists a single concept which that word is intended by the sender to express, regardless of whether it appears unambiguous to the receiver. However, unlike in mathematics, theoretical assumptions may not always hold in practice; for example, puns often exploit multiple meanings of a word for humorous effect. While such cases are not frequently considered in lexical semantics, we can expect exceptions to almost any assumption or conclusion regarding human languages.

Considering the field of lexical semantics as a whole, it is not sufficient to simply strive for ever-greater performance on benchmark datasets. Rather, scientific understanding is necessary to ensure that the resources and results we produce remain interpretable and open to analysis and improvement. Beyond this thesis, future work could apply methods for wordnet-complete problems to the construction of wordnets given only basic resources, such as text corpora and machine readable dictionaries and thesauri. This could be particularly beneficial for low-resource languages, those with little or no representation in standard knowledge bases, and, more broadly, could move the field away from the "English first" (or perhaps "WordNet first") paradigm which has dominated lexical semantics since its inception.

Another avenue for future work would be theory-driven analysis of language models and contextualized representations. Our taxonomy of semantic tasks provides insight into the relative hardness of various problems: if a solution to problem Q can be used to construct a solution to task P, then P is no harder than Q. We postulate that, when language models are tested on various tasks, the relative hardness of those tasks should be taken into consideration.

Finally, it must be noted that, while this thesis has focused on lexical semantics, the study of semantics does not end at the meaning of words. The study of meaning at the level of sentences, and even entire documents and discourses, is very much relevant to modern natural language processing, and, we believe, is just as much in need of careful theoretical treatment and analysis. How, for instance, would our insights into lexical translation generalize to the task of translating entire sentences? How could the link between word meanings and discrete concepts be applied to improve the state of semantic parsing? These and other questions are likewise fertile ground for future work.

In closing, this thesis has demonstrated that, despite promising recent advances in the tools, methods, and resources applied to semantic tasks, and despite the rapid pace of improvement on increasingly challenging benchmarks, the study of lexical semantics is far from complete. The need for scientific understanding, theoretical analysis, explainability, and predictability are greater than ever. Even with the long and venerable history of lexical semantics, dating to the earliest days of natural language processing, there is still much to be done.

References

Eneko Agirre, Oier Lopez de Lacalle, Christiane Fellbaum, Shu-Kai Hsieh, Maurizio Tesconi, Monica Monachini, Piek Vossen, and Roxanne Segers. 2010. SemEval-2010 task 17: All-words word sense disambiguation on a specific domain. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 75–80, Uppsala, Sweden, July. Association for Computational Linguistics.

Marianna Apidianaki and Li Gong. 2015. LIMSI: Translations as source of indirect supervision for multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 298–302, June.

Marianna Apidianaki. 2008. Translation-oriented word sense induction based on parallel corpora. In *Language Resources and Evaluation (LREC)*.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 597–604, June.

Mohit Bansal, John DeNero, and Dekang Lin. 2012. Unsupervised translation sense clustering. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 773–782, June.

Hongchang Bao, Bradley Hauer, and Grzegorz Kondrak. 2021. On universal colexifications. In *Proceedings of the 11th Global Wordnet Conference (GWC2021)*, pages 1–7, January.

Edoardo Barba, Luigi Procopio, Niccolo Campolungo, Tommaso Pasini, and Roberto Navigli. 2020. MuLaN: Multilingual label propagation for word sense disambiguation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3837–3844, 7.

Edoardo Barba, Tommaso Pasini, and Roberto Navigli. 2021a. ESC: Redesigning WSD with extractive sense comprehension. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4661–4672, June.

Edoardo Barba, Luigi Procopio, Caterina Lacerra, Tommaso Pasini, and Roberto Navigli. 2021b. Exemplification modeling: Can you give me an example, please? In *Proceedings of 30th International Joint Conference on Artificial Intelligence (IJCAI 2021)*, August.

Edoardo Barba, Luigi Procopio, and Roberto Navigli. 2021c. ConSeC: Word sense disambiguation as continuous sense comprehension. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1492–1503, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Barend Beekhuizen, Sasa Milic, Blair C Armstrong, and Suzanne Stevenson. 2018. What company do semantically ambiguous words keep? Insights from distributional word vectors. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, pages 1347–1352.

Luisa Bentivogli and Emanuele Pianta. 2000. Looking for lexical gaps. In *Proceedings of the ninth EURALEX International Congress*, pages 8–12. Stuttgart: Universität Stuttgart.

Luisa Bentivogli and Emanuele Pianta. 2005. Exploiting parallel texts in the creation of multilingual semantically annotated resources: The MultiSemCor Corpus. *Natural Language Engineering*, 11(3):247–261.

Michele Bevilacqua, Marco Maru, and Roberto Navigli. 2020. Generationary or "how we went beyond word sense inventories and learned to gloss". In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7207–7221, Online, November. Association for Computational Linguistics.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006–1017, July.

Terra Blevins, Mandar Joshi, and Luke Zettlemoyer. 2021. FEWS: Large-scale, low-shot word sense disambiguation with the dictionary. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 455–465, April.

Giulia Bonansinga and Francis Bond. 2016. Multilingual sense intersection in a parallel corpus with diverse language families. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 44–49.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria, August.

Francis Bond, Timothy Baldwin, Richard Fothergill, and Kiyotaka Uchimoto. 2012. Japanese SemCor: A sense-tagged corpus of Japanese. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, pages 56–63.

Anna Breit, Artem Revenko, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. WiC-TSV: An evaluation benchmark for target sense verification of words in context. In *Proceedings of the 16th Conference of* *the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1635–1645, April.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1991. Word-sense disambiguation using statistical methods. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 264–270, June.

Susan Windisch Brown. 2008. Choosing sense distinctions for WSD: Psycholinguistic evidence. In *Proceedings of ACL-08: HLT, Short Papers*, pages 249–252, June.

Marine Carpuat. 2009. One translation per discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 19–27, June.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 33–40, June.

John Woldemar Cowan. 1997. *The complete Lojban language*, volume 15. Logical Language Group.

Ido Dagan, Alon Itai, and Ulrike Schwall. 1991. Two languages are more informative than one. In *29th Annual Meeting of the Association for Computational Linguistics*, pages 130–137, June.

Bharath Dandala, Chris Hokamp, Rada Mihalcea, and Razvan Bunescu. 2013. Sense clustering using Wikipedia. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP) 2013*, pages 164–171.

Adis Davletov, Nikolay Arefyev, Denis Gordeev, and Alexey Rey. 2021. LIORI at SemEval-2021 Task 2: Span prediction and binary classification approaches to word-in-context disambiguation. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 780–786, August.

Gerard De Melo and Gerhard Weikum. 2009. Towards a universal wordnet by learning from combined evidence. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 513–522.

Gerard De Melo, Collin F Baker, Nancy Ide, Rebecca J Passonneau, and Christiane Fellbaum. 2012. Empirical comparisons of MASC word sense annotations. In *LREC*, pages 3036–3043.

Claudio Delli Bovi, Jose Camacho-Collados, Alessandro Raganato, and Roberto Navigli. 2017. EuroSense: Automatic harvesting of multilingual sense annotations from parallel text. In *Proceedings of the 55th Annual Meeting of* *the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 594–600, July.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186.

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 255–262, July.

Mona Diab. 2004. The feasibility of bootstrapping an Arabic WordNet leveraging parallel corpora and an English wordnet. In *Proceedings of the Arabic Language Technologies and Resources, NEMLAR, Cairo*.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, June.

Helge Dyvik. 2004. Translations as semantic mirrors: From parallel corpus to wordnet. In *Advances in corpus linguistics*, pages 309–326. Brill Rodopi.

Philip Edmonds and Scott Cotton. 2001. Senseval-2: Overview. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 1–5, July.

Philip Edmonds and Graeme Hirst. 2002. Near-synonymy and lexical choice. *Computational Linguistics*, 28(2):105–144.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Darja Fišer and Benoît Sagot. 2015. Constructing a poor man's wordnet in a resource-rich world. *Language Resources and Evaluation*, 49(3):601–635.

Alexandre François. 2008. Semantic maps and the typology of colexification: Intertwining polysemous networks across languages. *From Polysemy to Semantic change: Towards a Typology of Lexical Semantic Associations*, 163-215.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992a. One sense per discourse. In *Proceedings of the workshop on Speech and Natural Language*, pages 233–237.

William A Gale, Kenneth W Church, and David Yarowsky. 1992b. Using bilingual materials to develop word sense disambiguation methods. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, volume 112. Citeseer.

Weiwei Guo and Mona Diab. 2010. COLEPL and COLSLM: An unsupervised WSD approach to multilingual lexical substitution, Tasks 2 and 3 SemEval 2010. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 129–133, Uppsala, Sweden, July. Association for Computational Linguistics.

Rohan Gupta, Jay Mundra, Deepak Mahajan, and Ashutosh Modi. 2021. MCL@IITK at SemEval-2021 Task 2: Multilingual and cross-lingual word-in-context disambiguation using augmented data, signals, and transformers. In *Proceedings of the Fifteenth Workshop on Semantic Evaluation*.

Amir Ahmad Habibi, Bradley Hauer, and Grzegorz Kondrak. 2021. Homonymy and polysemy detection with multilingual information. In *Proceedings of the 11th Global Wordnet Conference*, pages 26–35.

Christian Hadiwinoto, Hwee Tou Ng, and Wee Chung Gan. 2019. Improved word sense disambiguation using pre-trained contextualized word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5297–5306, Hong Kong, China, November. Association for Computational Linguistics.

Bradley Hauer and Grzegorz Kondrak. 2020a. One homonym per translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7895–7902.

Bradley Hauer and Grzegorz Kondrak. 2020b. Synonymy = translational equivalence. *arXiv preprint arXiv:2004.13886*.

Bradley Hauer and Grzegorz Kondrak. 2021. One sense per translation. *arXiv preprint arXiv:2106.06082*.

Bradley Hauer and Grzegorz Kondrak. 2022. WiC = TSV = WSD: On the equivalence of three semantic tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2478–2486, Seattle, United States, July. Association for Computational Linguistics.

Bradley Hauer and Grzegorz Kondrak. 2023. Taxonomy of problems in lexical semantics. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9833–9844, Toronto, Canada, July. Association for Computational Linguistics.

Bradley Hauer, Garrett Nicolai, and Grzegorz Kondrak. 2017. Bootstrapping unsupervised bilingual lexicon induction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 619–624, April.

Bradley Hauer, Yixing Luan, and Grzegorz Kondrak. 2019. You shall know the most frequent sense by the company it keeps. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, pages 208–215. IEEE.

Bradley Hauer, Amir Ahmad Habibi, Yixing Luan, Arnob Mallik, and Grzegorz Kondrak. 2020a. Low-resource G2P and P2G conversion with synthetic training data. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 117–122, Online, July. Association for Computational Linguistics.

Bradley Hauer, Amir Ahmad Habibi, Yixing Luan, Arnob Mallik, and Grzegorz Kondrak. 2020b. UAlberta at SemEval-2020 task 2: Using translations to predict cross-lingual entailment. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 263–269, Barcelona (online), December.

Bradley Hauer, Hongchang Bao, Arnob Mallik, and Grzegorz Kondrak. 2021a. UAlberta at SemEval-2021 Task 2: Determining sense synonymy via translations. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 763–770, August.

Bradley Hauer, Grzegorz Kondrak, Yixing Luan, Arnob Mallik, and Lili Mou. 2021b. Semi-supervised and unsupervised sense annotation via translations. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 504–513, Held Online, September. INCOMA Ltd.

Bradley Hauer, Grzegorz Kondrak, Yixing Luan, Arnob Mallik, and Lili Mou. 2021c. Semi-supervised and unsupervised sense annotation via translations. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 504–513, Held Online, September. INCOMA Ltd.

Bradley Hauer, Seeratpal Jaura, Talgat Omarov, and Grzegorz Kondrak. 2022. UAlberta at SemEval 2022 task 2: Leveraging glosses and translations for multilingual idiomaticity detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 145–150, Seattle, United States, July. Association for Computational Linguistics.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3507–3512, November.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for word sense disambiguation: An evaluation study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 897–907, August.

Nancy Ide and Yorick Wilks. 2007. Making sense about sense. In *Word sense disambiguation*, pages 47–73. Springer.

Nancy Ide. 2000. Cross-lingual sense determination: Can it work? *Computers and the Humanities*, 34(1-2):223–234.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing*. Prentice Hall, 2nd edition.

Eric Kafe. 2023. Mapping wordnets on the fly with permanent sense keys.

Adam Kilgarriff. 1997. I don't believe in word senses. *Computers and the Humanities*, 31(2):91–113.

Robert Krovetz. 1998. More than one sense per discourse. *NEC Princeton NJ Labs., Research Memorandum*, 23.

Oi Yee Kwong. 2018. Translation equivalence and synonymy: Preserving the synsets in cross-lingual wordnets. In *Proceedings of the 9th Global WordNet Conference (GWC 2018)*, page 201.

Els Lefever and Véronique Hoste. 2010. SemEval-2010 Task 3: Cross-lingual word sense disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 15–20, July.

Els Lefever, Véronique Hoste, and Martine De Cock. 2011. ParaSense or how to use parallel corpora for word sense disambiguation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 317–322, June.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26. ACM.

Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2020. SenseBERT: Driving some sense into BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4656–4667, July.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016*, pages 923–929. European Language Resources Association.

Frederick Liu, Han Lu, and Graham Neubig. 2018. Handling homographs in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1336–1345.

Daniel Loureiro and Jose Camacho-Collados. 2020. Don't neglect the obvious: On the role of unambiguous words in word sense disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3514–3520, November.

Daniel Loureiro and Alípio Jorge. 2019. LIAAD at SemDeep-5 challenge: Word-in-context (WiC). In *Proceedings of the 5th Workshop on Semantic Deep Learning (SemDeep-5)*, pages 1–5, Macau, China, August.

Daniel Loureiro, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2020. Analysis and evaluation of language models for word sense disambiguation.

Daniel Loureiro, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. Analysis and evaluation of language models for word sense disambiguation. *Computational Linguistics*, 47(2):387–443.

Daniel Loureiro, Alípio Mário Jorge, and Jose Camacho-Collados. 2022. LMMS reloaded: Transformer-based sense embeddings for disambiguation and beyond. *Artificial Intelligence*, 305.

Yixing Luan, Bradley Hauer, Lili Mou, and Grzegorz Kondrak. 2020. Improving word sense disambiguation with translations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4055–4065.

John Lyons. 1995. *Linguistic semantics: An introduction*. Cambridge University Press.

Arnob Mallik and Grzegorz Kondrak. 2023. Correcting sense annotations using wordnets and translations. In *12th International Global Wordnet Conference*, Donostia / San Sebastian, Basque Country, January. Global Wordnet Association.

Massimiliano Mancini, Jose Camacho-Collados, Ignacio Iacobacci, and Roberto Navigli. 2017. Embedding words and senses together via joint knowledge-enhanced training. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 100–111, August.

Federico Martelli, Najla Kalach, Gabriele Tola, and Roberto Navigli. 2021. SemEval-2021 Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation (MCL-WiC). In *Proceedings of the Fifteenth Workshop on Semantic Evaluation (SemEval-2021)*.

David Martinez and Eneko Agirre. 2000. One sense per collocation and genre/topic variations. In *2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 207–215, October.

Abelardo Carlos Martínez Lorenzo, Marco Maru, and Roberto Navigli. 2022. Fully-Semantic Parsing and Generation: the BabelNet Meaning Representation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1727–1741, Dublin, Ireland, May. Association for Computational Linguistics.

Marco Maru, Simone Conia, Michele Bevilacqua, and Roberto Navigli. 2022. Nibbling at the hard core of word sense disambiguation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4724–4737, Dublin, Ireland, May. Association for Computational Linguistics.

Rowan Hall Maudslay and Simone Teufel. 2022. Metaphorical polysemy detection: Conventional metaphor meets word sense disambiguation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 65–77, Gyeongju, Republic of Korea, October. International Committee on Computational Linguistics.

Diana McCarthy and Roberto Navigli. 2007. SemEval-2007 Task 10: English lexical substitution task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, Prague, Czech Republic, June. Association for Computational Linguistics.

Diana McCarthy, Marianna Apidianaki, and Katrin Erk. 2016. Word sense clustering and clusterability. *American Journal of Computational Linguistics*, 42(2):245–275, June.

John Philip McCrae, Alexandre Rademaker, Ewa Rudnicka, and Francis Bond. 2020. English WordNet 2020: Improving and extending a WordNet for English using an open-source methodology. In *Proceedings of the LREC 2020 Workshop on Multimodal Wordnets (MMW2020)*, pages 14–19, May.

Igor Mel'čuk. 2013. *Semantics: From meaning to text*, volume 2. John Benjamins.

Rada Mihalcea, Ravi Sinha, and Diana McCarthy. 2010. SemEval-2010 Task 2: Cross-lingual lexical substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 9–14, Uppsala, Sweden, July.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*. Technical report.

George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to WordNet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.

George A. Miller, Claudia Leacock, Randee I. Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 303–308.

George A Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

George A. Miller. 1998. Foreword. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages xv–xxii. MIT Press.

Andrea Moro and Roberto Navigli. 2015. SemEval-2015 Task 13: Multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, pages 288–297, June.

M. Lynne Murphy and Anu Koskela. 2010. *Key terms in semantics*. London: Continuum.

Saeed Najafi, Bradley Hauer, Rashed Rubby Riyadh, Leyuan Yu, and Grzegorz Kondrak. 2018. Combining neural and non-neural methods for low-resource morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 116–120, Brussels, October. Association for Computational Linguistics.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, July.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. Semeval-2013 task 12: Multilingual word sense disambiguation. In *Proceedings of the Seventh International Workshop on Semantic Evaluation*, pages 222–231, June.

Roberto Navigli. 2006. Meaningful clustering of senses helps boost word sense disambiguation performance. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 105–112, July.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10.

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

Hwee Tou Ng, Bin Wang, and Yee Seng Chan. 2003. Exploiting parallel texts for word sense disambiguation: An empirical study. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 455–462, July.

Made Nindyatama Nityasya, Haryo Wibowo, Alham Fikri Aji, Genta Winata, Radityo Eko Prasojo, Phil Blunsom, and Adhiguna Kuncoro. 2023. On “scientific debt” in NLP: A case for more rigour in language model pre-training research. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8554–8572, Toronto, Canada, July. Association for Computational Linguistics.

Michael Ogezi, Bradley Hauer, Talgat Omarov, Ning Shi, and Grzegorz Kondrak. 2023. UAlberta at SemEval-2023 task 1: Context augmentation and translation for multilingual visual word sense disambiguation. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2043–2051, Toronto, Canada, July. Association for Computational Linguistics.

Talgat Omarov and Grzegorz Kondrak. 2023. Grounding the lexical substitution task in entailment. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2854–2869, Toronto, Canada, July. Association for Computational Linguistics.

Riccardo Orlando, Simone Conia, Fabrizio Brignone, Francesco Cecconi, and Roberto Navigli. 2021. AMuSE-WSD: An all-in-one multilingual system for easy word sense disambiguation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 298–307, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Simone Papandrea, Alessandro Raganato, and Claudio Delli Bovi. 2017. SupWSD: A flexible toolkit for supervised word sense disambiguation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 103–108, Copenhagen, Denmark, September. Association for Computational Linguistics.

Tommaso Pasini and Roberto Navigli. 2017. Train-O-Matic: Large-scale supervised word sense disambiguation in multiple languages without manual training data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 78–88, September.

Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. XL-WSD: An extra-large and cross-lingual evaluation framework for word sense disambiguation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13648–13656.

Tommaso Pasini. 2021. The knowledge acquisition bottleneck problem in multilingual word sense disambiguation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 4936–4942.

Tommaso Petrolito and Francis Bond. 2014. A survey of WordNet annotated corpora. In Heili Orav, Christiane Fellbaum, and Piek Vossen, editors, *Proceedings of the Seventh Global WordNet Conference*, pages 236–245.

Emanuele Pianta and Luisa Bentivogli. 2004. Knowledge intensive word alignment with KNOWA. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 1086–1092, Geneva, Switzerland, aug 23–aug 27. COLING.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: developing an aligned multilingual database. In *First international conference on global WordNet*, pages 293–302.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of NAACL-HLT*, pages 1267–1273.

Mohammad Taher Pilehvar, Jose Camacho-Collados, Roberto Navigli, and Nigel Collier. 2017. Towards a seamless integration of word senses into downstream NLP applications. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1857–1869, July.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. Semeval-2007 task-17: English lexical sample, srl and all words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations*, pages 87–92, June.

Krunoslav Puškar. 2015. Esperanto (s) en perspektivo? Croatian Esperantists on the international language Esperanto. *Interdisciplinary Description of Complex Systems: INDECS*, 13(2):322–341.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110, April.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. XL-WiC: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206, November.

Alessandro Raganato, Iacer Calixto, Jose Camacho-Collados, Asahi Ushio, and Mohammad Taher Pilehvar. 2023. SemEval-2023 Task 1: Visual word sense disambiguation (Visual-WSD)). In *FORTHCOMING: Proceedings of the Seventeenth Workshop on Semantic Evaluation (SemEval-2023).*

Philip Resnik and David Yarowsky. 1997. A perspective on word sense disambiguation methods and their evaluation. In *Tagging Text with Lexical Semantics: Why, What, and How?*, pages 79–86.

Philip Resnik and David Yarowsky. 1999. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(2):113–133.

Ewa Rudnicka, Marek Maziarz, Maciej Piasecki, and Stan Szpakowicz. 2012. A strategy of mapping Polish Wordnet onto Princeton Wordnet. In *Proceedings of COLING 2012: Posters*, pages 1039–1048, Mumbai, India, December. The COLING 2012 Organizing Committee.

Benoît Sagot and Darja Fišer. 2008. Building a free French wordnet from multilingual resources. *OntoLex 2008 Programme*.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2019. Just "OneSeC" for producing multilingual sense-annotated data. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 699–709.

Helmut Schmid. 1999. Improvements in part-of-speech tagging with an application to German. In *Natural language processing using very large corpora*, pages 13–25. Springer.

Helmut Schmid. 2013. Probabilistic part-of-speech tagging using decision trees. In *New methods in language processing*, page 154.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational linguistics*, 24(1):97–123.

Rion Snow, Sushant Prakash, Daniel Jurafsky, and Andrew Y. Ng. 2007. Learning to merge word senses. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Benjamin Snyder and Martha Palmer. 2004. The English all-words task. In Rada Mihalcea and Phil Edmonds, editors, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43, July.

Angus Stevenson. 2010. *Oxford dictionary of English*. Oxford University Press, USA.

Kaveh Taghipour and Hwee Tou Ng. 2015. One million sense-tagged instances for word sense disambiguation and induction. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 338–344.

Simone Tedeschi, Johan Bos, Thierry Declerck, Jan Hajič, Daniel Hershcovich, Eduard Hovy, Alexander Koller, Simon Krek, Steven Schockaert, Rico Sennrich, Ekaterina Shutova, and Roberto Navigli. 2023. What's the meaning of superhuman performance in today's NLU? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12471–12491, Toronto, Canada, July. Association for Computational Linguistics.

Zdeňka Urešová, Eva Fučíková, Eva Hajičová, and Jan Hajič. 2018. Synonymy in bilingual context: The CzEngClass lexicon. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2456–2469, Santa Fe, New Mexico, USA, August.

Jason Utt and Sebastian Padó. 2011. Ontology-based distinction between polysemy and homonymy. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*, pages 265–274.

Sven van den Beukel and Lora Aroyo. 2018. Homonym detection for humor recognition in short text. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 286–291.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Piek Vossen. 1996. Right or wrong: Combing lexical resources in the EuroWordNet project. In *M. Gellerstam, J. Jarborg, S. Malmgren, K. Noren, L. Rogstrom, CR Papmehl, Proceedings of Euralex-96, Goetheborg, 1996*, pages 715–728. Vrije Universiteit.

Piek Vossen. 2004. EuroWordNet: a multilingual database of autonomous and language-specific wordnets connected via an inter-lingual index. *International Journal of Lexicography*, 17(2):161–173.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *NeurIPS*.

Warren Weaver. 1949. Translation. *Machine Translation of Languages*.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. OntoNotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*.

Xuchen Yao, Benjamin Van Durme, and Chris Callison-Burch. 2012. Expectations of word sense in parallel corpora. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 621–625.

David Yarowsky. 1993. One sense per collocation. In *Proceedings of the workshop on Human Language Technology*, pages 266–271.

David Yarowsky. 1995. Unsupervised WSD rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, Massachusetts Institute of Technology, Cambridge, MA*.

Tatu Ylonen. 2022. Wiktextract: Wiktionary as machine-readable structured data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1317–1325, Marseille, France, June. European Language Resources Association.

Hee Suk Yoon, Eunseop Yoon, John Harvill, Sunjae Yoon, Mark Hasegawa-Johnson, and Chang-Dong Yoo. 2022. Smsmix: Sense-maintained sentence mixup for word sense disambiguation. In *The 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*. EMNLP.

Zhi Zhong and Hwee Tou Ng. 2010. It Makes Sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations*, pages 78–83, July.