Lexical Semantic Analysis in Natural Language Text

Nathan Schneider

CMU-LTI-14-001

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213
www.lti.cs.cmu.edu

<u>Thesis Committee:</u>

Noah A. Smith *(chair)*, Carnegie Mellon University
Chris Dyer, Carnegie Mellon University
Eduard Hovy, Carnegie Mellon University
Lori Levin, Carnegie Mellon University
Timothy Baldwin, University of Melbourne

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
In Language and Information Technologies

Lexical Semantic Analysis in Natural Language Text

Nathan Schneider

Language Technologies Institute ◇ School of Computer Science
Carnegie Mellon University

September 30, 2014

Submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN LANGUAGE AND INFORMATION TECHNOLOGIES

Abstract

Computer programs that make inferences about natural language are easily fooled by the often haphazard relationship between words and their meanings. This thesis develops **Lexical Semantic Analysis** (LxSA), a general-purpose framework for describing word groupings and meanings in context. LxSA marries comprehensive linguistic annotation of corpora with engineering of statistical natural language processing tools. The framework does not require any lexical resource or syntactic parser, so it will be relatively simple to adapt to new languages and domains.

The contributions of this thesis are: a formal *representation* of lexical segments and coarse semantic classes; a well-tested linguistic *annotation* scheme with detailed guidelines for identifying multi-word expressions and categorizing nouns, verbs, and prepositions; an English web *corpus* annotated with this scheme; and an open source NLP system that *automates* the analysis by statistical sequence tagging. Finally, we motivate the applicability of lexical semantic information to sentence-level language technologies (such as semantic parsing and machine translation) and to corpus-based linguistic inquiry.

I, Nathan Schneider, do solemnly swear that this thesis is my own work, and that everything therein is, to the best of my knowledge, true, accurate, and properly cited, so help me Vinken.*

*Except for the footnotes that are made up. If you cannot handle jocular asides, please recompile this thesis with the `\jnote` macro disabled.

To my parents, who gave me language.

In memory of Charles “Chuck” Fillmore, one of the great linguists.

Acknowledgments

As a fledgling graduate student, one is confronted with obligatory words of advice, admonitions, and horror stories from those who have already embarked down that path. At the end of the day,[2] though, I am forced to admit that my grad school experience has been perversely *happy*. Some of that can no doubt be attributed to my own bizarre priorities in life, but mostly it speaks to the qualities of those around me at CMU, who have been uniformly brilliant, committed, enthusiastic, and generous in their work.

First and foremost, I thank my Ph.D. advisor, Noah Smith, for serving as a researcher role model, for building and piloting a resplendent research group (ARK), and for helping me steer my research while leaving me the freedom to develop my own style. I thank the full committee (Noah, Chris, Lori, Ed, and Tim), not just for their thoughtful consideration and feedback on the thesis itself, but for their teaching and insights over the years that have influenced how I think about language and its relationship to technology.

It was my fellow ARK students that had the biggest impact on my day-to-day experiences at CMU. Mike Heilman, Shay Cohen, Dipanjan Das, Kevin Gimpel, André Martins, Tae Yano, Brendan O'Connor, Dani Yogatama, Jeff Flanigan, David Bamman, Yanchuan

[2] "day" being a euphemism for 6 years

Sim, Waleed Ammar, Sam Thomson, Lingpeng Kong, Jesse Dodge, Swabha Swayamdipta, Rohan Ramanath, Victor Chahuneau, Daniel Mills, and Dallas Card—it was an honor. My officemates in GHC 5713—Dipanjan, Tae, Dani, Waleed, and (briefly) Sam and Lingpeng—deserve a special shout-out for making it an interesting place to work. It was a privilege to be able to work with several different ARK postdocs—Behrang Mohit, (pre-professor) Chris Dyer, and Fei Liu—and always a pleasure to talk with Bryan Routledge at group meetings.

But oh, the list doesn't stop there. At CMU I enjoyed working with amazing annotators—Spencer Onuffer, Henrietta Conrad, Nora Kazour, Mike Mordowanec, and Emily Danchik—whose efforts made this thesis possible, and with star undergraduates including Desai Chen and Naomi Saphra. I enjoyed intellectual exchanges (in courses, reading groups, collaborations, and hallway conversations) with Archna Bhatia, Yulia Tsvetkov, Dirk Hovy, Meghana Kshirsagar, Chu-Cheng Lin, Ben Lambert, Marietta Sionti, Seza Doğruöz, Jon Clark, Michael Denkowski, Matt Marge, Reza Bosagh Zadeh, Philip Gianfortoni, Bill McDowell, Kenneth Heafield, Narges Razavian, Nesra Yannier, Wang Ling, Manaal Faruqui, Jayant Krishnamurthy, Justin Cranshaw, Jesse Dunietz, Prasanna Kumar, Andreas Zollmann, Nora Presson, Colleen Davy, Brian MacWhinney, Charles Kemp, Jacob Eisenstein, Brian Murphy, Bob Frederking, and many others. And I was privileged to take courses with the likes of Noah, Lori, Tom Mitchell, William Cohen, Yasuhiro Shirai, and Leigh Ann Sudol, and to have participated in projects with Kemal Oflazer, Rebecca Hwa, Ric Crabbe, and Alan Black.

I am especially indebted to Vivek Srikumar, Jena Hwang, and everyone else at CMU and elsewhere who has contributed to the preposition effort described in ch. 5.

The summer of 2012 was spent basking in the intellectual warmth of USC's Information Sciences Institute, where I did an internship under the supervision of Kevin Knight, and had the good fortune to cross paths with Taylor Berg-Kirkpatrick (and his cats), Daniel Bauer, Bevan Jones, Jacob Andreas, Karl Moritz Hermann, Christian Buck, Liane Guillou, Ada Wan, Yaqin Yang, Yang Gao, Dirk Hovy, Philipp Koehn, and Ulf Hermjakob. That summer was the beginning of my involvement in the AMR design team, which includes Kevin and Ulf as well as Martha Palmer, Claire Bonial, Kira Griffitt, and several others; I am happy to say that the collaboration continues to this day.

The inspiration that pushed me to study computational linguistics and NLP in the first place came from teachers and mentors at UC Berkeley, especially Jerry Feldman and Dan Klein. I am happy to have remained connected to that community during my Ph.D., and to have learned a great deal about semantics from the likes of Collin Baker, Miriam Petruck, Nancy Chang, Michael Ellsworth, and Eve Sweetser.

Finally, I thank my family for being there for me and putting up with my eccentricities. The same goes for my friends mentioned above, as well as from the All University Orchestra (Smitha Prasadh, Tyson Price, Deepa Krishnaswamy), from Berkeley linguistics (Stephanie Shih, Andréa Davis, Kashmiri Stec, Cindy Lee), and from NLP conferences.

This thesis would not have been possible without the support of research funding sources;[3] all who have contributed to the smooth functioning of CMU SCS and LTI, especially Jaime Carbonell (our fearless leader) and all the LTI staff; as well as Pittsburgh eating

[3] The research at the core of this thesis was supported in part by NSF CAREER grant IIS-1054319, Google through the Reading is Believing project at CMU, DARPA grant FA8750-12-2-0342 funded under the DEFT program, and grant NPRP-08-485-1-083 from the Qatar National Research Fund (a member of the Qatar Foundation). The statements made herein are solely the responsibility of the authors.

establishments, especially Tazza d'Oro, the 61C Café, and Little Asia.

There are undoubtedly others that I should have mentioned above, so I hereby acknowledge them.

If in this document contains any errors, omissions, bloopers, dangling modifiers, wardrobe malfunctions, etc., it is because I screwed up. It is emphatically *not* the fault of my committee or reviewers if a bug in my code caused low numbers to round up, high numbers to round down, or expletives to be inserted in phonologically suspect syllable positions.

Contents

Computers are getting smarter all the time: scientists tell us that soon they will be able to talk to us. (By "they" I mean "computers": I doubt scientists will ever be able to talk to us.)

Dave Barry, *Dave Barry's Bad Habits*: "The Computer: Is It Terminal?"

CHAPTER 1

Setting the Stage

The seeds for this thesis were two embarrassing realizations.[1]

The first was that, despite established descriptive frameworks for syntax and morphology, and many vigorous contenders for relational and compositional semantics, I did not know of any general-purpose linguistically-driven computational scheme to represent the contextual meanings of *words*—let alone resources or algorithms for putting such a scheme into practice.

My embarrassment deepened when it dawned on me that I knew of no general-purpose linguistically-driven computational scheme for even *deciding* which groups of characters or tokens in a text qualify as meaningful words!

Well. Embarrassment, as it turns out, can be a constructive

[1] In the Passover Haggadah, we read that two Rabbis argued over which sentence should begin the text, whereupon everyone else in the room must have cried that four cups of wine were needed, stat.

This thesis is similar, only instead of wine, you get footnotes.

motivator. Hence this thesis, which chronicles the why, the what, and the how of **analyzing** (in a general-purpose linguistically-driven computational fashion) the lexical semantics of natural language text.

The intricate communicative capacity we know as "language" rests upon our ability to learn and exploit conventionalized associations between *patterns* and *meanings*. When in *Annie Hall* Woody Allen's character explains, "My raccoon had hepatitis," English-speaking viewers are instantly able to identify sound patterns in the acoustic signal that resemble sound patterns they have heard before. The patterns are *generalizations* over the input (because no two acoustic signals are identical). For one acquainted with English vocabulary, they point the way to basic meanings like the concepts of 'raccoon' and 'hepatitis'—we will call these *lexical* meanings—and the *grammatical* patterns of the language provide a template for organizing words with lexical meanings into sentences with complex meanings (e.g., a pet raccoon having been afflicted with an illness, offered as an excuse for missing a Dylan concert). Sometimes it is useful to contrast denoted *semantics* ('the raccoon that belongs to me was suffering from hepatitis') and *pragmatics* ('…which is why I was unable to attend the concert') with meaning inferences ('my *pet* raccoon'; 'I was unable to attend the concert *because I had to nurse* the ailing creature'). Listeners draw upon vast stores of world knowledge and contextual knowledge to deduce coherent interpretations of necessarily limited utterances.

As effortless as this all is for humans, understanding language production, processing, and comprehension is the central goal of the field of linguistics, while automating human-like language behavior in computers is a primary dream of artificial intelligence. The field of natural language processing (NLP) tackles the language au-

tomation problem by decomposing it into subproblems, or *tasks*; NLP tasks with natural language text input include grammatical analysis with linguistic representations, automatic knowledge base or database construction, and machine translation.[2] The latter two are considered *applications* because they fulfill real-world needs, whereas automating linguistic analysis (e.g., syntactic parsing) is sometimes called a "core NLP" task. Core NLP systems are those intended to provide modular functionality that could be exploited for many language processing applications.

This thesis develops linguistic descriptive techniques, an English text dataset, and algorithms for a core NLP task of analyzing the lexical semantics of sentences in an integrated and general-purpose way. My hypothesis is that the approach is conducive to rapid high-quality human annotation, to efficient automation, and to downstream application.

A synopsis of the task definition, guiding principles, methodology, and contributions will serve as the entrée of this chapter, followed by an outline of the rest of the document for dessert.

1.1 Task Definition

We[3] define **Lexical Semantic Analysis** (LxSA) to be the task of segmenting a sentence into its lexical expressions, and assigning semantic labels to those expressions. By **lexical expression** we mean a word or group of words that, intuitively, has a "basic" meaning or function. By **semantic label** we mean some representation of

[2] I will henceforth assume the reader is acquainted with fundamental concepts, representations, and methods in linguistics, computer science, and NLP. An introduction to computational linguistics can be found in Jurafsky and Martin (2009).

[3] Since the time of Aristotle, fledgling academics have taken to referring themselves in the plural in hopes of gaining some measure of dignity. Instead, they were treated as grad students.

the expression's contextual meaning, selected from a predefined categorization scheme.

The flavor of LxSA pursued here incorporates multiword expression identification (to determine the lexical segmentation) and supersense classification (to choose labels for noun, verb, and preposition expressions). For example, Groucho Marx's famous aphorism is analyzed thusly:[4]

(1) a.

Time	flies	like	an	arrow	.
TIMEˆ	MOTIONˇ	MANNER›		ARTIFACTˆ	

b.

Fruit flies	like	a	banana	.
ANIMALˆ	COGNITIONˇ		FOODˆ	

The lexical segmentation is indicated by underlining: every token belongs to exactly one lexical expression.[5] Observe that *fruit flies* is analyzed as a multiword expression because it is deemed to have a sufficiently "atomic" meaning (see ch. 3 for criteria). This expression and other nouns, verbs, and prepositions in the sentence receive supersense labels. There are 26 supersense categories for nouns, 15 for verbs, and 70 for prepositions (see ch. 4: p. 65 and ch. 5: table 5.1 on p. 102). These coarse senses provide a measure of word sense disambiguation (contrast the two senses of *like* in (1)), even beyond the part-of-speech disambiguation (e.g., *liking a post on Facebook* would involve a COMMUNICATION sense of the verb). Supersenses also group together semantically related tokens: *like*, *prefer*, *think*, *decide*, etc. can also function as COGNITION verbs.

[4]Supersenses are part-of-speech–specific. To avoid visual clutter, we render parts of speech as symbols: ˇ for verbs, ˆ for nouns, and › for prepositions.

[5]Ch. 3 introduces a distinction between *strong* and *weak* multiword expressions. The definition of "lexical expression" assumed here disregards weak groupings.

1.2 Approach

To build a system that automatically performs lexical semantic analysis, we adopt a workflow of human annotation and supervised machine learning. The research described in this thesis employs the following **methodologies**:

1. **Representation:** Designing the formal scheme for encoding analyses. We specify a space of possible multiword expression groupings and a mapping from lexical expression tokens to possible semantic labels.

2. **Annotation:** Devising a usable annotation scheme and applying it to a corpus with human annotators. This includes providing linguistic guidelines (categories, definitions, criteria, and examples), as well as an annotation interface and quality control procedures. Our result is a fully annotated 56,000-word corpus of English reviews from the web.

3. **Automation:** Developing a system that performs the analysis given new sentences. We implement a discriminative sequence tagger, train it on the labeled corpus, and evaluate its accuracy compared to human annotations on held-out data.

While this endeavor requires a certain amount of human expertise and intuition, we can test our hypotheses in part by quantifying several aspects of the process, including: the number of sentences that seem to violate our representational constraints; the degree to which annotators agree with one another when working independently; the extent to which system output on held-out data matches the gold standard; the runtime of the system; and, ultimately, the impact of lexical semantic analyses on performance measures for subsequent tasks.

1.3 Guiding Principles

We believe LxSA is worth pursuing as a general-purpose NLP task for several reasons:

- **Coverage & Informativeness:** Our approach integrates three existing representations/tasks, or **components**: MWE identification, noun and verb supersense tagging, and preposition classification. All of these can be independently motivated and are associated with existing resources. The depth and coverage of semantic labels lies somewhere between named entity annotation (limited to a small proportion of tokens) and fine-grained word sense annotation (broad-coverage in principle, but expensive to produce). We hypothesize that interactions between lexical grouping and kinds of lexical meaning can be exploited in a joint model for improved accuracy, and that predicting a single analysis that is consistent between the facets of meaning will be more useful to applications than multiple possibly inconsistent layers of analysis.

- **Corpus-comprehensiveness:** Ideally, all of the text in a corpus should be analyzed so corpus statistics can be maximally useful to machine learning algorithms. In our data, all sentences have been annotated in full for multiword expressions, and all expressions meeting simple syntactic criteria (headed by a noun, verb, or preposition) have been annotated with supersenses. This contrasts with resources whose annotations are limited to certain types (e.g., particular high-frequency words, or expressions realizing concepts from a domain ontology).

- **Annotateability:** In our experience, it takes some training—but not an advanced linguistics degree—to learn the annotation schemes as we have formulated them. The annotation process is fairly rapid, as it involves mostly local (rather than long-distance or structural) decisions; uses a relatively small, interpretable label vocabulary; and does not require reference to external resources or layers apart from the tokenized sentence.

- **Universality:** The principle of identifying and classifying "semantic words" should apply crosslinguistically. Without making light of the typological differences between languages that would affect the methodology (e.g., a polysynthetic language would likely require *splitting* words into lexical morphemes), some kind of mismatch between units of sound or writing and units of meaning is expected to be universal, along with many of the categories in our supersense inventory.

- **Robustness:** Our annotation scheme is reasonably domain-general, though the guidelines and even the supersense inventory could be customized to meet the needs of a particular domain. The cost of coarse LxSA annotation in a new domain (on which new domain-specific models could be trained) should not be prohibitive.

- **Simplicity:** Formally, the representation has straightforward properties and well-formedness conditions. This matters for verifying the structural consistency of annotations, measuring inter-annotator agreement, and establishing a search space for automated tools.

- **Accuracy & Efficiency:** Computationally, the LxSA framework permits leveraging and extending well-known efficient and accurate supervised discriminative sequence modeling techniques.

- **Evaluability:** Because the representation permits a closed set of labels and grouping operations, the similarity between two analyses of a text can be quantified automatically.

1.4 Contributions

The contributions of this thesis are primarily methodological and practical rather than theoretical. There are no radical new ideas about computational lexical semantics in this thesis. Rather, several previous lines of work are woven together into a novel framework; the prior approaches are revised, expanded, and integrated into what we hope is a helpful conceptual paradigm within the broader landscape of computational semantics, as well as a useful practical tool in the NLP toolbox. Through trial and error, we have developed an annotation scheme that can be applied rapidly with acceptable inter-coder agreement, and shown that statistical models trained on the annotated data obtain far superior performance to heuristic lexicon lookup procedures. The artifacts produced in this work—datasets, annotation guidelines, software—are shared publicly for the benefit of further research on this and related tasks.

1.5 Organization

Following some background on computational lexical semantics tasks and techniques in ch. 2, we will turn to the core contributions of the thesis. Their primary mode of organization is methodological; the secondary mode is by analysis component:

		COMPONENTS		
		MWEs	Nouns & Verbs	Prepositions
METHODOLOGIES	I Representation & Annotation	Ch. 3	Ch. 4	Ch. 5
	II Automation	Ch. 6	Ch. 7	

Ch. 8 concludes with a discussion of broader issues, future work, and the prospects of applying lexical semantic analysis to downstream tasks.

"Good morning!" said Bilbo, and he meant it. The sun was shining, and the grass was very green. But Gandalf looked at him from under long bushy eyebrows that stuck out further than the brim of his shady hat.

"What do you mean?" he said. "Do you wish me a good morning, or mean that it is a good morning whether I want it or not; or that you feel good this morning; or that it is a morning to be good on?"

J.R.R. Tolkien, *The Hobbit*

CHAPTER 2

General Background: Computational Lexical Semantics

This chapter:

- Gives a high-level overview of several areas of computational semantics
- Introduces WordNet, word sense disambiguation, and named entity recognition
- Contrasts WSD-style and NER-style representations with respect to granularity
- Introduces linear models and algorithms for multi-class classification and sequence tagging in the context of lexical semantics

2.1 The Computational Semantics Landscape

Human language facilitates communication of just about any idea that can be imagined—and each language does so in its own peculiar way and with a finite set of symbols. Accordingly, part of the enterprise of building a scientific understanding of human language is understanding understanding of language. The study of linguistic meaning is called **semantics**.[1] There are many nuances and facets to meaning, and different semantic frameworks approach these from different angles.

Computational semantics, which aims to enable computers to detect aspects of meaning in language and to encode formally represented ideas as language, is similarly diverse. While it is impossible to do justice here to such a broad (and growing) field, it is worth mentioning a few of the paradigms in use:

- **Lexical semantics and ontologies:** The question of how to model entities and concepts and their relationship to one another and to language. As this area is directly relevant to this thesis, we discuss it in some detail below.

- **Grammaticalization semantics:** Several recent studies have been concerned with isolating aspects of linguistic meaning/function that are conveyed by grammatical categories and constructions, where languages differ in the mappings between these functions and their morphological and syntactic categories. Examples in NLP for English include semantic classification for tense/aspect (Reichart and Rappoport, 2010; Friedrich and Palmer, 2014), modality (Prabhakaran et al., 2012), definiteness functions (Bhatia et al., 2014), core vs. non-core arguments (Abend and Rappoport, 2010), argument structure constructions (Hwang et al., 2010b), and preposition senses/functions (O'Hara and Wiebe, 2003; Hovy et al., 2010; Srikumar and Roth, 2013b, *inter alia*; see ch. 5 for details).

- **Relational semantics:** This covers various attempts to model the nature of links between (usually lexically-associated) concepts in a sentence or discourse. Canonical NLP tasks include semantic role labeling (Gildea and Jurafsky, 2002; Palmer et al., 2010), relational semantic parsing (Tratz and Hovy, 2011; Das et al., 2014; Flanigan et al., 2014; Oepen et al., 2014), and coreference resolution (Stede, 2011, ch. 3).

- **Logical semantics:** An outgrowth of analytic philosophy, these approaches (sometimes called "formal semantics" in linguistics) represent concepts as predicates in formal logic and seek to describe the linguistic correlates of compositional operations that allow for predicates to be combined to form complex meanings. The logical expressions allow for truth relationships to be deduced, possibly with regard to a world model or database. In NLP, logic parsing (a variety of semantic parsing) seeks to produce logical forms for sentences, much like (and often in tandem with) syntactic parsing (e.g., Zelle and Mooney, 1996; Asudeh and Crouch, 2001; Bos et al., 2004; Zettlemoyer and Collins, 2005; Copestake et al., 2005).

- **Deep meaning and reasoning systems:** These are most closely associated with the label "artificial intelligence," and involve substantial human-like reasoning about the world, with natural language as the input or output. Such systems are of-

[1] The semantics of individual words and utterances in isolation is sometimes distinguished from meaning that requires wider communicative context (**pragmatics**). We do not need to reach the theoretical question of whether a line can be drawn between semantics and pragmatics, but the approaches taken in this thesis generally treat sentences in isolation.

ten associated with applications such as question answering, computer-assisted dialogue, and language-directed robotics (e.g., Carbonell, 1978; Narayanan, 1999; Branavan et al., 2010; Tellex et al., 2011; Ferrucci, 2012). They may use representations based on expert knowledge, mined from language data, or grounded in the physical/sensory domain.

2.2 Lexical Semantic Categorization Schemes

The contributions of this thesis belong to the area of **lexical semantics**, i.e., accounting for natural language words (**lexical items**) and their individual meanings. The inventory of lexical items available to speakers of a language, whether in the abstract or documented in dictionary form, is known as the **lexicon**. Lexicons hold information about word **types**; instances in context are called **tokens**. For example, a lexicon may record that the word type *seal* is polysemous (has multiple **senses**), and it might therefore be useful to disambiguate which of those senses is meant by a particular token in context. Lexical semantics includes both the study of the organization of the lexicon, and the study of how words convey meaning in context.

In this thesis, we will propose/adapt categorization schemes for lexical items and apply those categories (manually or automatically) in corpora. A primary consideration in developing a categorization is **granularity**. This is true in linguistics whether the categorization is grammatical (Croft, 2001, ch. 2) or semantic. When it comes to categorizing the meanings of lexical items, there are two major traditions in NLP. These are illustrated in figure 2.1. Traditionally, **word sense disambiguation (WSD)** is concerned with choosing among multiple senses of a word in a lexicon given a use of the word in context. The semantic representation adds information by refining the word into multiple **lexicalized** senses (figure 2.1a). **Named entity**

(a) lexicalized *seal*`.n`: `02` 'impression-making device', `09` 'kind of marine mammal'
(b) hybrid {ARTIFACT: {`seal.n.02`: *seal*, *stamp*}, . . . }, {ANIMAL: {`seal.n.09`: *seal*}, {`tasmanian_devil.n.01`: *Tasmanian devil*}, . . . }
(c) unlexicalized {ARTIFACT: *seal*, *stamp*, . . . }, {ANIMAL: *seal*, *Tasmanian devil*, *Burmese python*, . . . }

Figure 2.1: Categorization schemes for two senses of the noun *seal* and related concepts.

recognition (NER), on the other hand, is concerned with marking and classifying proper names, most of which will not be listed in a lexicon; in this way the task is **unlexicalized** and contributes information by grouping together multiple lexical items that belong to the same (coarse) semantic class.

The following sections will elaborate on those traditions and introduce machine learning techniques that can be applied to automatically categorize lexical meanings in text.

2.3 Lexicons and Word Sense Classification

This section provides a brief introduction to fundamentals of word sense lexicons and word sense disambiguation that are essential to understanding the thesis. (For a recent survey of corpora and tools for word senses, see Agirre et al., 2013.)

2.3.1 WordNet

Princeton **WordNet** (Fellbaum, 1998, `http://wordnet.princeton.edu/`) is a free and open source computational semantic network of the lexicon of English. A graph-structured database of lexical concepts organized into **synsets** (synonym sets), it includes natural

The	materials	for	compromise	are	at hand	:	The	Nation	,	Walter Lippman
DT	NN	IN	NN	VB	JJ	P	DT	NN	P	NNP
	material.2		compromise.1	be.3	at_hand.1			OTHER		PERSON

and	other	sober	commentators	(	see	Alan Clark	on	p.	367	)	have	spelled
CC	JJ	JJ	NN	P	RB	NNP	IN	NN	CD	P	VBP	VB
	other.1	sober.2	commentator.1		see.1	PERSON						spell_out.1

them	out	again and again	.
PRP	IN	RB	P
		again_and_again.1	

Figure 2.2: A sentence from SemCor with its part-of-speech and lexical sense annotations. Note the coarse labels for named entities lacking a specific entry in WordNet.

language descriptions and examples, and several types of taxonomic links between concepts (such as inheritance and part–whole relations). As of version 3.0, 118,000 synsets account for 155,000 lexical entries of nouns (including some proper names), verbs, adjectives, and adverbs.

Figure 2.1b is a flattened, partial view of the taxonomy of the WordNet lexicon. This approach both groups and refines lexical items in mapping them to synsets and defining groupings over synsets. WordNet is fundamentally lexicalized: every semantic category is associated with at least one lexical item.

WordNet additionally defines categorizations of noun and verb senses that are known as "supersenses"; we introduce these in §4.2. In our systems, we make use of WordNet and several other available lexical resources, as discussed in §6.4.2.1 and §7.3.2.

Princeton WordNet has been imitated or adapted in many other languages: a list of these projects and related resources can be found at http://globalwordnet.org.

2.3.2 SemCor

SemCor (Miller et al., 1993) is a 360,000 word sense-tagged subset of the Brown Corpus (Kučera and Francis, 1967) that was created as part of the development of WordNet. Miller et al. contrast two approaches to developing a lexicon and sense-tagged corpus: a "targeted" approach, traditional in lexicography, of considering one word type at a time to develop a sense inventory and label all instances in a corpus with the appropriate sense—we will call this a **type-driven** strategy; and a "sequential" (in our terms, **token-driven**) approach which proceeds token by token in a corpus, labeling each with an existing sense or revising the sense inventory as necessary. This second approach was preferred for constructing SemCor. Miller et al. observe that the token-by-token strategy naturally prioritizes corpus coverage. Nearly all of SemCor's content words are tagged with a fine-grained WordNet sense: figure 2.2 shows an example annotated sentence. Named entities not in WordNet (most of them) were tagged with a coarse class.

2.3.3 Beyond WordNet

Though WordNet and similar resources record relationships such as taxonomic inheritance between senses, they do not offer an account of how a listener can infer details of the meaning that depend on multiple senses being exploited in combination—e.g., they can explain that *have a banana* can decompose into *have* 'consume' + *banana* 'long yellow edible fruit', and similarly for *have a milkshake* and *have a cigarette*, but not that consuming a banana entails *eating* it, consuming a milkshake entails *drinking* it, consuming a cigarette entails *smoking* it, and so forth. WordNet-style lexicons also do not explain word choice constraints/preferences, e.g., that *heavy rain* is idiomatic, while *big rain* may seem marked or nonnative—as descriptors of intensity, *heavy* and *big* are simply not on an equal footing with respect to *rain*. Generative Lexicon Theory (Pustejovsky, 1998) and the theory of lexical functions (Mel'čuk, 1998) posit richer structure to lexical entries in order to explain their combinatorial behavior.

2.3.4 Word Sense Disambiguation as Classification

Assuming we have a list of known senses for one or more ambiguous words, we can set out to build algorithms that will choose the one that is most contextually appropriate. This is the classic **word sense disambiguation (WSD)** task (see Agirre and Edmonds, 2006 and Navigli, 2009 for comprehensive surveys). WSD is a type of discrete multi-class **classification** problem, where every input (such as a word within a sentence) is associated with a desired output, or **label**, to be predicted automatically. Mathematically, we can represent a classifier as a function $h(x)$ that, when presented with an input x, chooses a label $\hat{y} \in \mathcal{Y}$ (for now, we assume the set of labels $\mathcal{Y}$ is discrete, finite, and predetermined) as the prediction. For datasets where the "true" label, y^*, is known for every input, the accuracy of the classifier can be estimated as the proportion of inputs for which $h(x) = y^*$.

Classifiers can use various characteristics of the input that may be predictive of particular labels; the input–output combinations that factor into classification decisions are known as the classifier's **features**. For example, a word token's **distributional context** (in terms of the words that appear nearby) is known to be an important cue of its sense. If *seal* occurs near the word *aquatic* in a text, chances are it is being used in its animal sense. These cues could be hardcoded into the classification algorithm; or, they can be specified at an abstract level (e.g., "all words up to 5 words away from

the word being classified") and the predictive ones for each label learned from data. In particular, **supervised** learning algorithms mine labeled examples for statistical associations between inputs and outputs. Because of the complexities of natural language, data-driven semantic classifiers are generally much more robust than non-statistical (**rule-based**) systems.

For purposes of this thesis, we focus on **linear** classifiers, which model the prediction function $h(x)$ as:

$$h(x) = \arg\max_{y} score(x, y; \boldsymbol{\theta}) \tag{2.1}$$

where

$$score(x, y; \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \mathbf{g}(x, y) = \sum_{j=1}^{|\boldsymbol{\theta}|} \theta_j \cdot g_j(x, y) \tag{2.2}$$

I.e., the classifier chooses a label that maximizes a real-valued scoring function that quantifies the compatibility between x and y as a weighted sum of feature values $g_j(x, y)$.[2] The weighting $\boldsymbol{\theta}$ of the features is what is learned from data: thus, the weights are the **parameters** of the model. In our example, the weight linking context word *aquatic* to the animal sense of *seal* will receive a large positive weight if the feature is a good predictor in the training data. Of course, even if *aquatic* is contextually present, the classification function aggregates evidence from all its features, so it is possible that enough of the other active features will influence the classifier to choose some other label.

Many learning algorithms estimate parameters for linear classifiers given labeled data. The perceptron (Freund and Schapire, 1999) is one such algorithm; Smith (2011) discusses others and elucidates

[2] For discrete data, most of these feature values are binary: the characteristics described in the feature (e.g., a particular word appearing nearby) either apply for the current instance, or not.

how they effectively solve different optimization problems. For this thesis, we require *structured* learning and classification algorithms, which will be introduced in the next section.

2.4 Named Entity Recognition and Sequence Models

2.4.1 NER

Named entity recognition (NER) is the task of detecting and categorizing named entities (primarily proper names) in text.[3] The precise set of entity categories of interest varies with different formulations of the task (and annotated corpora), but the set of categories is typically small: one canonical scheme categorizes names as PERSON, ORGANIZATION, LOCATION, or MISCELLANEOUS (Tjong Kim Sang and De Meulder, 2003); others (e.g., Sekine et al., 2002; Weischedel and Brunstein, 2005; Grouin et al., 2011) consist of dozens of types.[4] NER therefore uses an unlexicalized grouping scheme analogous to figure 2.1c.

2.4.2 Chunking

Many instances of names contain multiple words: thus, detecting such named entities requires reasoning across spaces. A **chunking** representation encodes how sequence elements (tokens) group together into units. The most popular flavor, **BIO chunking**, accomplishes this by assigning each token one of three tags: *B* indicates that the token begins a chunk; *I* ("inside") indicates that it continues a multi-token chunk; and *O* ("outside") indicates that it is not a part

[3] See Nadeau and Sekine (2007) for a general survey of NER and Agirre et al. (2013) for a survey of resources.

[4] **Named entity disambiguation**, by contrast, is the task of resolving instances of a name to a canonical entry in an entity database.

of any chunk (Ramshaw and Marcus, 1995). The distinction between *B* and *I* allows for a boundary between adjacent chunks. Only contiguous chunks are allowed by this representation (a constraint that we relax in §6.3). This representation facilitates statistical models that make token-level predictions, though most commonly in a joint fashion.[5]

For tasks such as NER, in-chunk tags are commonly decorated with a class label categorizing the chunk: for example, "non-initial word of a PERSON chunk" can be denoted as $_I$PERSON, and this is only permitted to follow $_B$PERSON or $_I$PERSON. When a statistical model is used to predict the tags (and therefore the chunking), the decoding algorithm is constrained to only consider compatible tag bigrams. With C classes, the number of tags is $2C+1$, and the number of legal token tag bigrams is $2C^2+5C+1$. At each time step the Viterbi algorithm considers all tag bigrams, so decoding time is linear in the number of possible bigrams and also linear in the length of the sentence.

2.4.3 Structured Perceptron

The linear classifier described in §2.3.4 can be modified to incorporate a scoring function over *structures* such as a sequence of tags for a sentence: this is known as **structured prediction**. Let $\mathbf{x}$ denote the observed sequence of inputs (tokens) and $\mathbf{y}$ the sequence of predicted tags. The goodness of the tagging for the observed sequence is modeled as a linear function (again, with a real vector–valued

[5] To elaborate: Typically, adjacent predictions within a sequence are tied together to be interdependent—that is, the scoring function has (hard or soft) beliefs about the compatibility of adjacent predictions. Fortunately, with a Markov assumption that *non-adjacent* predictions are not directly interdependent, dynamic programming (e.g., the Viterbi algorithm) can recover the globally optimal sequence of predictions in polynomial time.

feature function **g** and parametrized by a real weight vector $\boldsymbol{\theta}$):

$$score(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \mathbf{g}(\mathbf{x}, \mathbf{y}) \tag{2.3}$$

The **decoding** (structured classification) problem given the weights $\boldsymbol{\theta}$ and input **x** is to construct the tag sequence **y** which maximizes this score. To facilitate efficient exact dynamic programming inference with the Viterbi algorithm we make a Markov assumption, stipulating that the scoring function factorizes into local functions over label bigrams:[6]

$$\mathbf{g}(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{|\mathbf{x}|+1} \mathbf{f}(\mathbf{x}, y_j, y_{j-1}, j) \tag{2.4}$$

Various supervised structured learning algorithms are available for linear models (Smith, 2011). The input to such an algorithm is a training corpus of labeled sequences, $\mathcal{D} = \left\langle \langle \mathbf{x}^{(1)}, \mathbf{y}^{(1)} \rangle, \ldots, \langle \mathbf{x}^{(N)}, \mathbf{y}^{(N)} \rangle \right\rangle$; the output is the feature weight vector $\boldsymbol{\theta}$.

One popular structured learning algorithm is the **structured perceptron** (Collins, 2002).[7] Its learning procedure, algorithm 1, generalizes the classic perceptron algorithm (Freund and Schapire, 1999) to incorporate a structured decoding step (for sequences, the Viterbi algorithm) in the inner loop. Thus, training requires only max inference, which is fast with a first-order Markov assumption. In training, features are adjusted where a tagging error is made. The result of learning is a weight vector that parametrizes a feature-rich scoring function over candidate labelings of a sequence.

[6]Note that in contrast to the independence assumptions of a generative hidden Markov model, local feature functions are allowed to see the entire observed sequence **x**.

[7]Conditional random fields (Lafferty et al., 2001) are another popular technique for discriminative sequence modeling with a convex loss function. We prefer the structured perceptron for its speed: learning and inference depend mainly on the runtime of the Viterbi algorithm, whose asymptotic complexity is linear in the length of the input and (with a first-order Markov assumption) quadratic in the number of tags.

Input: data $\left\langle \langle \mathbf{x}^{(n)}, \mathbf{y}^{(n)} \rangle \right\rangle_{n=1}^{N}$; number of iterations M
$\boldsymbol{\theta} \leftarrow \mathbf{0}$
$\bar{\boldsymbol{\theta}} \leftarrow \mathbf{0}$
$t \leftarrow 1$
for $m = 1$ **to** M **do**
  for $n = 1$ **to** N **do**
    $\langle \mathbf{x}, \mathbf{y} \rangle \leftarrow \langle \mathbf{x}^{(n)}, \mathbf{y}^{(n)} \rangle$
    $\hat{\mathbf{y}} \leftarrow \arg\max_{\mathbf{y}'} \left(\boldsymbol{\theta}^\top \mathbf{g}(\mathbf{x}, \mathbf{y}') + cost(\mathbf{y}, \mathbf{y}', \mathbf{x}) \right)$
    if $\hat{\mathbf{y}} \neq \mathbf{y}$ **then**
      $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \mathbf{g}(\mathbf{x}, \mathbf{y}) - \mathbf{g}(\mathbf{x}, \hat{\mathbf{y}})$
      $\bar{\boldsymbol{\theta}} \leftarrow \bar{\boldsymbol{\theta}} + t\mathbf{g}(\mathbf{x}, \mathbf{y}) - t\mathbf{g}(\mathbf{x}, \hat{\mathbf{y}})$
    end
    $t \leftarrow t + 1$
  end
end
Output: $\boldsymbol{\theta} - (\bar{\boldsymbol{\theta}}/t)$

Algorithm 1: Training with the averaged perceptron. (Adapted from Daumé, 2006, p. 19.)

2.4.4 Cost Functions

Algorithm 1 is a *cost-augmented* version of the structured perceptron: it assigns different values to different kinds of errors made during training. The **cost function**, $cost(\mathbf{y}, \mathbf{y}', \mathbf{x}) \geq 0$, encourages the learner to be especially careful to avoid certain kinds of mislabelings; worse errors incur a greater cost than milder errors during training (correct predictions should have a cost of 0). What counts as a "better" or "worse" error is stipulated according to the needs of the application. For example, in NER, the cost function can be defined to encourage recall by penalizing false negatives more than false positives (Mohit et al., 2012). The cost for each possible erroneous labeling is the minimum difference in score, or **margin**, between that labeling and the true labeling for the learner's prediction to be considered acceptable. The optimization performed by the cost-augmented structured perceptron algorithm approaches the same result as that of structured SVMs (Tsochantaridis et al., 2005)—in both cases the learning objective is the structured hinge loss.

2.5 Conclusion

This chapter has introduced the field of computational semantics and provided background on concepts in lexical semantics that will be crucial to understanding the rest of the thesis. These include the representation of semantic senses and classes in lexicons, the canonical tasks of word sense disambiguation and named entity recognition tasks, and linear models and algorithms for multi-class classification and sequence tagging.

PART I

Representation & Annotation

SONJA: It warmed the cockles of my heart!
BORIS: That's just great, nothing like hot cockles!

Love and Death

Q. Please explain the expression “this does not bode well.”
A. It means something is not boding the way it should. It could be boding better.

Dave Barry, *Dave Barry Is Not Making This Up*: “Punctuation 'R Easy”

CHAPTER 3

Multiword Expressions

This chapter:

- Reviews the literature on the linguistics and annotation of multiword expressions
- Proposes a formal representation of multiword lexical units in context that allows for (a) gaps, and (b) a strength distinction
- Develops a resource-agnostic linguistic understanding of which multiword combinations cohere strongly enough to count as units
- Designs a corpus annotation procedure for MWEs, documented with exemplar-based guidelines
- Describes a comprehensively annotated corpus of multiword expressions

3.1 Introduction

Language has a knack for defying expectations when put under the microscope. For example, there is the notion—sometimes referred to as *compositionality*—that words will behave in predictable ways, with individual meanings that combine to form complex meanings according to general grammatical principles. Yet language is awash with examples to the contrary: in particular, idiomatic expressions such as *awash with* NP, *have a knack for* VP$_{\text{-}ing}$, *to the contrary*, and *defy expectations*. Thanks to processes like metaphor and grammaticalization, these are (to various degrees) semantically opaque, structurally fossilized, and/or statistically idiosyncratic. In other words, idiomatic expressions may be exceptional in form, function, or distribution. They are so diverse, so unruly, so difficult to circumscribe, that entire theories of syntax are predicated on the notion that constructions with idiosyncratic form-meaning mappings (Fillmore et al., 1988; Goldberg, 1995) or statistical properties (Goldberg, 2006) offer crucial evidence about the grammatical organization of language.

Here we focus on **multiword expressions** (MWEs): *lexicalized* combinations of two or more words that are exceptional enough to be considered as single units in the lexicon. As figure 3.1 illustrates, MWEs occupy diverse syntactic and semantic functions. Within MWEs, we distinguish (a) proper names and (b) lexical idioms. The latter have proved themselves a "pain in the neck for NLP" (Sag et al., 2002). Automatic and efficient detection of MWEs, though far from solved, would have diverse applications including machine translation (Carpuat and Diab, 2010), information retrieval (Acosta et al., 2011; Newman et al., 2012), opinion mining (Berend, 2011), and second language learning (Ellis et al., 2008); see also §8.3.

It is difficult to establish any comprehensive taxonomy of mul-

1. **MW named entities:** *Chancellor of the Exchequer Gordon Brown*
2. **MW compounds:** *red tape, motion picture, daddy longlegs, Bayes net, hot air balloon, skinny dip, trash talk*
3. **conventionally SW compounds:** *snapdragon, overlook* (v. or n.), *blackjack, shootout, sunscreen, somewhere*
4. **verb-particle:** *pick up, dry out, take over, cut short, hold hostage, take seriously*
5. **verb-preposition:** *refer to, depend on, look for, prevent from*
6. **verb-noun(-preposition):** *pay attention (to), go bananas, lose it, break a leg, make the most of*
7. **support verb:** *make decisions, take breaks, take pictures, have fun, perform surgery*
8. **other phrasal verb:** *put up with, miss out (on), get rid of, look forward to, run amok, cry foul, add insult to injury*
9. **predicative or modifier PP:** *above board, beyond the pale, under the weather, at all, from time to time*
10. **coordinated phrase:** *cut and dried/dry, more or less, up and leave*
11. **conjunction/connective:** *as well as, let alone, in spite of, on the face of it/on its face*
12. **semi-fixed VP:** *smack* <one>*'s lips, pick up where* <one> *left off, go over* <thing> *with a fine-tooth(ed) comb, take* <one>*'s time, draw* <oneself> *up to* <one>*'s full height*
13. **fixed phrase:** *easy as pie, scared to death, go to hell in a handbasket, bring home the bacon, leave of absence*
14. **phatic:** *You're welcome. Me neither!*
15. **proverb:** *Beggars can't be choosers. The early bird gets the worm. To each his own. One man's* <thing$_1$> *is another man's* <thing$_2$>.

Figure 3.1: Some of the classes of idioms in English. The examples included here contain multiple lexicalized words—with the exception of those in (3), if the conventional single-word (SW) spelling is used.

tiword idioms, let alone develop linguistic criteria and corpus resources that cut across these types. Consequently, the voluminous literature on MWEs in computational linguistics—see §3.4.1, Baldwin and Kim (2010), and Ramisch (2012) for surveys—has been fragmented, looking (for example) at subclasses of phrasal verbs or nominal compounds in isolation. To the extent that MWEs have been annotated in existing corpora, it has usually been as a secondary aspect of some other scheme. Traditionally, such resources have prioritized certain kinds of MWEs to the exclusion of others, so they are not appropriate for evaluating general-purpose identification systems.

This chapter introduces a shallow form of analysis for MWEs that is neutral to expression type, and that facilitates free text annotation without requiring a prespecified MWE lexicon. The scheme applies to gappy (discontinuous) as well as contiguous expressions, and allows for a qualitative distinction of association strengths. We apply this scheme to fully annotate a 56,000-word corpus of English web reviews (Bies et al., 2012), a conversational genre in which colloquial idioms are highly salient.

We start off with background on the linguistics of MWEs (§3.2) and on available corpus resources (§3.3). Then we motivate the gist of our approach (§3.4), formalize the representational space of structures that can comprise an analysis (§3.5), and give an overview of the annotation guidelines (§3.6). The annotation process is discussed in §3.7, and the resulting dataset in §3.8. The chapter incorporates material from Schneider et al. (2014b), which described the MWE-annotated corpus, and Schneider et al. (2014a), which includes a formal description of the representation.

3.2 Linguistic Characterization

Much ink has been spilt over the definition of multiword expressions, idioms, collocations, and the like.[1] The general consensus is that many combinations of two or more wordforms are "word-like" in function. Following Baldwin and Kim (2010), we broadly construe the term **idiomatic** to apply to any expression with an exceptional form, function, or distribution; we will say such an expression has **unit** status.[2] Idiomaticity can be viewed relative to a constellation of criteria, including:

syntactic criteria: For example, if the combination has a syntactically anomalous form or is **fossilized** (resistant to morphological or syntactic transformation), then it is likely to be considered a unit (Huddleston, 2002; Baldwin and Kim, 2010). A construction exemplifying the former is *the X-er, the Y-er* (Fillmore et al., 1988); an example of the latter is the idiom *kick the bucket*, which only behaves like an ordinary verb phrase with respect to the verb's inflection: **the bucket was kicked/??kick swiftly the bucket/??the kicking of the bucket.*

[1] Gries (2008) discusses the closely related concepts of **phraseologism** in phraseology, **word cluster** and **n-gram** in corpus linguistics, **pattern** in Pattern Grammar, **symbolic unit** in Cognitive Grammar, and **construction** in Construction Grammar. In the language acquisition literature various terms for multiword expressions include **formula(ic sequence)**, **lexical phrase**, **routine**, **pattern**, and **prefabricated chunk** (Ellis, 2008). See also Moon (1998); Wray (2000).

[2] Moon's (1998, p. 6) criteria of "**institutionalization**, **lexicogrammatical fixedness**, and **non-compositionality**" correspond roughly to our criteria of distribution, form, and function, respectively. "**Institutionalization** is the process by which a string or formulation becomes recognized and accepted as a lexical item of the language" (Moon, 1998, p. 7). Moon requires all three criteria to be present in a multiword sequence to consider it a unit; we treat sequences meeting *any* of these criteria as MWEs, but further distinguish those that are exceptional only in distribution from those that are idiosyncratic in form and/or function.

semantic criteria: These often fall under the umbrella of **compositionality** vs. **lexicality**, which can refer to the notion that an expression's meaning may differ from the natural combination of the meanings of its parts.[3] This may be interpreted as a categorical or gradient phenomenon. More specifically, the meaning of the whole expression vis-a-vis its parts is said to be **transparent** (or **analyzeable**) vs. **opaque** when considered from the perspective of a hypothetical listener who is unfamiliar with it, and **predictable** vs. **unpredictable** from the perspective of a hypothetical speaker wishing to express a certain meaning. The expressions *kick the bucket* and *make sense* are neither predictable nor transparent, whereas *spill the beans* and *let slip* are unpredictable but likely to be fairly transparent in context. We will count all unpredictable or opaque expressions as units. The term **idiom** is used especially for an expression exhibiting a high degree of **figurativity** or **proverbiality** (Nunberg et al., 1994).

statistical criteria: An expression may be considered a unit because it enjoys unusually high token frequency, especially in comparison with the frequencies of its parts. Various **association measures** aim to quantify this in corpora; the most famous is the information-theoretic measure **mutual information (MI)** (Pecina, 2010). The term **collocation** generally applies to combinations that are statistically idiomatic, and an **institutionalized phrase** is idiomatic on purely statistical grounds (Baldwin and Kim, 2010).

[3]Whether an expression is "compositional" or "noncompositional" may be considered either informally, or more rigorously in the context of a formalism for compositional semantics.

psycholinguistic criteria: Some studies have found psycholinguistic correlates of other measures of idiomaticity (Ellis et al., 2008). Idiomatic expressions are expected to be memorized and retrieved wholesale in production, rather than composed on the fly (Ellis, 2008).

Some examples from Baldwin and Kim (2010) are as follows:

		Semantically idiomatic
	salt and pepper (cf. *?pepper and salt*); *many thanks*; *finish up*[4]	*traffic light*; *social butterfly*; *kick the bucket*; *look up* (= 'search for')
Syntactically idiomatic	*to and fro*	*by and large*

Unlike *eat chocolate* and *swallow down*, which are not regarded as idiomatic, all of the above expressions exhibit *statistical* idiomaticity (Baldwin and Kim, 2010). For instance, *traffic light* is more frequent than plausible alternatives like *traffic lamp*/*road light*/*intersection light* (none of which are conventional terms) or *streetlight*/*street lamp* (which have a different meaning). While *traffic light*, being an instance of the highly productive noun-noun compound construction, is not *syntactically* idiomatic, it is *semantically* idiomatic because that construction underspecifies the meaning, and *traffic light* has a conventionalized "ordinary" meaning of something like 'electronic light signal installed on a road to direct vehicular traffic'. It could conceivably convey novel meanings in specific contexts—e.g., 'glow emanating from car taillights' or 'illuminated wand used by a traffic officer for signaling'—but such usages have not been conventionalized.

[4]The completive meaning of 'up' is redundant with 'finish' (Gonnerman and Blais, 2012).

'create, constitute' (4): *make you drinks, make an army of [corpses], the kind of thing [potion] you ought to be able to make, tricky to make [potion]*
'cause (event, result, or state)' (9): *make your ears fall off, make a nice loud noise, make your brain go fuzzy, make a sound, make himself seem more important than he is, make Tom Riddle forget, make anyone sick, make you more confident, make trouble*
'be good or bad in a role' (2): *make a good witch, make a good Auror*
verb-particle constructions (2): *from what Harry could make out* (*make out* = 'reckon'), *make up to well-connected people* (*make up to* = 'cozy/kiss/suck up to'; this idiom is not present in WordNet)
light verb with eventive noun (13): *make any attempt, make the Unbreakable Vow* (×2), *make a suggestion, make the introduction, odd comment to make, make a joke, make a quick escape, make further investigations, make an entrance, make a decent attempt, make mistakes* (×2)
miscellaneous multiword expressions (9): *make different arrangements, make sure* (×5), *make do, make sense, make any sign of recognition*

Figure 3.2: Occurrences of the bare verb *make* in a small text sample.

3.2.1 Polysemy

Figure 3.2 lists the occurrences of the highly polysemous verb *make* in the first 10 chapters (about 160 pages) of *Harry Potter and the Half-Blood Prince* (Rowling, 2005).[5] Of the 39 occurrences in this sample, no more than 15 ought to be considered non-idiomatic.

Even knowing the extent of the MWE is often not sufficient to determine its meaning. The verb lemma *make up* has no fewer than 9 sense entries in WordNet, as shown in figure 3.3. Some of these senses are radically different: making up a story, a bed, a missed

[5]These were found by simple string matching; morphological variants were not considered.

1. {STATIVEˇ} form or compose
2. {CREATIONˇ} devise or compose
3. {POSSESSIONˇ} do or give something to somebody in return
4. {SOCIALˇ} make up work that was missed due to absence at a later point
5. {CREATIONˇ} concoct something artificial or untrue
6. {CHANGEˇ} put in order or neaten
7. {STATIVEˇ} adjust for
8. {COMMUNICATIONˇ} come to terms
9. {BODYˇ} apply make-up or cosmetics to one's face to appear prettier

Figure 3.3: WordNet senses of *make up*. The supersense label is shown alongside the definition.

exam, one's face, and (with) a friend have very little in common![6] Reassuringly, the supersenses attest to major differences, which suggests that the MWE grouping and supersense tags offer complementary information (in ch. 7 we exploit this complementarity in a unified model).

3.2.2 Frequency

Sources in the literature agree that multiword expressions are numerous and frequent in English and other languages (Baldwin and Kim, 2010; Ellis et al., 2008; Ramisch, 2012). Table 3.1 (p. 40) quantifies the frequency of MWEs in three corpora from different domains. (Appendix A takes an in-depth look at a subset of the **SEMCOR** data.) These corpora use different criteria for marking MWEs, so the relative frequencies are not directly comparable, but they show that frequency of MWE tokens is nothing to sneeze at. For example, in

[6]Arguably, senses 7 and 8 ought to be listed as prepositional verbs: *make up for* and *make up with*, respectively.

our corpus (§3.8), the proportion of word tokens belonging to an MWE is nearly as large as the proportion of words that are common nouns.

3.2.3 Syntactic Properties

Multiword expressions are diverse not only in function, but also in form. As noted above, some idioms are anomalous or highly inflexible in their syntax. But more commonly they exploit productive syntactic patterns. In the computational literature, studies generally focus on individual classes of English MWEs, notably:

- complex nominals, especially noun-noun and adjective-noun compounds (Lapata and Lascarides, 2003; Michelbacher et al., 2011; Hermann et al., 2012a,b)
- determinerless prepositional phrases (Baldwin et al., 2006)
- verbal expressions, including several non-disjoint subclasses: **phrasal verbs** (Wulff, 2008; Nagy T. and Vincze, 2011; Tu and Roth, 2012), generally including **verb-particle constructions** (where the particle is intransitive, like *make up*) (Villavicencio, 2003; McCarthy et al., 2003; Bannard et al., 2003; Cook and Stevenson, 2006; Kim and Baldwin, 2010) and **prepositional verbs** (with a transitive preposition, like *wait for*); **light verb constructions/support verb constructions** like *make... decision* (Calzolari et al., 2002; Fazly et al., 2007; Tu and Roth, 2011; Bonial et al., 2014b); and **verb-noun constructions** like *pay attention* (Ramisch et al., 2008; Diab and Bhutada, 2009; Diab and Krishna, 2009; Boukobza and Rappoport, 2009; Wulff, 2010)

By convention, the constructions referred to as multiword expressions have two or more lexically fixed morphemes. Some are

completely frozen in form, or allow for morphological inflection only. Other MWEs permit or require other material in addition to the lexically specified portions of the expression. Of particular interest in the present work are **gappy multiword expressions**. In our terminology, gappiness is a property of the surface mention of the expression: a mention is gappy if its lexicalized words are interrupted by one or more additional words. This happens in the following scenarios:

- When the expression takes a lexically unspecified argument, such as an object or possessive determiner, occurring between lexicalized parts (the **argument gap** column of figure 3.4);[7]
- When an internal modifier such as an adjective, adverb, or determiner is present (the **modifier gap** column of figure 3.4);
- When the expression is transformed via some syntactic process such that other words intervene. This is relatively rare; examples we found in the SemCor involved fronting of prepositional verb complements (e.g. *those if any <u>on</u> ⟨ whom we can ⟩ <u>rely</u>*) and coordination (*<u>grade</u> ⟨ and high ⟩ <u>schools</u>*).[8]

One final point worth making is that multiword expressions create syntactic ambiguity. For example, somone might *make [up to*

[7]This is not to suggest that the syntactic arguments MWEs always fall between lexicalized words: with prepositional verbs and verb-particle constructions, for instance, the open argument typically follows the verb and preposition (*<u>make up</u> a story*, *<u>rely on</u> someone*)—but we will not refer to these as *gaps* so long as the lexically fixed material is contiguous.

[8]In the coordination example the word *schools* is really shared by two MWEs. Another case of this might be a phrase like *fall fast asleep*, where *fall asleep* and *fast asleep* are arguably MWEs. But this sharing is extremely rare, so in the interest of simplicity our representation will prevent any word token from belonging to more than one MWE mention.

construction	argument gap	modifier gap
Complex nominal		*a great head of ⟨ brown ⟩ hair*
Verb-particle	*leave ⟨ his mother ⟩ behind*	
Prepositional verb	*kept ⟨ me ⟩ from painting*	*look ⟨ just ⟩ like a set,* *coming ⟨ with a friend ⟩ upon*
Verb-noun	*caught ⟨ her ⟩ breath,* *made up ⟨ her ⟩ mind*	*runs ⟨ too great ⟩ a risk,* *paid ⟨ no ⟩ attention*
Verb-PP	*put ⟨ many persons ⟩ to death*	*falls ⟨ hopelessly ⟩ in love*
Verb-adverb		*stood ⟨ very ⟩ still*

Figure 3.4: Examples of gappy MWEs in the SemCor corpus. See §A.1 for further analysis.

a million dollars] or *make up [to a friend]*. This is further complicated by expressions that license gaps. In the context of describing one's ascent of Kilimanjaro, *make the climb up* probably cannot be paraphrased as *make up the climb*. Heuristic matching techniques based on n-grams are likely to go awry due to such ambiguity—for some kinds of MWEs, more sophisticated detection strategies are called for (see ch. 6).

3.2.4 Multiword Expressions in Other Languages

Though our presentation of multiword expressions has focused on English, MWEs are hardly an English-specific phenomenon. Studies in other languages have included Basque compound prepositions (Díaz de Ilarraza et al., 2008), German determinerless PPs (Dömges et al., 2007; Kiss et al., 2010), German complex prepositions (Trawinski, 2003), Hebrew noun compounds (Al-Haj and Wintner, 2010), Japanese and English noun-noun compounds (Tanaka and Baldwin, 2003), Japanese compound verbs (Uchiyama and Ishizaki, 2003), Korean light verb constructions (Hong et al., 2006), Persian compound verbs (Rasooli et al., 2011), and Persian light verb constructions

(Salehi et al., 2012). The new multiword datasets we propose below will be in English, but we intend to evaluate our system on the multiword expressions in the French Treebank (Abeillé et al., 2003), as discussed below.

3.3 Existing Resources

Annotated corpora do not pay much attention to multiword expressions. On the one hand, MWEs are typically not factored into the syntactic and morphological representations found in treebanks.[9] On the other, the MWE literature has been driven by lexicography: typically, the goal is to acquire an MWE lexicon with little or no supervision, or to apply such a lexicon to corpus data. Studies of MWEs in context have focused on various subclasses of constructions in isolation, necessitating special-purpose datasets and evaluation schemes.

Without getting into the details of automatic multiword analysis tasks here just yet (they will appear in ch. 6), we take the position that a comprehensive treatment requires corpora annotated for a broad variety of multiword expressions. A canonical corpus resource would offer to the the multiword expressions community a benchmark dataset comparable to datasets used for problems such as NER and parsing.

To our knowledge, only a few existing corpora approach this goal of marking heterogeneous MWEs in context:

[9]Some datasets mark shallow phrase chunks (Tjong Kim Sang and Buchholz, 2000), but these are not the same as multiword expressions: syntactically, *green dye* and *green thumb* are both noun phrases, yet only the second is idiomatic.

[10]Estimates are for version 2.0, which is annotated for MWEs and noun and verb supersense. Version 3.0, which will add preposition supersenses, is under development. Statistics for CMWE version 1.0 (MWEs only) appear in §3.8.

[11]**SemCor** counts include 166 documents/17k sentences/386k words that have

	SemCor (Miller et al., 1993)	Wiki50 (Vincze et al., 2011)	STREUSLE[10] (*this work*)
text source	Brown Corpus	Wikipedia	EWTB
genre	published texts	crowdsourced articles	user reviews
docs · sents · words	352 · 37k · 820k[11]	50 · 4.4k · 114k[12]	723 · 3.8k · 56k
words/sents	22	26	15
syntactic parses	139 docs in PTB	—	EWTB
NE instances	9700	9000	—
MW NEs	3900	3600	≈500[13]
other MWEs		3900	3000
contiguous	30,000	3600	2500
gappy	*not explicit*[14]	220 LVCs, 40 VPCs	500
total LEs	780k	100k	51k
NE classes	PER, ORG, LOC, MISC[15]		*not distinguished from supersenses*
other classes	NOTAG, COMPLEXPREP, FOREIGNWORD, IDIOM, METAPHOR, NONCEWORD[16]	COMPOUND_ADJ, COMPOUND_NOUN, IDIOM, LVC, VPC, OTHER[17]	strong, weak
semantic senses	32k WordNet synsets	—	41 supersenses[18]
labeled instances	235k synsets+NEs	9k NEs	17k N/V mentions

Table 3.1: Comparison of two existing English lexical semantic corpora with the one created in this work. Counts of documents, sentences, and space-separated tokens are rounded.

been sense-tagged only for verbs.

[12]**Wiki50** counts are from the "distilled" version of the data.

[13]Our **STREUSLE** dataset does not explicitly mark NEs, but we estimate the number of multiword NEs by counting the number of strong MWEs containing a proper noun.

[14]**SemCor**'s file format does not directly mark gappy MWEs, though sometimes only the first part of one is tagged with the synset (e.g., of a particle verb). See

3.3.1 SemCor

As discussed in §2.3.2, **SemCor**[19] includes named entities and many other multiword expressions, most of which are tagged with WordNet senses. Exactly how the lexicographic decisions were made is not documented, but WordNet seems to prioritize complex nominals and verb-particle constructions over other kinds of multiword constructions.

Statistics and further details on lexical expressions in SemCor appear in table 3.1 and appendix A.

3.3.2 Wiki50

The **Wiki50** corpus (Vincze et al., 2011)[20] consists of 50 English Wikipedia articles fully annotated for named entities as well as several classes of other MWEs—principally compound nominals,

discussion in §A.1.

[15]In SemCor the respective labels are PERSON, GROUP, LOCATION, and OTHER.

[16]These special designations apply to lexical expressions with no appropriate WordNet synset. Token frequencies in SemCor: 39,401 NOTAG (1,827 of which are contiguous MWEs), 320 COMPLEXPREP (312 contiguous MWEs), 77 FOREIGNWORD (18 contiguous MWEs), 270 IDIOM (215 contiguous MWEs), 155 METAPHOR (29 contiguous MWEs), 21 NONCEWORD (4 contiguous MWEs). Two caveats: (a) NOTAG is not a well-defined class. It contains some miscellaneous lexical expressions (function words and MWEs such as *a lot* and *such as*), as well as many apparently spurious multiword chunks, such as *many of* and *must not*. It also contains many named entities and other expressions that presumably belong to another category but were never reviewed. (b) A handful of these labels apply to an obviously gappy MWE—e.g., *due* ⟨ *in large part* ⟩ *to* as COMPLEXPREP—but these cannot be counted automatically because the part after the gap is not annotated.

[17]See table 3.2.

[18]26 noun supersenses and 15 verb supersenses, including a special verb category for auxiliaries. The WEATHERˇ supersense for verbs is not attested in the corpus.

[19]http://lit.csci.unt.edu/~rada/downloads/semcor/semcor3.0.tar.gz

[20]http://www.inf.u-szeged.hu/rgai/project/nlp/research/mwe/attachments.zip

		NEs				
		SW	MW			
P. G. Wodehouse	PER	2743	1344	78	CMPD_ADJ	*brand new*
				2926	CMPD_NOUN	*hunger strike*
Monty Python	ORG	647	849	19	IDIOM	*is on cloud nine*
Kii Province	LOC	978	579	368	LVC	*raise ⟨ the ⟩ issue*
				446	VPC	*let ⟨ Max ⟩ down*
Kentucky Derby	MISC	975	844	21	OTHER	*alter ego*
total		5343	3616	3858		
				MWEs		

Table 3.2: Categories and tokens in the **WIKI50** corpus for named entities (single-word and multiword) and other MWEs. (Above, COMPOUND is abbreviated as CMPD.)

light verb constructions (LVCs),[21] and verb-particle constructions (VPCs). The LVC (VPC) annotations specifically designate the word tokens that belong to the verb and the words that belong to the noun (particle)—there may be a gap between the verb part and the noun (particle) part, but nested annotations within the gap are not allowed. There are also two rare categories: phrasal idioms such as *come out of the closet* (IDIOM), and OTHER, which consists of compound verbs and foreign phrases.[22] Examples and counts of these expressions appear in table 3.2, and a comparison to other corpora in table 3.1. In §6.5.8 we use this corpus for out-of-domain

[21] The definition of LVC is controversial. Vincze et al. (2011) broadly define LVCs as constructions "where the noun is usually taken in one of its literal senses but the verb usually loses its original sense to some extent e.g. *to give a lecture, to come into bloom, the problem lies (in)*." Some examples in the data, such as *change... mind* and *draw... ire*, might better be classified as support verb constructions (Calzolari et al., 2002) or verb-noun idiomatic combinations; see discussion in Baldwin and Kim (2010).

[22] Namely: *down played, drink and drive, double teamed, free fall, test fire, voice acted, alter ego, de facto, fait accompli, modus operandi, non sequitur, per capita*, and *status quo*.

evaluation of our MWE identification system.

3.3.3 Other English Corpora

The **Prague Dependency Treebank (PDT)** (Hajič, 1998) and the **Prague Czech-English Dependency Treebank (PCEDT)** (Čmejrek et al., 2005)[23] contain rich annotations at multiple levels of syntactic, lexical, and morphological structure. Bejček and Straňák (2010) describe the technical processes involved in multiword expression annotation in the (Czech) PDT; notably, their corpus annotation benefited from and informed a multiword lexicon, SemLex, whereas our annotation procedure is not tied to any lexicon. The PCEDT contains parallel annotations for English (source) and Czech (translated) versions of the WSJ corpus (Marcus et al., 1993). Morphosyntactic structures for several classes of multiword expressions are detailed in the manual for the English tectogrammatical annotation layer (Cinková et al., 2006). These annotations are complex, but it is possible to automatically extract some shallow multiword groupings for named entities, light verb constructions (marked `CPHR`), phrasal idioms (`DPHR`), and certain other MWEs (for an explanation of the `CPHR` and `DPHR` functors, see Urešová et al., 2013).

Several other corpus resources describe English **LVCs** (Tan et al., 2006; Hwang et al., 2010a; Tu and Roth, 2011; Vincze, 2012; Rácz et al., 2014; Bonial et al., 2014a) or **named entities** (§2.4.1).

3.3.4 The French Treebank

Though this thesis focuses on English, the **French Treebank** is notable for having facilitated several evaluations of MWE identification systems (Constant and Sigogne, 2011; Constant et al., 2012; Candito

[23] `http://ufal.mff.cuni.cz/pcedt2.0/index.html`

and Constant, 2014; Green et al., 2011, 2012; see §6.6). It designates syntactic constituents that correspond to a subclass of MWEs, which it terms *compounds*:

> Compounds also have to be annotated since they may comprise words which do not exist otherwise (e.g. *insu* in the compound preposition *à l'insu de* = unbeknownst to) or exhibit sequences of tags otherwise non-grammatical (e.g. *à la va vite* = Prep + Det + finite verb + adverb, meaning 'in a hurry'), or sequences with different grammatical properties than expected from those of the parts: peut-être is a compound adverb made of two verb forms, a *peau rouge* (American Indian) can be masculine (although *peau* (skin) is feminine in French) and a *cordon bleu* (master chef) can be feminine (although *cordon* (ribbon) is masculine in French). (Abeillé et al., 2003, p. 167)

Contiguity up to simple internal modification is given as a criterion (Abeillé and Clément, 2003, p. 44):

> Les composants sont contigus. Seule quelques petites insertions sont possibles (en général un petit adverbe ou adjectif).
>
> The compounds are contiguous. Only some small insertions are possible (in general a short adverb or adjective).
>
> *à force de* [by repeated action of, due to]
> *un maillot <doré> deux-pièces* [a <gold> bikini/2-piece swimsuit]
> *?? un maillot <de ma soeur> deux pièces* [a 2-piece <my sister's> swimsuit]

Compounds are semi-automatically detected based on preexisting lexical resources (Abeillé et al., 2003, pp. 170, 172). This category appears to be rather narrow, excluding (for example) support verbs except in frozen idioms (Abeillé et al., 2004, p. 23):

> Dans certains cas on peut trouver un verbe support suivi d'un nom et d'un complément prépositionnel : *avoir peur de*, *avoir envie de*, etc. On ne compose pas le verbe parce que le nom peut former un syntagme plus complexe (*avoir une peur bleue de*, *avoir la plus grande envie de*), et parce qu'on peut le déplacer (*la peur que j'ai eue*), ce qui montre que ce type de construction n'est pas figé.
>
> In some cases you can find a support verb followed by a noun and a prepositional complement: 'be afraid of' [lit. 'have fear of'], 'want to' [lit. 'have desire of'], etc. We do not compose the verb because the noun can form a more complex phrase ('be deathly afraid of' lit. 'have a blue fear of', 'have the greatest desire to'), and because it can be moved ('the fear that I had'), which shows that this type of construction is not fixed.

The morphosyntactic guidelines further elaborate on the limitations of the compound verb category (Abeillé and Clément, 2003, p. 52):

> **Les verbes composés:** On a choisi d'en retenir très peu dans le corpus, car la plupart sont dis-
>
> **Compound verbs:** We chose to retain very few in the corpus, as most are discontinuous, and

continus, et suivent une syntaxe régulière.

On a retenu les expressions verbales qui mettent en jeu un composant n'existant pas par ailleurs (*faire fi de*) ou celles qui sont très figées (V N sans déterminant possible: *faire partie de*) et non compositionnelles. [...]

On ne fige pas la Prep à la fin de l'expression [...]

follow a regular syntax.

We retained verbal expressions that involve a component that does not otherwise exist [i.e., a fossil word] or those which are firmly frozen [...] and non-compositional. [...]

We do not fix the Prep at the end of the expression, because [it syntactically heads a PP complement].

The last sentence apparently indicates that prepositional verbs are not marked as compounds.

3.4 Taking a Different Tack

The remainder of this chapter is devoted to the design of an annotation scheme and corpus that offers a more comprehensive treatment of multiword expressions in context. Applied to a 56,000-word corpus of English web text with the aim of full corpus coverage, our novel scheme emphasizes:

- *heterogeneity*—the annotated MWEs are not restricted by syntactic construction;
- *shallow but gappy grouping*—MWEs are simple groupings of tokens, which need not be contiguous in the sentence; and

- *expression strength*—the most idiomatic MWEs are distinguished from (and can belong to) weaker collocations.

We examine these characteristics in turn below. Details of the formal representation appear in §3.5, the annotation guidelines in §3.6, and the annotation process in §3.7. §3.8 gives an overview of the resulting annotated corpus. The annotations are available for download at `http://www.ark.cs.cmu.edu/LexSem`.

3.4.1 Heterogeneity

By "multiword expression," we mean a group of tokens in a sentence that cohere more strongly than ordinary syntactic combinations: that is, they are idiosyncratic in *form*, *function*, or *frequency*. As figure 3.1 shows, the intuitive category of MWEs or idioms cannot be limited to any syntactic construction or semantic domain. The sheer number of multiword types and the rate at which new MWEs enter the language make development of a truly comprehensive lexicon prohibitive. Therefore, we set out to build a corpus of MWEs without restricting ourselves to certain candidates based on any list or syntactic category. Rather, annotators are simply shown one sentence at a time and asked to mark all combinations that they believe are multiword expressions. Examples from our corpus appear in figures 3.5 (below) and 3.6 (p. 56).

3.4.2 Shallow token groupings

Concretely, we represent each MWE as a grouping of tokens within a sentence. The tokens need not be contiguous: **gappy** (discontinuous) uses of an expression may arise due to internal arguments, internal modifiers, and constructions such as passives (see §3.2.3). For example, sentence (2) in figure 3.5 contains a gappy instance of

(2) My wife had taken_1 ⟨ her '07_2 Ford_2 Fusion_2 ⟩ in_1 for a routine oil_3 change_3 .

(3) he was willing to budge_1 ⟨ a_2 little_2 ⟩ on_1 the price which $\underline{\text{means}^4\ \text{a}_3^4\ \text{lot}_3^4\ \text{to}^4}$ $\underline{\text{me}^4}$.

Figure 3.5: Two sentences from the corpus. Subscripts and text coloring indicate strong multiword groupings; superscripts and underlining indicate weak groupings. Angle brackets indicate gaps.

the verb-particle construction *take in*. It also contains two contiguous MWEs, the named entity *'07 Ford Fusion* and the noun-noun compound *oil change*. Syntactic annotations are not used or given as part of the MWE annotation, though MWEs can be syntactically categorized with part-of-speech tags (as in appendix C and figure 3.7) or syntactic parses.

3.4.3 Strength

Qualitatively, the strength of association between words can vary on a continuum of lexicality, ranging from fully transparent collocations to completely opaque idioms (Bannard et al., 2003; Baldwin et al., 2003; McCarthy et al., 2003; Baldwin, 2006, *inter alia*). In the interest of simplicity, we operationalize this distinction with two kinds of multiword groupings: **strong** and **weak**. For example, the expression *close call* describes a situation in which something bad nearly happened but was averted (*He was late and nearly missed the performance—it was a* ***close call***). This semantics is not readily predictable from the expression: the motivation for *call* in this expression is opaque; and moreover, **near call* and **far call* are not acceptable variants,[24] nor can the danger be described as **closely calling*

[24] But note that *close shave* and *near miss* are other idioms using the same "proximity to danger" metaphor.

or **calling close*. We therefore would treat *close call* as a strong MWE. On the other hand, the expression *narrow escape* is somewhat more transparent and flexible—one can *narrowly escape/avoid* an undesirable eventuality, and the alternative formulation *close escape* is acceptable, though less conventional—so it would therefore qualify as a weak MWE. Along the same lines, *abundantly clear* and *patently obvious* (?*patently clear*, ?*abundantly obvious*) would be considered mostly compositional but especially frequent collocations/phrases, and thus marked as weak MWEs.

While there are no perfect criteria for judging MWE-hood, several heuristics tend to be useful when a phrase's status is in doubt. The strongest cues are semantic opacity and morphosyntactic idiosyncrasy: if a word has a function unique to a particular expression, or an expression bucks the usual grammatical conventions of the language, the expression is almost certainly an MWE. It often helps to test how fixed/fossilized the expression is, by substituting words with synonyms/antonyms, adding or removing modifiers, or rearranging the syntax. Another strategy is to search large corpora for the expression to see if it is much more frequent than alternatives. In practice, it is not uncommon for annotators to disagree even after considering these factors, and to compromise by marking something as a weak MWE.

For purposes of annotation, the only constraints on MWE groupings are: (a) a group must consist of two or more tokens; (b) all tokens in a group must belong to the same sentence; (c) a given token may belong to at most one strong group and at most one weak group; and (d) strong groups must cohere when used inside weak groups—i.e., if a token belongs to both a strong group and a weak group, all other tokens in the strong group must belong to the same weak group.

3.5 Formal Representation

With these principles in mind, it is time to lay out formally the space of possible MWE analyses given a sentence.

We define a **lexical segmentation** of a sentence as a partitioning of its tokens into segments such that each segment represents a single unit of lexical meaning. A *multiword* lexical expression may contain **gaps**, i.e. interruptions by other segments. We impose two restrictions on gaps that appear to be well-motivated linguistically:

- **Projectivity:** Every expression filling a gap must be completely contained within that gap; gappy expressions may not interleave.
- **No nested gaps:** A gap in an expression may be filled by other single- or multiword expressions, so long as those expressions do not themselves contain gaps.

Formal grammar. Our scheme corresponds to the following extended context-free grammar (Thatcher, 1967), where S is the full sentence and terminals w are word tokens:

$$\begin{aligned} S &\rightarrow X^+ \\ X &\rightarrow \underline{w}^+ \ (Y^+ \ \underline{w}^+)^* \\ Y &\rightarrow \underline{w}^+ \end{aligned}$$

Each expression X or Y is lexicalized by the words in one or more underlined variables on the right-hand side. An X constituent may optionally contain one or more gaps filled by Y constituents, which must not contain gaps themselves.[25]

[25] MWEs with multiple gaps are rare but attested in data: e.g., ***putting*** *me* ***at*** *my* ***ease***. We encountered one violation of the gap nesting constraint in the reviews data: *I* $\textit{have}^2_1$ $\textit{nothing}^2_1$ $\textit{but}^2_1$ *fantastic* $\textit{things}^2$ $\textit{to}^2_1$ $\textit{say}^2_1$. Additionally, the interrupted phrase *great gateways* $\textit{never}^1$ $\textit{before}^1$, $\textit{so}^2_3$ $\textit{far}^2_3$ $\textit{as}^2_3$ *Hudson* $\textit{knew}^2$, $\textit{seen}^1$ *by Europeans* was annotated in another corpus.

Denoting multiword groupings with subscripts, *My wife had* $\textit{taken}_1$ $\langle$ *her* $\textit{'07}_2$ $\textit{Ford}_2$ $\textit{Fusion}_2$ $\rangle$ $\textit{in}_1$ *for a routine* $\textit{oil}_3$ $\textit{change}_3$ contains 3 multiword groups—{*taken, in*}, {*'07, Ford, Fusion*}, {*oil, change*}—and 7 single-word groups. The first MWE is gappy; a single word and a contiguous multiword group fall within the gap. This corresponds to the following derivation in terms of the formal grammar:

$$\underbrace{\underbrace{\text{My}}_{X}\ \underbrace{\text{wife}}_{X}\ \underbrace{\text{had}}_{X}\ \underbrace{\text{taken}\ \underbrace{\text{her}}_{Y}\ \underbrace{\text{'07 Ford Fusion}}_{Y}\ \text{in}}_{X}\ \underbrace{\text{for}}_{X}\ \underbrace{\text{a}}_{X}\ \underbrace{\text{routine}}_{X}\ \underbrace{\text{oil change}}_{X}\ \underbrace{\text{.}}_{X}}_{S}$$

The projectivity constraint forbids an analysis like $\textit{taken}_1$ *her* $\textit{'07}_2$ $\textit{Ford}_1$ $\textit{Fusion}_2$, while the gap nesting constraint forbids $\textit{taken}_1$ $\langle$ $\textit{her}_2$ $\langle$ *'07* $\rangle$ $\textit{Ford}_2$ $\textit{Fusion}_2$ $\rangle$ $\textit{in}_1$.

3.5.1 Two-level Scheme: Strong vs. Weak MWEs

Our annotated data distinguish two strengths of MWEs as discussed in §3.6. Augmenting the grammar of the previous section, we therefore designate nonterminals as strong ($\bar{X}$, $\bar{Y}$) or weak ($\tilde{X}$, $\tilde{Y}$):

$$\begin{aligned} S &\rightarrow \tilde{X}^+ \\ \tilde{X} &\rightarrow \underline{\bar{X}}^+ \ (\tilde{Y}^+ \ \underline{\bar{X}}^+)^* \\ \bar{X} &\rightarrow \underline{w}^+ \ (\tilde{Y}^+ \ \underline{w}^+)^* \\ \tilde{Y} &\rightarrow \underline{\bar{Y}}^+ \\ \bar{Y} &\rightarrow \underline{w}^+ \end{aligned}$$

A weak MWE may be lexicalized by single words and/or strong multiwords. Strong multiwords cannot contain weak multiwords except in gaps. Further, the contents of a gap cannot be part of any multiword that extends outside the gap.[26]

[26] This was violated 6 times in our annotated data: modifiers within gaps are sometimes collocated with the gappy expression, as in $\textit{on}^1_2$ $\textit{a}^1_2$ $\textit{tight}^1$ $\textit{budget}^1_2$ and

For example, consider the segmentation: *he was willing to* $\textit{budge}_1$ $\textit{a}_2$ $\textit{little}_2$ $\textit{on}_1$ *the price which* $\textit{means}^4$ $\textit{a}^4_3$ $\textit{lot}^4_3$ $\textit{to}^4$ $\textit{me}^4$. Subscripts denote strong MW groups and superscripts weak MW groups; unmarked tokens serve as single-word expressions. The MW groups are thus {*budge*, *on*}, {*a*, *little*}, {*a*, *lot*}, and {*means*, {*a*, *lot*}, *to*, *me*}. As should be evident from the grammar, the projectivity and gap-nesting constraints apply here just as in the 1-level scheme.

3.6 Annotation Scheme

The previous section outlined a fairly simple formal representation to describe what annotations *can* encode: A sentence's lexical segmentation is formed by grouping together space-separated tokens, subject to a few constraints, with the option to distinguish between strong and weak groupings. To keep the scheme fully general-purpose, the annotator is not tied to any particular taxonomy or syntactic structure when marking MWEs. This simplifies the number of decisions that have to be made for each sentence.

Now we turn to a much thornier issue: what our annotations *should* encode. How is the annotator to decide which tokens should belong to the same MWE instance? This is a question of linguistic conventions; the contours of our answer were arrived at over time and set down in roughly a dozen pages of annotation guidelines rife with examples.

Reproduced in appendix B, the guidelines document describes general issues and considerations (e.g., inflectional morphology; the spans of named entities; date/time/address/value expressions; overlapping expressions), then briefly discusses about 40 categories of constructions such as comparatives (*as X as Y*), age descriptions (*N years old*), complex prepositions (*out of*, *in front of*), discourse

$\textit{have}^1_2$ $\textit{little}^1$ $\textit{doubt}^1_2$.

connectives (*to start off with*), and support verb constructions (*make a decision*, *perform surgery*).

Some further instructions to annotators include:

- Groups should include only the lexically fixed parts of an expression (modulo inflectional morphology); this generally excludes determiners and pronouns: ***made*** *the* ***mistake***, ***pride*** *themselves* ***on***.[27]

- Multiword proper names count as MWEs.

- Misspelled or unconventionally spelled tokens are interpreted according to the intended word if clear.

- Overtokenized words (spelled as two tokens, but conventionally one word) are joined as multiwords. Clitics separated by the tokenization in the corpus—negative *n't*, possessive *'s*, etc.—are joined if functioning as a fixed part of a multiword (e.g., ***T 's Cafe***), but not if used productively.

- Some constructions require a possessive or reflexive argument (see semi-fixed VP examples in figure 3.1). The possessive or reflexive marking is included in the MWE only if available as a separate token; possessive and reflexive pronouns are excluded because they contain the argument and the inflection in a single token. This is a limitation of the tokenization scheme used in the corpus.[28]

[27] In some cases idiosyncratic constructions were rejected because they did not contain more than one lexicalized element: e.g., the construction *have* + <evaluative adjective> + <unit of time> (*have an excellent day*, *had a bad week*, etc.).

[28] MWE annotators were not permitted to modify the sentence and word tokenizations supplied by the treebank. Because we use treebank data, syntactic parses are available to assist in post hoc analysis. Syntactic information was not shown to annotators.

- A handful of cases of apparent MWE **overlap** emerged during the course of our annotation: e.g., for *threw a surprise birthday party*, the groups {*threw*, *party*}, {*surprise*, *party*}, and {*birthday*, *party*} all would have been reasonable; but, as they share a token in common, the compromise decision was to annotate {*birthday*, *party*} as a strong MWE and {*threw*, {*birthday*, *party*}} as a weak MWE.

While annotators' understanding of the task and conventions developed over time, we hope to have documented the conventions well enough that a new annotator could learn them reasonably well without too much difficulty.

3.7 Annotation Process

Over the course of 5 months, we fully annotated the 56,000-word REVIEWS section of the English Web Treebank (Bies et al., 2012). MWE annotation proceeded document by document, sentence by sentence. **Annotators** were the first six authors of (Schneider et al., 2014b). All are native speakers of English, and five hold undergraduate degrees in linguistics.

The annotation took three forms: (a) **individual** annotation (a single annotator working on their own); (b) **joint** annotation (collaborative work by two annotators who had already worked on the sentence independently); and (c) **consensus** annotation (by negotiation among three or more annotators, with discussion focused on refining the guidelines). In joint and consensus annotation, differences of opinion between the individual annotations were discussed and resolved (often by marking a weak MWE as a compromise). Initially, consensus annotation sessions were held semi-weekly; the rate of these sessions decreased as agreement improved. Though consensus annotations are only available for 1/5 of the sentences,

every sentence was at least reviewed independently and jointly. The annotation software recorded the full version history of each sentence; during some phases of annotation this was exposed so that analyses from different annotators could be compared.

The judgment of whether an expression should qualify as an MWE relied largely on the annotator's intuitions about its semantic coherence, idiosyncrasy, and entrenchment in the language. As noted in §3.4.3, the decision can be informed by heuristics. Judgments about the acceptability of syntactic manipulations and substitution of synonyms/antonyms, along with informal web searches, were often used to investigate the fixedness of candidate MWEs; a more systematic use of corpus statistics (along the lines of Wulff, 2008) might be adopted in the future to make the decision more rigorous.

Annotation guidelines. The annotation conventions discussed in §3.6 were developed and documented on an ongoing basis as the annotation progressed.

Annotation interface. A custom web interface, figure 3.6, was used for this annotation task. Given each pretokenized sentence, annotators added underscores (_) to join together strong multiwords and tildes (~) for weak MWEs. During joint annotation, the original annotations were displayed, and conflicts were automatically detected.

Inter-annotator agreement. Blind inter-annotator agreement figures show that, although there is some subjectivity to MWE judgments, annotators can be largely consistent. E.g., for one measurement over a sample of 200 sentences, the average inter-annotator

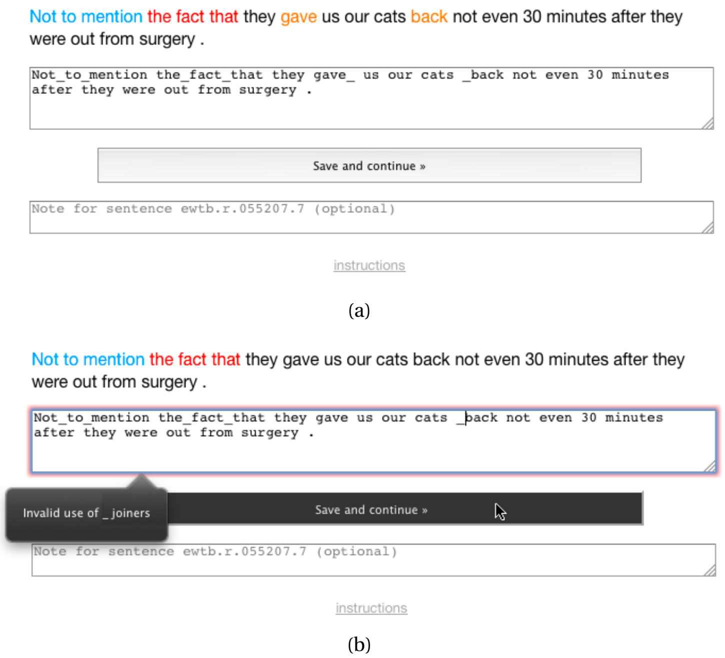


Figure 3.6: MWE annotation interface. The user joins together tokens in the textbox, and the groupings are reflected in the color-coded sentence above. Invalid markup results in an error message (b). A second textbox is for saving an optional note about the sentence. The web application also provides capabilities to see other annotations for the current sentence and to browse the list of sentences in the corpus (not shown).

	constituent tokens					*number of gaps*			*gap length*		
	2	**3**	**4**	**≥5**	**total**	**0**	**1**	**2**	**1**	**2**	**≥3**
strong	2,257	595	126	46	3,024	2,626	394	4	259	98	45
weak	269	121	44	25	459	322	135	2	93	38	8
	2,526	716	170	71	3,483	2,948	529	6	352	136	53

(a) MWE instances by number of constituent word tokens (b) MWEs by number of gaps (c) Gaps by length (in tokens)

Table 3.3: Annotated corpus statistics over 723 documents (3,812 sentences). $8,060/55,579$=15% of tokens belong to an MWE; in total, there are 3,024 strong and 459 weak MWE instances. 82 weak MWEs (18%) contain a strong MWE as a constituent (e.g., *means **a lot** to me* in figure 3.5 and *get **in touch** with* in figure 4.5).

F_1 over all 10 pairings of 5 annotators was 65%.[29] When those annotators were divided into two pairs and asked to negotiate an analysis with their partner, however, the agreement between the two *pairs* was 77%, thanks to reductions in oversights as well as the elimination of eccentric annotations.

Difficult cases. Prepositions were challenging throughout; it was particularly difficult to identify prepositional verbs (*speak with*? *listen to*? *look for*?). We believe a more systematic treatment of preposition semantics is necessary, and undertake to provide one in ch. 5. Nominal compounds (*pumpkin spice latte*?) and alleged support verbs (especially with *get*: *get busy*? *get a flat*?) were frequently controversial as well.

[29] Our measure of inter-annotator agreement is the precision/recall–based MUC criterion (Vilain et al., 1995), described in §6.2. Originally developed for coreference resolution, it gives us a way to award partial credit for partial agreement on an expression.

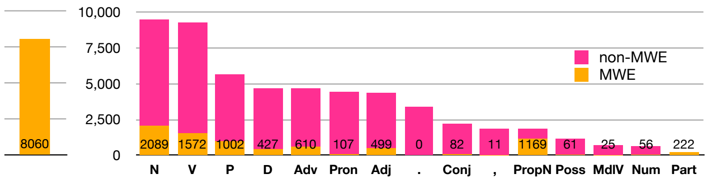


Figure 3.7: Distribution of tokens in the corpus by gold POS grouping and whether or not they belong to an MWE. Overall, 8,060 tokens are within an MWE; this not much less than the total number of common nouns (left). The rarest POS categories are not shown; of these, the only ones with large proportions of MWE tokens are hyphens (79/110) and incomplete words (28/31).

3.8 The Corpus

The MWE corpus (Schneider et al., 2014b)[30] consists of the full REVIEWS subsection of the English Web Treebank (Bies et al., 2012), comprising 55,579 words in 3,812 sentences. Each of the 723 documents is a user review of a service such as a restaurant, dentist, or auto repair shop. The reviews were collected by Google, tokenized, and annotated with phrase structure trees in the style of the Penn Treebank (Marcus et al., 1993). Most reviews are very short; half are 1–5 sentences long and only a tenth of the reviews contain 10 or more sentences. The writing style of these reviews is informal, so we would expect a lot of colloquial idioms, perhaps for dramatic effect (especially given the strong opinions expressed in many reviews).[31]

[30] Released as the Comprehensive Multiword Expressions (CMWE) Corpus version 1.0: `http://www.ark.cs.cmu.edu/LexSem/`

[31] See, e.g., Nunberg et al. (1994, p. 493: "idioms are typically associated with relatively informal or colloquial registers and with popular speech and oral culture"), Moon (1998, p. 267: "[fixed expressions/idioms] can be seen as part of a discourse

Topical category	# docs
Food/restaurant	207
Retail	115
Home services	74
Automotive	73
Medical/dental	52
Entertainment/recreation	45
Travel	44
Health/beauty	30
Pet	16
Other	65
Unsure	2

Perceived sentiment	# docs
++ strongly positive	310
+ positive	214
− negative	88
−− strongly negative	111

Table 3.4: Distribution of review topics and sentiment as coded by one of the annotators.

As the Web Treebank does not provide metadata for reviews, one of our annotators coded all the documents for topic and perceived sentiment. The distribution is shown in table 3.4.

Summary statistics of the MWEs in the corpus are given in table 3.3. Among the highlights:

- The 3,483 MWEs include 15% of all tokens in the corpus. As a point of reference, 17% of all tokens are common nouns.
- 57% of sentences (72% of sentences over 10 words long) and 88% of documents contain at least one MWE.
- 87% of the MWEs are strong/13% are weak.

of familiarity. . . [they can] increase solidarity between the speaker/writer and hearer/reader"), and Simpson and Mendis (2003, p. 434: "possible communicative effects [of idioms] include exaggeration, informality, and rhetorical flair").

- 16% of the MWEs are strong and contain a gold-tagged proper noun—most of these are proper names.
- 73% of the MWEs consist of two tokens; another 21% consist of three tokens.
- 15% of the MWEs contain at least one gap. (Only 6 contain two gaps.[32])
- 65% of the gaps are one word long; another 25% are two words long.
- 1.5% of tokens fall within a gap; 0.1% of tokens belong to an MWE nested within a gap (like *'07 Ford Fusion* and *a little* in figure 3.5).

These figures demonstrate (i) that MWEs are quite frequent in the web reviews genre, and (ii) that annotators took advantage of the flexibility of the scheme to encode gappy expressions and a strength distinction.

Figure 3.7 shows the distribution of intra-MWE and extra-MWE words by part of speech. The MWE words are syntactically diverse: common nouns, verbs, proper nouns, prepositions, adverbs, adjectives, determiners, and particles account for most of them. Nearly all particles and nearly two thirds of proper nouns were marked as part of an MWE.

Categorizing MWEs by their coarse POS tag sequence, we find only 8 of these patterns that occur more than 100 times: common noun–common noun, proper noun–proper noun, verb-preposition, verb-particle, verb-noun, adjective-noun, and verb-adverb. But

[32]They are: *offers*^{1} *a decent bang*$^{1}_{2}$ *for*$^{1}_{2}$ *the buck*$^{1}_{2}$; *take*$_{3}$ *this as*$_{3}$ *far*$_{3}$ *as*$_{3}$ *we can*$_{3}$; *passed*$^{5}_{4}$ *away*$^{5}_{4}$ *silently in*^{5} *his sleep*^{5}; *asked*$_{6}$ *Pomper for*$_{6}$ *my money back*$_{6}$; *putting*$_{7}$ *me at*$_{7}$ *my ease*$_{7}$; *tells*$_{8}$ *me BS to*$_{8}$ *my face*$_{8}$

there is a very long tail—460 patterns in total. For the interested reader, appendix C shows the most frequent patterns, with examples of each.

Many patterns are attested with and without gaps; a handful occur more frequently with gaps than without. About 78% of gaps are immediately preceded by a verb.

There are 2,378 MWE types.[33] 82% of these types occur only once; just 183 occur three or more times. The most frequent are *highly recommend(ed)*, *customer service*, *a lot*, *work with*, and *thank you*. The longest are 8 lemmas long, e.g. *do n't get catch up in the hype* and *do n't judge a book by its cover*.

3.9 Conclusion

We have described a process for shallow annotation of heterogeneous multiword expressions in running text. With this processs, we have created a dataset of informal English web text that has been specifically and comprehensively annotated for MWEs, without reference to any particular lexicon. 6 annotators referred to and improved the guidelines document on an ongoing basis. Every sentence was seen independently by at least 2 annotators, and differences of opinion were discussed and resolved collaboratively. The annotation guidelines and our annotations for the English Web Treebank can be downloaded at: `http://www.ark.cs.cmu.edu/LexSem`.[34]

To the best of our knowledge, this corpus is the first to be freely annotated for more than a handful of kinds of MWEs (without refer-

[33]Our operational definition of *MWE type* combines a strong or weak designation with an ordered sequence of lemmas, using the WordNet API in NLTK (Bird et al., 2009) for lemmatization.

[34]Licensing restrictions prevent us from publishing the full text of every sentence, so we provide annotations in terms of token offsets in the original corpus. Tokens within the span of an MWE are retained.

ence to a lexicon or a set of targeted constructions). The most similar English corpora with shallow lexical semantic representations are not quite as comprehensive in their treatment of MWEs because they focused on a few subclasses (**Wiki50**, §3.3.2) or were created primarily for sense annotation with an existing lexicon (**SemCor**, §3.3.1). Our representation, though also shallow, allows far more flexibility in the configuration of MWEs (arbitrary gaps with limited nesting) and also provides for subclassing in the form of a strong/weak contrast. Our corpus thus creates an opportunity to tackle *general-purpose* MWE identification, such as would be desirable for use by high-coverage downstream NLP systems. An MWE identification system trained on our corpus is presented in ch. 6.

Ch. 4 and 5 offer an approach to enriching lexical segments (single-word or multiword) with semantic class annotations. Future work includes extending the annotation scheme to new datasets; developing semi-automatic mechanisms to detect or discourage inconsistencies across sentences; and integrating complementary forms of annotation of the MWEs (such as syntactic classes). These improvements will facilitate NLP tools in more accurately and informatively analyzing lexical semantics for the benefit of downstream applications.

SPAULDING: I was sittin' in front of the cabin, smoking some meat—
RITTENHOUSE: *Smoking some meat?*
SPAULDING: Yes, there wasn't a cigar store in the neighborhood.

Animal Crackers

BORIS: Sonja—are you scared of dying?
SONJA: Scared is the wrong word. I'm frightened of it.
BORIS: Interesting distinction.

Love and Death

CHAPTER 4

Noun and Verb Supersenses

This chapter:

- Motivates the use of WordNet's *supersense* labels for coarse lexical semantic analysis in context
- Repurposes the existing noun and verb supersense inventories for direct human annotation
- Provides detailed descriptions of the supersense categories in annotation guidelines
- Demonstrates the practicality of supersense annotation in two languages
- Enriches the English multiword expressions corpus (§3.8) with noun and verb supersenses

4.1 Introduction

The previous chapter concerned the determination of *units* of lexical meaning in context. Now we turn to the issue of *categorizing* each lexical expression with semantic labels.

As detailed in ch. 2, two major traditions of lexical semantic labeling are (a) lexicon senses, and (b) named entity classes. In this work we instead use **supersenses**, which like named entities are coarse in *granularity*, making them practical for rapid token annotation with high coverage in a variety of languages/domains. Supersenses, however, are neither restricted to names nor tied to lexicon coverage, which makes for high annotation *density*. This chapter elaborates on an existing inventory of supersenses for **nouns and verbs**, turning superordinate categories within the WordNet hierarchy into a practical annotation scheme that we then tested for Arabic (nouns only, §4.3) and English (nouns and verbs, §4.4). The approach to nouns as applied to Arabic has already been published (Schneider et al., 2012). Working within the same framework, the next chapter tackles the considerably thornier problem of prepositions.

4.2 Background: Supersense Tags

WordNet's **supersense** categories are the top-level hypernyms in the taxonomy (sometimes known as **semantic fields**) which are designed to be broad enough to encompass all nouns and verbs (Miller, 1990; Fellbaum, 1990).[1]

[1] WordNet synset entries were originally partitioned into **lexicographer files** for these coarse categories, which became known as "supersenses." The `lexname` function in WordNet/attribute in NLTK returns the lexicographer file of a given synset. A subtle difference is that a special file called `noun.Tops` contains each noun supersense's root synset (e.g., `group.n.01` for GROUP^) as well as a few miscellaneous synsets, such as `living_thing.n.01`, that are too abstract to fall under any single supersense. In §4.4 we treat the latter cases under an OTHER^ supersense category and merge the former under their respective supersense when processing SemCor (which uses the top-level synsets to mark named entities that are not in WordNet).

The 25 main noun supersense categories are:

(4) NATURAL OBJECT, ARTIFACT, LOCATION, PERSON, GROUP, SUBSTANCE, TIME, RELATION, QUANTITY, FEELING, MOTIVE, COMMUNICATION, COGNITION, STATE, ATTRIBUTE, ACT, EVENT, PROCESS, PHENOMENON, SHAPE, POSSESSION, FOOD, BODY, PLANT, ANIMAL

Appendix D gives several examples for each of these noun tags. (A very small category, OTHER, is sometimes used for miscellaneous cases like *organism*, which include both plants and animals; see footnote 1.)

There are 15 tags for verbs:

(5) BODY, CHANGE, COGNITION, COMMUNICATION, COMPETITION, CONSUMPTION, CONTACT, CREATION, EMOTION, MOTION, PERCEPTION, POSSESSION, SOCIAL, STATIVE, WEATHER

Though WordNet synsets are associated with lexical entries, the supersense categories are unlexicalized. The PERSON category, for instance, contains synsets for *principal*, *teacher*, and *student*. A different sense of *principal* falls under the category POSSESSION. The supersense categories are listed with examples in table 4.1.

As far as we are aware, the supersenses were originally intended only as a method of organizing the WordNet structure. But Ciaramita and Johnson (2003) pioneered the coarse WSD task of **supersense tagging**, noting that the supersense categories provided a natural broadening of the traditional named entity categories to encompass all nouns. Ciaramita and Altun (2006) later expanded the task to include all verbs, and applied a supervised sequence modeling

Noun			
GROUP	1469	*place*	6
PERSON	1202	*people*	1
ARTIFACT	971	*car*	2
COGNITION	771	*way*	4
FOOD	766	*food*	21
ACT	700	*service*	3
LOCATION	638	*area*	8
TIME	530	*day*	9
EVENT	431	*experience*	14
COMMUNICATION	417	*review*	5
POSSESSION	339	*price*	16
ATTRIBUTE	205	*quality*	7
QUANTITY	102	*amount*	13
ANIMAL	88	*dog*	18
BODY	87	*hair*	11
STATE	56	*pain*	10
NATURAL OBJECT	54	*flower*	15
RELATION	35	*portion*	19
SUBSTANCE	34	*oil*	12
FEELING	34	*discomfort*	20
PROCESS	28	*process*	22
MOTIVE	25	*reason*	25
PHENOMENON	23	*result*	17
SHAPE	6	*square*	24
PLANT	5	*tree*	23
OTHER	2	*stuff*	26
all 26 NSSTs	**9018**		

Verb			
STATIVE	2922	*is*	1
COGNITION	1093	*know*	4
COMMUNICATION	974	*recommend*	2
SOCIAL	944	*use*	5
MOTION	602	*go*	6
POSSESSION	309	*pay*	7
CHANGE	274	*fix*	3
EMOTION	249	*love*	11
PERCEPTION	143	*see*	9
CONSUMPTION	93	*have*	12
BODY	82	*get. . . done*	13
CREATION	64	*cook*	10
CONTACT	46	*put*	8
COMPETITION	11	*win*	14
WEATHER	0	—	15
all 15 VSSTs	**7806**		

N/A		
`a	1191	*have*
`	821	*anyone*
`j	54	*fried*

Table 4.1: Summary of noun and verb supersense tagsets. Each entry shows the label, the count and the most frequent lexical item in the **STREUSLE** 2.0 corpus, and the frequency rank of the supersense in the SEMCOR corpus.

framework adapted from NER. (We return to the supersense tagging task in ch. 7.) Evaluation was against manually sense-tagged data that had been automatically converted to the coarser supersenses. Similar taggers have since been built for Italian (Picca et al., 2008) and Chinese (Qiu et al., 2011), both of which have their own WordNets mapped to English WordNet.

We choose WordNet supersenses primarily for continuity with the literature on supersense tagging, though other taxonomies of semantic classes have been explored for coarse WSD (Rayson et al., 2004; Huang and Riloff, 2010; Qadir and Riloff, 2012; Izquierdo et al., 2014), and ad hoc categorization schemes not unlike supersenses have been developed for purposes ranging from question answering (Li and Roth, 2002) to animacy hierarchy representation for corpus linguistics (Zaenen et al., 2004). We believe the interpretation of the supersenses articulated and applied here can serve as a single starting point for diverse resource engineering efforts and applications, especially when fine-grained sense annotation is not feasible.

4.3 Supersense Annotation for Arabic

In Schneider et al. (2012), we decided to test whether the supersense categories offered a practical scheme for direct lexical semantic *annotation* by humans, especially in a language and domain where no high-coverage WordNet is available.[2] Our annotation of Arabic

[2] Even when a high-coverage WordNet is available, we have reason to believe supersense annotation as a first pass would be faster and yield higher agreement than fine-grained sense tagging (though we did not test this). WordNet has a reputation for favoring extremely fine-grained senses, and Passonneau et al.'s (2010) study of the fine-grained annotation task found considerable variability among annotators for some lexemes. The OntoNotes corpus (Hovy et al., 2006) addresses this problem by iteratively merging fine-grained WordNet senses and measuring inter-annotator agreement until an acceptable rate is achieved (Yu et al., 2010), but such an approach is difficult to scale to the full vocabulary.

History	Science	Sports	Technology
Crusades	Atom	2004 Summer Olympics	Computer
Damascus	Enrico Fermi	Christiano Ronaldo	Computer Software
Ibn Tolun Mosque	Light	Football	Internet
Imam Hussein Shrine	Nuclear power	FIFA World Cup	Linux
Islamic Golden Age	Periodic Table	Portugal football team	Richard Stallman
Islamic History	Physics	Raúl Gonzáles	Solaris
Ummayad Mosque	Muhammad al-Razi	Real Madrid	X Window System
434s, 16,185t, 5,859m	*777s, 18,559t, 6,477m*	*390s, 13,716t, 5,149m*	*618s, 16,992t, 5,754m*

Figure 4.1: Domains, (translated) article titles, and sentence, token, and mention counts in the Arabic Wikipedia Supersense Corpus.

Wikipedia articles validated this approach.[3]

The Arabic Wikipedia dataset has subsequently been used to evaluate noun supersense tagging in Arabic via a machine translation projection method (Schneider et al., 2013). That work is not discussed in this thesis—the system presented in ch. 7 is only trained and evaluated for English—but we examine the Arabic annotation process because the methodology was then adapted for English (§4.4).

4.3.1 Arabic Data

28 Arabic Wikipedia articles in four topical domains (history, science, sports, and technology) were selected from Mohit et al.'s (2012) named entity corpus for supersense annotation. The corpus is summarized in figure 4.1.

[3] In an unpublished experiment, Stephen Tratz, Dirk Hovy, Ashish Vaswani, and Ed Hovy used crowdsourcing to collect supersense annotations for English nouns and verbs in specific syntactic contexts (Dirk Hovy, personal communication).

يعتبر	كتاب	جينيس	للأرقام	القياسية	أن	جامعة	القيروان	في
considers	book	Guinness	for-records	the-standard	that	university	Al-Karaouine	in
	COMMUNICATION^					ARTIFACT^		

فاس	المغرب	أقدم	جامعة	في	العالم	حيث	تم	تأسيسها	في
Fez	Morocco	oldest	university	in	the-world	where	was	established	in
LOCATION^			GROUP^		LOCATION^			ACT^	

سنة	859	ميلادي	.
year		AD	
TIME^			

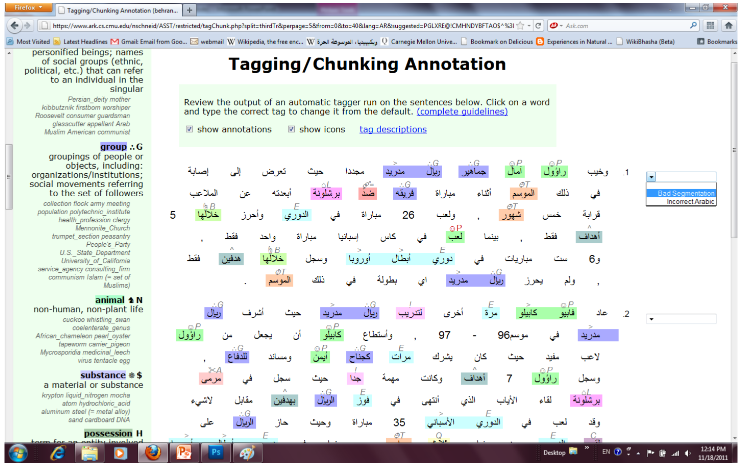
'*The Guinness Book of World Records* considers the University of Al-Karaouine in Fez, Morocco, established in the year 859 AD, the oldest university in the world.'

Figure 4.2: A sentence from the article "Islamic Golden Age," with the supersense tagging from one of two annotators. The Arabic is shown left-to-right.

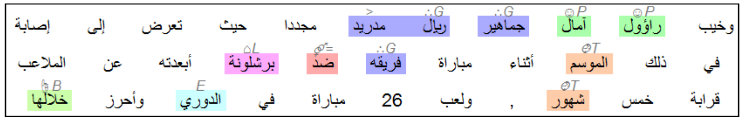
4.3.2 Arabic Annotation Process

This project focused on annotating the free text Arabic Wikipedia data with the 25 noun supersenses of (4) and appendix D. The goal was to mark all common and proper nouns, including (contiguous) multiword names and terms. Following the terminology of NER, we refer to each instance of a supersense-tagged unit as a **mention**. Figure 4.2 shows an annotated sentence (the English glosses and translation were not available during annotation, and are shown here for explanatory purposes only).

We developed a browser-based interactive annotation environment for this task (figure 4.3). Each supersense was assigned an



(a)



(b)

Figure 4.3: Annotation interface for noun supersenses in Arabic.

ASCII symbol; typing that symbol would apply the tag to the currently selected word. Additional keys were reserved for untagging a word, for continuing a multiword unit, and for an "unsure" label. Default tags were assigned where possible on the basis of the previously annotated named entities as well as by heuristic matching of entries in Arabic WordNet (Elkateb et al., 2006) and OntoNotes (Hovy et al., 2006).

Annotators were two Arabic native speakers enrolled as under-

graduates at CMU Qatar. Neither had prior exposure to linguistic annotation. Their training, which took place over several months, consisted of several rounds of practice annotation, starting with a few of the tags and gradually expanding to the full 25. Practice annotation rounds were interspersed with discussions about the tagset. The annotation guidelines, appendix E, emerged from these discussions to document the agreed-upon conventions. The centerpiece of these guidelines is a 43-rule decision list describing and giving (English) examples of (sub)categories associated with each supersense. There are also a few guidelines regarding categories that are particularly salient in the focus domains (e.g., pieces of software in the TECHNOLOGY subcorpus).

Inter-annotator mention F_1 scores after each practice round were measured until the agreement level reached 75%; at that point we started collecting "official" annotations. For the first few sentences of each article, the annotators worked cooperatively, discussing any differences of opinion. Then the rest of the article was divided between them to annotate independently; in most cases they were assigned a few common sentences, which we use for the final inter-annotator agreement measures. This process required approximately 100 annotator-hours to tag 28 articles. The resulting dataset is available at: `http://www.ark.cs.cmu.edu/ArabicSST/`

4.3.2.1 Inter-Annotator Agreement

Agreement was measured over 87 independently-annotated sentences (2,774 words) spanning 19 articles (none of which were used in practice annotation rounds). Our primary measure of agreement, strict inter-annotator mention F_1 (where mentions are required to match in both boundaries and label to be counted as correct), was 70%. Boundary decisions account for a major portion of the disagreement: F_1 increases to 79% if the measure is relaxed to count a match for every pair of mentions that overlap by at least one word. Token-level F_1 was 83%. Further analysis of the frequent tags revealed that the COGNITION category—probably the most heterogeneous—saw much lower agreement rates than the others, suggesting that revising the guidelines to further clarify this category would be fruitful. We also identified some common confusions, e.g. for words like *book* annotators often disagreed whether the physical object (ARTIFACT) or content (COMMUNICATION) was more salient.[4]

4.4 Supersense Annotation for English

As suggested above, supersense tags offer a practical semantic label space for an integrated analysis of lexical semantics in context. For English, we have created the **STREUSLE**[5] dataset, version 2.0 of which fully annotates the **REVIEWS** corpus for WordNet's noun and verb supersenses as well as multiword expressions (ch. 3). (A new inventory of supersenses for prepositions will be applied to the same corpus: ch. 5.)

In developing the methodology for supersense annotation with Arabic Wikipedia, we predicted that it would port well to other languages and domains. Experience with English web reviews has borne this out. We generally adhered to the same supersense annotation process; the most important difference was that the data had already been annotated for MWEs, and supersense labels apply to any strong MWEs as a whole. The same annotators had already done the MWE annotation; whenever they encountered an apparent mistake from an earlier stage (usually an oversight), they were encouraged to correct it. The more sophisticated annotation interface used for English supports modification of MWEs as well as supersense labels in one view.

[4]Additional details and analysis are reported by Schneider et al. (2012).

[5]Supersense-Tagged Repository of English with a Unified Semantics for Lexical Expressions

Most of the supersense annotation was broken into separate rounds: first we annotated nearly the entire **REVIEWS** corpus for noun supersenses; then we made another pass to annotate for verbs. This was decided to minimize cognitive load when reasoning about the tagsets. Roughly a tenth of the sentences were saved for a combined noun+verb annotation round at the end; annotators reported that constantly switching their attention between the two tagsets made this mode of annotation more difficult.

4.4.1 Nouns

Targets. According to the annotation standard, all noun singletons[6] and noun-headed[7] MWEs should receive a noun supersense label. Annotation targets were determined heuristically from the gold (PTB-style) POS tags in the corpus: all lexical expressions containing a noun[8] were selected. This heuristic overpredicts occasionally because it does not check the syntactic head of MWEs. For this round, the backtick symbol (`) was therefore reserved for MWEs (such as light verb constructions) that should not receive a noun label.[9] The annotation interface prohibited submission of blank annotation targets to avoid oversights.

Interface. Instead of the interface used for Arabic annotation, we extended the online MWE annotation tool (figure 3.6) to also sup-

[6]But not pronouns.

[7]Headedness within lexical expressions is not marked as part of our annotation scheme, but annotators are expected to be able to recognize the head in order to determine the set of supersense candidates for the full expression.

[8]Specifically, any POS tag starting with N or ADD (web addresses).

[9]Pronouns like *anything* also fall into this category because they are POS-tagged as nouns.

port supersense labeling of units. This is visualized in figure 4.4. Specifically, singletons and strong MWEs may receive labels (subject to a POS filter). This allows the two types of annotation to be worked on in tandem, especially when a supersense annotator wishes to change a multiword grouping. Additionally, the tool provides a complete version history of the sentence and a "reconciliation" mode that merges two users' annotations of a sentence, flagging any differences for manual resolution; these features are extremely useful when breaking the annotation down into several rounds among several annotators.

Before any annotation is saved, the tool will validate that its MWE analysis and labels are valid and compatible with one another. The set of valid labels is prespecified to consist of the supersenses, ` to mark the supersenses as not applicable to a lexical expression, and ? to indicate uncertainty. As the user begins to type a label, an autocomplete dropdown menu with possible matches will be displayed. Every identified target must receive some label.

Tagset conventions. Even though the annotation guidelines were already established from the Arabic effort, the English annotators were new to the scheme, so we devoted several brief annotation rounds to practicing it and reaching agreement in the reviews domain. Metonymy posed the chief difficulty in this domain: institutions with a premises (such as restaurants, hotels, schools, and offices) are frequently ambiguous between a GROUP^ reading (institution as a whole), an ARTIFACT^ reading (the building), and a LOCATION^ (site as a whole). Our convention was to choose whichever reading seemed most salient in context: for example, a statement about the the quality of a restaurant's service would be

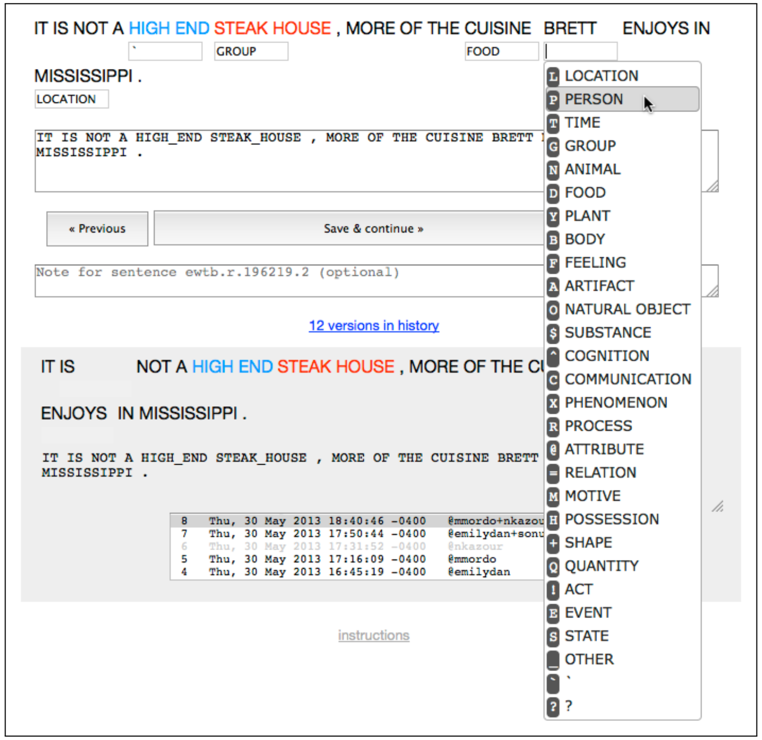


Figure 4.4: Interface for noun supersense annotation. Previous annotation versions can be browsed in the gray box.

the GROUP^ reading of *restaurant*.[10] Many of these decisions may be subjective, however, which probably indicates a limitation of the scheme in that it cannot always capture multifaceted frame-based concepts.

4.4.2 Verbs

Targets. The set of lexical expressions that should receive a verb supersense label consists of (a) all verb singletons that are not auxiliaries, and (b) all verb-headed MWEs. Again, simple but overly liberal heuristics were used to detect annotation targets,[11] so wherever the heuristics overpredicted, annotators entered:

- `` `a `` for auxiliary verbs
- `` `j `` for adjectives (some *-ing* and *-ed* adjectives are POS-tagged as VBG and VBD, respectively)
- `` ` `` for all other cases

Interface. The same interface was used as for nouns. Figure 4.5 shows the dropdown list for verbs. For MWEs containing both a noun and a verb, all the noun and verb labels were included in the dropdown and accepted as valid.

Tagset conventions. We developed new guidelines to characterize the verb supersenses for use by annotators. The guidelines document appears in appendix F. It is similar to the guidelines for nouns (appendix E), but is shorter (as there are only 15 verb supersenses) and formulates priorities as precedence relations between the cat-

[10] This rule is sometimes at odds with WordNet, which only lists ARTIFACT for *hotel* and *restaurant*.

[11] All lexical expressions containing a POS tag starting with V.

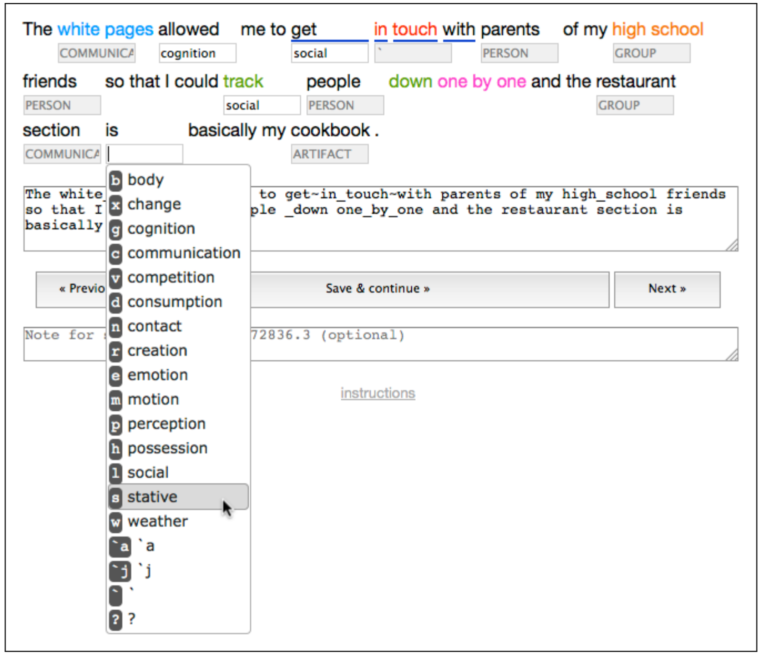


Figure 4.5: Annotation interface, with dropdown menu for verb supersenses.

egories rather than as a decision list. For instance, the precedence rule

(6) {PERCEPTIONˇ, CONSUMPTIONˇ} > BODYˇ > CHANGEˇ

demands that verbs of perception or consumption (*hear*, *eat*, etc.) be labeled as such rather than the less specific class BODYˇ. The precedence rules help to resolve many of the cases of meaning overlap between the categories. These guidelines were developed over several weeks and informed by annotation difficulties and disagreements.

4.4.3 Corpus Statistics

A total of 9,000 noun mentions and 8,000 verb mentions incorporating 20,000 word tokens are annotated. Table 4.1 displays supersense mention counts as well as the most frequent example of each category in the corpus. As a point of reference, it also shows the frequency rank of the supersense in **SemCor**—note, for instance, that FOOD^ is especially frequent in the **REVIEWS** corpus, where it ranks fifth among noun supersenses (vs. 21st in SemCor).

4.5 Related Work: Copenhagen Supersense Data

An independent English noun+verb supersense annotation effort targeting the Twitter domain was undertaken by the COASTAL lab at the University of Copenhagen (Johannsen et al., 2014). The overarching goal of annotating supersenses directly in running text was the same as in the present work, but there are three important differences. First, general-purpose MWE annotation was not a goal in that work; second, sentences were pre-annotated by a heuristic system and then manually corrected, whereas here the MWEs and supersenses are supplied from scratch; and third, Johannsen et al. (2014) provided minimal instructions and training to their annotators, whereas here we have worked hard to encourage consistent interpretations of the supersense categories. Johannsen et al. have released their annotations on two samples of tweets (over 18,000 tokens in total).[12]

[12] https://github.com/coastalcph/supersense-data-twitter

Johannsen et al.'s (2014) dataset provides a good illustration of why supersense annotation by itself is not the same as the full scheme for lexical semantic analysis proposed here. Many of the expressions that they have supersense-annotated as single-word nouns/verbs probably would have been considered larger units in MWE annotation. However, examining the Johannsen et al.'s in-house sample, multiword chunks arguably should have been used for verb phrases such as *gain entry*, *make sure*, and *make it* ('succeed'), and for verb-particle constructions such as *take over*, *find out*, and *check out* ('ogle'). In the traditional supersense annotation scheme, there are no chunks not labeled with a supersense; thus, e.g., PPs such as *on tap*, *of ALL-Time*, and *up to [value limit]* are not chunked.

Many of the nominal expressions in Johannsen et al.'s (2014) data appear to have overly liberal boundaries, grouping perfectly compositional modifiers along with their heads as a multiword chunk: e.g., *Panhandling Ban*, *Loudoun Firefighters*, *Panda Cub*, *farm road crash*, *Sri Lanka's west coast*, and *Tomic's dad*. Presumably, some of these were boundary errors made by the heuristic pre-annotation system that human annotators failed to notice.

4.6 Conclusion

This chapter has described WordNet-based noun and verb supersenses from the perspective of annotation. Supersenses offer coarse-grained and broadly applicable semantic labels for lexical expressions and naturally complement multiword expressions in lexical semantic analysis. We have developed detailed annotation criteria using the existing supersense categories and applied them to annotate text corpora in Arabic and English. Our representation of supersenses dovetails with the scheme for multiword expressions proposed in the previous chapter. English annotations for the **REVIEWS** corpus will be released as the **STREUSLE** 2.0 dataset, which forms the basis of our integrated lexical semantic analyzer in ch. 7.

You can leave in a taxi. If you can't get a taxi, you can leave in a huff. If that's too soon, you can leave in a minute and a huff.

Firefly in *Duck Soup*

JOSH: You went over my head, and you did it behind my back.
AMY: Quite the contortionist am I.

The West Wing, "Dead Irish Writers" (Season 3, Episode 15)

CHAPTER 5

Preposition Supersenses

This chapter:

- Illustrates the extreme semantic polysemy of English prepositions
- Summarizes resources and NLP approaches to preposition semantics developed to date, and their limitations
- Argues for treating preposition semantics in the framework of supersenses, which can be integrated with multiword expressions and noun and verb supersenses in a lexical semantic analysis
- Proposes a novel supersense taxonomy and annotation scheme for prepositions, drawing on an existing non-hierarchical coarse grouping of English preposition senses and an existing hierarchy for semantic roles of verbs
- Offers concrete criteria for selecting annotation targets
- Describes notable decisions in the new supersense scheme, including new portions of the hierarchy devoted to temporal, path, value, and comparison subcategories

Prepositions are perhaps the most beguiling yet pervasive lexicosyntactic class in English. They are everywhere (figure 5.1); their functional versatility is unrivaled and largely idiosyncratic (7). They are nearly invisible, yet present in some of the most quoted lines of text.[1] In a way, prepositions are the bastard children of lexicon and grammar, rising to the occasion almost whenever a noun-noun or verb-noun relation is needed and neither subject nor object is appropriate. Consider the many uses of the word *to*, just a few of which are illustrated in (7):

(7) a. My cake is **to** die for. *(nonfinite verb idiom)*
b. If you want I can treat you **to** some. *(prepositional verb idiom[2])*
c. How about this: you go **to** the store *(locative goal)*
d. **to** buy ingredients. *(nonfinite purpose)*
e. That part is up **to** you. *(responsibility)*
f. Then if you give the recipe **to** me *(recipient)*
g. I'm happy **to** make the batter *(nonfinite adjectival complement)*

[1] Observe that a preposition is:

- the 1st word in *Genesis* ("**In** the beginning"), the *Quran* ("**In** the name **of** the merciful and compassionate God."), *Paradise Lost* ("**Of** Mans First Disobedience"), *Oliver Twist* ("**Among** other public buildings **in** a certain town"), *Don Quixote* ("**En** un lugar **de** la Mancha"/"**In** some village **in** la Mancha"), and *The Hobbit* ("**In** a hole **in** the ground there lived a hobbit.")
- the 2nd word of the *Magna Carta* ("John, **by** the grace **of** God"), the *U.S. Declaration of Independence* ("When **in** the course **of** human events"), and *The Hitchhiker's Guide to the Galaxy* ("Far **out in** the uncharted backwaters")
- the 3rd word of *Peter Pan* ("All children, **except** one, grow **up**.") and *Lord of the Flies* ("The boy **with** fair hair lowered himself **down** the last few feet **of** rock")
- the 4th word of the *U.S. Constitution* ("We the people **of** the United States")
- the 5th word of *Romeo and Juliet* ("Two households, both alike **in** dignity") and *Richard III* ("Now is the winter **of** our discontent / Made glorious summer **by** this sun **of** York")

[2] The lexical item *treat... to* is from (Huddleston, 2002, p. 279).

.	2,650	is	633	**with**	384	n't	254	-	191
the	2,186	it	573	not	351	had	246	so	185
and	1,782	**for**	549	we	305	service	241	good	181
,	1,618	my	544	**on**	293	do	236	**from**	171
i	1,528	they	510	are	276	be	219	food	165
to	1,352	you	504	great	273	there	208	if	165
a	1,175	!	486	but	265	he	195	"	163
of	660	that	433	me	261	were	193	's	163
was	644	this	393	**at**	260	would	193	all	161
in	635	have	385	very	255	place	191	**as**	160

Figure 5.1: Counts of the top 50 most frequent words in REVIEWS. Prepositions are bolded; others in the top 100 include *up* (#61), *about* (#68), *back* (#74), *by* (#86), and *after* (#96).

h. and put it in the oven for 30 **to** 40 minutes *(range limit)*
i. so you will arrive **to** the sweet smell of chocolate. *(background event)*
j. That sounds good **to** me. *(affective/experiencer)*
k. I hope it lives up **to** your expectations. *(prepositional verb idiom)*
l. That's all there is **to** it. *(phrasal idiom)*

Sometimes a preposition specifies a relationship between two entities or quantities, as in (7h). In other scenarios it serves a case-marking sort of function, marking a complement or adjunct—principally to a verb, but also to an argument-taking noun or adjective (7g). As we have seen in ch. 3, prepositions play a key role in multiword expressions, as in (7a), (7l), the prepositional verbs in (7b) and (7k), and arguably (7e).

This chapter briefly introduces the preposition from syntactic, semantic, and multilingual perspectives, then reviews the literature on resources and NLP for preposition semantics (§5.1). §5.2 then introduces our approach. Its main components are (a) integrating the multiword expression analysis developed in ch. 3; (b) targeting a broad range of tokens based on syntactic and lexical criteria, as detailed in §5.3; and (c) proposing in §5.4 a hierarchical taxonomy of preposition supersenses that combines the preposition inventory of Srikumar and Roth (2013a) with VerbNet's (Kipper et al., 2008) taxonomy of thematic roles (Bonial et al., 2011; VerbNet Annotation Guidelines; Hwang, 2014, appendix C), while also drawing insights from AMR's (Banarescu et al., 2013) inventory of non-core roles (Banarescu et al., 2014).

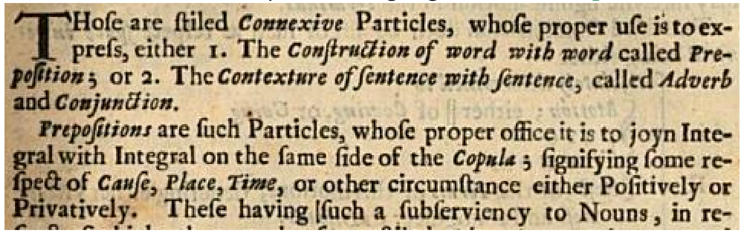
A wiki documenting our scheme in detail can be accessed at `http://tiny.cc/prepwiki`. It contains mappings from fine-grained senses (of 34 prepositions) to our supersenses, as well as numerous examples. The structured format of the wiki is conducive to browsing and to exporting the examples for display in our annotation tool. From our experience with pilot annotations, we believe that the scheme is fairly stable and broadly applicable: preposition tokens for which it is difficult to choose a supersense label are relatively rare. Ultimately, we aim to annotate the English REVIEWS corpus to augment the MWE and verb/noun supersense annotations in our **STREUSLE** dataset (§4.4) with preposition supersenses for version 3.0.

5.1 Background

5.1.1 What is an English Preposition?

It is generally understood that the category of English prepositions includes such words as *to*, *for*, and *with*. As with much in grammar, however, the precise contours of this category are difficult to pin down.

An early definition comes from a 1668 treatise by English clergyman and philosopher John Wilkins entitled *An Essay Towards a Real Character And a Philosophical Language*. We read on p. 309:[3]


THoſe are ſtiled *Connexive* Particles, whoſe proper uſe is to expreſs, either 1. The *Conſtruction of word with word* called *Prepoſition*; or 2. The *Contexture of ſentence with ſentence*, called *Adverb* and *Conjunction*.

Prepoſitions are ſuch Particles, whoſe proper office it is to joyn Integral with Integral on the ſame ſide of the *Copula*; ſignifying ſome reſpect of *Cauſe*, *Place*, *Time*, or other circumſtance either Poſitively or Privatively. Theſe having [ſuch a ſubſerviency to Nouns, in re-

That is, prepositions are connectives that join together content words[4] (as opposed to sentences) to express "some respect of *Cause*, *Place*, *Time*, or other circumstance."

A contemporary articulation of the traditional grammar view of prepositions can be found in the *Merriam-Webster Learner's Dictionary*, where it is defined as "a word or group of words that is used with a noun, pronoun, or noun phrase to show direction, location, or time, or to introduce an object." The *Oxford English Dictionary* gives a similar definition: "An indeclinable word or particle governing (and usu. preceding) a noun, pronoun, etc., and expressing a relation between it and another word."

The traditional definition applies uncontroversially to the underlined PPs in the following passage:

(8) She's gonna meet him **at** the top **of** the Empire State Building. Only she got hit **by** a taxi. . . . And he's too proud **to** {find **out** why she doesn't come}. But he comes **to** {see her} anyway. . . . he doesn't even notice that she doesn't get **up to** {say hello}. And

[3] Image from Google Books.

[4] Per the *Oxford English Dictionary*, the term *integral* is "Applied by Wilkins to those words or parts of speech which of themselves express a distinct notion, as distinct from those which express relations between notions."

he's very bitter. And you think that he's just gonna—walk **out** the door and never know why, she's just lying there, you know, like **on** the couch **with** {this blanket **over** her shriveled little legs}. . . . And he, he like goes **into** the bedroom. And he looks and he comes **out** and he looks **at** her, and he kinda just—they know, and then they hug. [*Sleepless in Seattle*]

Those PPs express several spatial relations (*on the couch*, *out the door*, etc.), and one expresses a causal relation (*hit by a taxi*). However, the traditional definition leaves out instances of *to*, *up*, *with*, and *out* in (8) that do not take a noun phrase complement:

- *to* followed by a verb phrase, where the *to* marks the verb as an infinitive
- *up* used following *get* (traditional grammar might deem this an adverb)
- *with* marking a state of affairs expressed in a clause (this is traditionally called a subordinating conjunction)
- *out* in *comes out* (of the bedroom), which is not followed by a noun phrase

Pullum and Huddleston (2002, p. 603) argue for a more inclusive definition of preposition, pointing out that alternations such as *walk out the door* vs. *come[] out*—where *out* has the same semantic function, expressing that the room was exited—are possible for many of the prepositions. Emphasizing the class of word types rather than the syntactic environment, they define preposition as:

> a relatively closed grammatically distinct class of words whose most central members characteristically express spatial relations or serve to mark various syntactic functions and semantic roles.

In Pullum and Huddleston's terminology, if *out*, *up*, *down*, *off*, etc. appears with no complement, it is an *intransitive* preposition forming a one-word PP—not a member of an entirely distinct word class. Some prepositions are always transitive (e.g., *at*, *beside*; Pullum and Huddleston, 2002, pp. 635–6), or almost always (e.g., *to*, *for*[5]). Others—such as *together* and *back*, both closed-class items with schematic spatial meanings—can never take NP complements (Pullum and Huddleston, 2002, p. 614). (This is similar to the case with verbs, some of which are always transitive, some of which are always intransitive, and some of which are, shall we say, transiflexible.[6]) Thus, they use the term **particle** to describe a word's role in a particular construction, not (*pace* Penn Treebank's RP) as a distinct part of speech.[7]

All of this is just a flavor of the descriptive challenges raised by prepositions; see Saint-Dizier (2006b) for an assortment of syntactic and semantic issues.

As noted above, prepositions figure prominently in several well-studied classes of multiword expressions: verb-particle constructions, prepositional verbs, and determinerless PPs (§3.2.3). In our MWE annotation, the category of prepositional verbs proved especially troublesome (§3.7). Undertaking to systematically analyze the functions of prepositions may therefore help us to separate the pro-

[5]The only exceptions that we know of are the lexical idioms *come to* 'regain consciousness' and *pull. . . to* 'pull (a door, etc.) closed' (Pullum and Huddleston, 2002, p. 613); *done for* 'doomed'; and *talking to* and *what for*, compound nouns meaning 'scolding'.

[6]Some linguists prefer the term **ambitransitive**. They are not much fun at parties.

[7]The literature on verb-particle constructions (see ch. 3) sometimes takes a more inclusive view, equating "particle" with "intransitive preposition." Huddleston (2002, p. 280) instead refer to the **verb-particle-object construction** (in which the particle may be positioned either before or after the verb's object NP), pointing out a few non-prepositions that can serve as particles in this construction. For instance: *cut **short*** (adjective) and *let **go*** (verb).

ductive and compositional cases from the exceptional, MWE-worthy cases.

5.1.2 Linguistic Approaches

Studies of prepositions appear to be relatively rare in the linguistics literature,[8] especially outside of the spatial domain. E.g., I am aware of but a few edited volumes on the subject (Rauh, 1991; Zelinsky-Wibbelt, 1993b; Cuyckens and Radden, 2002; Feigenbaum and Kurzon, 2002; Saint-Dizier and Ide, 2006; Kurzon and Adler, 2008).

The lexical-vs.-functional dimension and, relatedly, the degree of association between prepositions and other words (especially verbs) used in combination has received some theoretical attention (e.g., Bolinger, 1971; Vestergaard, 1977; Rauh, 1993; O'Dowd, 1998; Tseng, 2000) but without (it seems to me) any clear and robust diagnostics that could be incorporated into an annotation scheme.

The structured polysemy analysis of *over* put forward by Brugman (1981) and elaborated by Lakoff (1987, pp. 416–461), Dewell (1994), Tyler and Evans (2003, ch. 4), and Deane (2005) has been influential within cognitive linguistics. (See also footnote 28.) Working in this tradition, Lindstromberg (2010) examines over 90 English prepositions, considering in detail the schematic spatial situations that can be expressed and the ways in which these motivate non-spatial extensions. Chapter 21 gives an inventory of about 75 "non-spatial notions"—these are not unlike the categories we will adopt below, though some are quite fine-grained: e.g., BEING RESOLVED, FIXED as in *pin him* ***down*** vs. BEING UNRESOLVED, UNDECIDED as in *everything's still* ***up*** *in the air*. The extent to which annotators could

[8]This excludes the many dictionaries and pedagogical materials (especially, for second language learners) on preposition-bearing constructions such as phrasal verbs.

be trained to agree on Lindstromberg's detailed categorization is unknown.

5.1.3 Other Languages

Crosslinguistic variation in prepositions and spatial categorization systems has received considerable attention from theorists (Bowerman and Choi, 2001; Hagège, 2009; Regier, 1996; Xu and Kemp, 2010; Zelinsky-Wibbelt, 1993a) but is of practical interest as well, especially when it comes to machine translation (see §8.3.3) and second language acquisition (§8.3.4). A corpus creation project for German preposition senses (Müller et al., 2010, 2011) is similar in spirit to the supersense approach taken below. Finally, the PrepNet resource (Saint-Dizier, 2006a) aimed to describe the semantics of prepositions across several languages; however, it seems not to have progressed beyond the preliminary stages.

5.1.4 Preposition Resources

The following corpus resources contain semantic categorizations that apply to English prepositions:

The Penn Treebank. As detailed by O'Hara and Wiebe (2009), the PTB since version II (Marcus et al., 1994) has included a handful of coarse function tags (such as LOCATION and TIME) that apply to constituents, including PPs.

FrameNet. Semantic relationships in FrameNet (Baker et al., 1998) are organized according to scenes, known as **frames**, that can be evoked by predicates in a sentence. Each frame defines roles, or **frame elements**, which indicate possible facets along which the description of the scene can be elaborated with **arguments** in the sentence. Many roles are highly specific to a single frame, while others are quite generic. Arguments are often realized as PPs, thus the frame element labels can be interpreted as disambiguating the function of the preposition.

The Preposition Project (TPP). This is an English preposition lexicon and corpus project (Litkowski and Hargraves, 2005) that adapts sense definitions from the *Oxford Dictionary of English* and applies them to prepositions in sentences from corpora. A dataset for the SemEval-2007 shared task on preposition WSD (Litkowski and Hargraves, 2007) was created by collecting FrameNet-annotated sentences (originally from the BNC) and annotating 34 frequent preposition types with a total of 332 attested senses.[9] TPP now incorporates

[9]The SemEval-2007 sentences—of which there are over 25,000, each with a single preposition token annotated—were collected from FrameNet's lexicographic annotations that had been selected by a FrameNet lexicographer to illustrate, for a given lexical unit, a valence pattern with a particular kind of PP. In FrameNet data, such sentences are grouped under a `ppX` label (`ppto`, `ppfor`, etc.). Preposition types having too few of these PP exemplars were filtered out, leaving 34 types in the SemEval-2007 data (Litkowski, 2013).

The FrameNet lexicographic exemplars were handpicked from the BNC to illustrate, e.g., the range of valence patterns for a predicate; usages with few arguments are underrepresented and rare patterns are overrepresented. Biases in the lexicographic exemplars have been found to distort statistical models trained on them (e.g., Das et al., 2014). Further, only a fraction of PPs from these exemplars constitute frame element fillers (arguments to an annotated frame), and only a small proportion of *those* were highlighted under a `ppX` label. Therefore, while the SemEval-2007 sentences illustrate a great variety of preposition usages, it is important to note that the dataset is not statistically representative—as evaluation data it is not a realistic yardstick for performance on a real corpus, and it cannot be assumed to capture the full semantic range of PPs in FrameNet, let alone prepositions in English.

Recognizing this, Litkowski (2013) has initiated an effort to extend TPP annotations to a new, statistically representative corpus. Our approach is intended to complement that effort by facilitating rapid and comprehensive annotation of corpora at a coarser level of granularity. By recording many-to-many correspondences between TPP senses and supersenses, we can ensure partial (but nondeterministic) compatibility between the two annotation schemes, which should allow models to make use of

additional prepositions and resources, with new annotated corpora under development (Litkowski, 2013, 2014).

Dahlmeier et al. To learn and evaluate their joint model of semantic roles and preposition senses, Dahlmeier et al. (2009) annotated TPP senses in the PropBank WSJ corpus for 7 high-frequency prepositions (*of*, *in*, *for*, *to*, *with*, *on*, and *at*). This amounted to 3,854 statistically representative instances in the news domain. The inter-annotator agreement rate was estimated at 86%, which suggests that clearly applicable TPP senses are available for the preponderance of tokens, but gives little insight into TPP's suitability for rare or borderline usages.[10]

Tratz. Tratz (2011, ch. 4) refined the TPP sense inventory for the SemEval-2007 corpus with the goal of improving its descriptive adequacy and measuring inter-annotator agreement for all 34 prepositions. The refinement was performed by two annotators, who reorganized and reworded the sense definitions and reannotated instances in an iterative fashion until agreement was qualitatively high. The total number of senses was reduced from 332 to 278, though a few prepositions gained additional senses. A third annotator was then added for final estimation of inter-annotator agreement. Pairwise agreement rates, Fleiss' κ, and per-annotator sense entropies are reported for each preposition. Tratz also reports supervised classification results with the original vs. refined sense inventories.

Srikumar and Roth (S&R). Srikumar and Roth (2013b) confront the problem of predicting preposition token *relations*, i.e., the prepo-

both kinds of data.

[10]Like TPP but unlike our approach, Dahlmeier et al.'s (2009) annotations were restricted to prepositions heading transitive PPs.

sition's governor, object, and semantic label. For their experiments, Srikumar and Roth coarsen the original TPP SemEval-2007 sense annotations into 32 categories determined semi-automatically (the fine-grained senses were clustered automatically, then the clusters were manually refined and given names). Detailed in Srikumar and Roth (2013a), those categories cut across preposition types to combine related TPP senses for better data-driven generalization. Cohen's κ for inter-annotator agreement was an estimated 0.75, which is encouraging, though it is unclear whether the disagreements were due to systematic differences in interpretation of the scheme or to difficulty with rare preposition usages. We shall return to this scheme in §5.4 below.

5.1.5 Prepositions in NLP

Despite a steady trickle of papers over the years (see Baldwin et al., 2009 for a review), there is no apparent consensus approach to the treatment of preposition semantics in NLP. Studies have examined preposition semantics within multiword expressions (Cook and Stevenson, 2006), in spatial relations (Hying, 2007), across languages (Saint-Dizier, 2006a), in nonnative writing (Chodorow et al., 2007), in semantic role labeling (Dahlmeier et al., 2009), in vector space models (Zwarts and Winter, 2000), and in discourse (Denand and Rolbert, 2004).

Preposition sense disambiguation systems have been evaluated against one or more of the resources described in §5.1.4 (O'Hara and Wiebe, 2003, 2009; Ye and Baldwin, 2007; Dahlmeier et al., 2009; Tratz and Hovy, 2009; Hovy et al., 2010, 2011; Srikumar and Roth, 2013b). Unfortunately, all of these resources are problematic. Neither the PTB function tags nor the FrameNet roles were designed with prepositions in mind: the former set is probably not comprehensive enough to be a general-purpose account of prepositions, and the latter representation only makes sense in the broader analytical framework of frame semantics, which we believe should be treated as a separate problem (Das et al., 2014, §8.3.2). The Preposition Project data, though extensive, were selected and annotated from a lexicographic, type-driven perspective—i.e. with the goal of describing and documenting the uses of individual prepositions in a lexical resource rather than labeling a corpus with free-text preposition annotations (footnote 9; cf. **SemCor**, §2.3.2). A **token-driven** approach would be more in line with the philosophy advocated here for lexical semantic annotation and modeling.[11]

5.2 Our Approach to Prepositions

As a "sweet spot" between linguistic descriptiveness and practicality for annotation, we approach preposition semantics much like the noun and verb supersenses in the previous chapter. The descriptive steps are therefore:

1. **Lexical segmentation:** Mark any multiword expressions, as in ch. 3.

2. **Preposition targets:** Identify any single-word prepositions, as well as any MWEs headed by a preposition, as candidates for receiving a preposition tag.

[11] A technical reason that the type-driven approach to annotation is not ideal for learning NLP systems is the i.i.d. assumption typically made in machine learning. If a sample is not random but biased by an annotator's interest in covering as many phenomena as possible, this bias will be evident in predictions made by a learned model. As an example, Das et al. (2014) mention that including a large number of FrameNet lexicographic annotations (on handpicked sentences from a corpus) in the training data for a frame-semantic parser actually hurt performance when evaluated on a statistically representative corpus.

3. **Preposition tagging:** Assign a preposition supersense label to each of the preposition targets.

The procedure for identifying targets is described in §5.3, and the supersense inventory in §5.4.

5.3 Preposition Targets

From a lexical segmentation, we first have to decide which lexical expressions should be considered as candidates for receiving a preposition annotation. Though for the REVIEWS corpus (Bies et al., 2012) we have Penn Treebank–style part-of-speech tags, those tags do not entirely line up with our definition of preposition (see §5.1.1)—for example, *apart* is consistently tagged as an adverb, but most adverbs are not prepositions. Given that prepositions form a relatively closed class (Pullum and Huddleston, 2002), we have compiled a list of 235 single words that can function as prepositions.[12] It is as follows:

(9) 2, 4, a, abaft, aboard, about, above, abreast, abroad, absent, across, adrift, afore, aft, after, afterward, afterwards, against, agin, ago, aground, ahead, aloft, along, alongside, amid, amidst, among, amongst, an, anent, anti, apart, apropos, apud, around, as, ashore, aside, aslant, astraddle, astride, asunder, at, athwart, atop, away, back, backward, backwards, bar, barring, before, beforehand, behind, below, beneath, beside, besides, between, betwixt, beyond, but, by, c., cept, chez, circa, come, concerning, considering, counting, cum, dehors, despite, down, downhill, downstage, downstairs, downstream, downward, downwards, downwind, during, eastward, eastwards, ere, ex, except, excepting, excluding, failing, following, for, forbye, fore, fornent, forth, forward, forwards, frae, from, gainst, given, gone, granted, heavenward, heavenwards, hence, henceforth, home, homeward, homewards, in, including, indoors, inside, into, inward, inwards, leftward, leftwards, less, like, mid, midst, minus, mod, modulo, mongst, near, nearby, neath, next, nigh, northward, northwards, notwithstanding, o', o'er, of, off, on, onto, onward, onwards, opposite, out, outdoors, outside, outta, outward, outwards, outwith, over, overboard, overhead, overland, overseas, overtop, pace, past, pending, per, plus, pon, post, pro, qua, re, regarding, respecting, rightward, rightwards, round, sans, save, saving, seaward, seawards, since, skyward, skywards, southward, southwards, than, thenceforth, thro', through, throughout, thru, thruout, thwart, 'til, till, times, to, together, touching, toward, towards, under, underfoot, underground, underneath, unlike, until, unto, up, uphill, upon, upside, upstage, upstairs, upstream, upward, upwards, upwind, v., versus, via, vice, vis-a-vis, vis-à-vis, vs., w/, w/i, w/in, w/o, westward, westwards, with, withal, within, without

[12] This list was drawn primarily from the *Cambridge Grammar of the English Language* (Pullum and Huddleston, 2002), the *Oxford English Grammar* (Greenbaum, 1996), the Preposition Project (Litkowski and Hargraves, 2005), and the Wikipedia article "List of English Prepositions" (`https://en.wikipedia.org/wiki/List_of_English_prepositions`).

Note that this list includes alternate/nonstandard spellings (e.g., *2* for *to*) and words that are more commonly other parts of speech, but can act as prepositions in certain constructions (*like*, *post*, etc.). We therefore use POS tags in combination with lexical matching to automatically identify preposition candidates, according to the following rule:

(10) A single-word token is considered a preposition target if it meets either of the following criteria:

a. Its POS tag is RP (verb particle) or TO (the word *to*)

b. Its POS tag is IN (transitive preposition or subordinator) or RB (adverb), and the word is listed in (9).

We are also interested in analyzing multiword prepositions (i.e., multiword expressions that function as prepositions). While this is a more difficult class to circumscribe, it is difficult to come up with an example of a multiword preposition that does not contain a word from (9)—in fact, the TPP and Wikipedia lists include several dozen multiword prepositions, all of which indisputably contain a single-word preposition type: these include *out of*, *next to*, *on top of*, *in lieu of*, *along with*, *due to*, *thanks to*, *except for*, *such as*, and *as regards*. Therefore, we adopt the procedure in (11):

(11) A strong MWE instance is considered a preposition target if it meets either of the following criteria:

a. The MWE begins with a word that matches the criteria of (10).

b. The MWE contains a word matching the criteria of (10), and begins with one of the following words (all of which begin a multiword preposition in TPP): a, according, all, bare, because, but, care, complete, contrary, courtesy, depending, due, exclusive, inclusive, instead, irrespective, little, more, next, nothing, other, outboard, owing, preparatory, previous, prior, pursuant, regardless, relative, short, subsequent, thanks, this

The main reason to use a whitelist in (11b) is to avoid identifying prepositional verbs as preposition supersense candidates. Thus far, these heuristics applied to our data seem to be successful at identifying everything we want to annotate as a preposition (good recall) without too many false positives (good precision).

5.3.1 Special Syntactic Categories

There is a certain amount of lexical and semantic overlap between prepositions that serve as heads of prepositional phrases, and the category of **subordinators** (or **subordinating conjunctions**), which serve to link clauses. Words in the overlapping group include *for*, *with*, and *as*. The IN POS category includes such cases; however, we have decided to prioritize in our annotation (a) prepositions with NP complements, (b) intransitive prepositions, and (c) infinitival *to*. For all other cases automatically flagged as targets—including

words with clausal complements, and connectives like *as well as*—annotators are instructed to mark the expression as not applicable to preposition annotation. Special cases include:

5.3.1.1 Infinitival *to*

We are interested in infinitival *to* where it marks a PURPOSE’ or FUNCTION’.[13] More commonly, however, infinitival *to* serves a purely syntactic function, which we designate with a special label (`` `i ``).

5.3.1.2 Subordinating *for*, *with*, and *as*

In sentences like *Unity is not possible* ***with*** *John sitting on the throne* and ***For*** *him to abdicate would have been unpredecented*, we analyze *with* and *for* as subordinators: these constructions are unlike intransitive particles or transitive PPs.[14] The ` label is used to mark these as non-prepositions.

5.3.1.3 Discourse Connectives

We do not consider discourse connectives as prepositions, even those that are prepositional phrases (e.g., *apart from that*, *in other*

[13] Where *to* is governed by a verb, we use the acceptability of an *in order to* paraphrase as a diagnostic for PURPOSE’: *I bought a TV in order to watch the election returns*, but **I vant in order to suck your blood*. Governed by a noun, infinitival *to* can also mark a function for which an entity can serve as an instrument (*a bed to lie on*): these receive the label FUNCTION’, a subtype of PURPOSE’.

[14] In the first example, the complement of *with* is arguably a clause. Likewise with *as* in *Unity is not possible* ***as*** *John is on the throne*. For *with* but not *as*, the clause may be verbless: ***with/*as*** *[John on the throne]*.

The second example is not to suggest that *for* can never head an adjective’s PP complement—in *Abdication would have been unbearable* ***for*** *him*, we consider *for* to be a preposition with the EXPERIENCER’ function. Further discussion of the subtleties of *for* as subordinator can be found on the wiki.

words, of course). Annotators are instructed to annotate such expressions with the ` label.

5.4 Preposition Tags

To characterize preposition meanings/functions in context with broad coverage, we would like an inventory of a manageable number of coarse categories in the manner of nouns and verbs (§4.2). We take Srikumar and Roth's (2013a) inventory (hereafter, **S&R**) as a starting point: as noted in §5.1.4, it clusters fine-grained dictionary senses of 34 English prepositions into 32 labeled classes. Many of the classes resemble semantic roles (e.g., TEMPORAL, LOCATION, AGENT) or spatial relations (PHYSICALSUPPORT, SEPARATION).

We[15] revise and extend **S&R** to improve its descriptive power and deploy it directly as an annotation scheme.[16] The main areas of improvement are highlighted below; final annotation guidelines will be published at a later date.

5.4.1 Broadening

S&R provides a type-level resource: a labeled clustering of dictionary senses for 34 prepositions. Besides improving these sense groupings, we ultimately intend to annotate all preposition tokens in a corpus. Disregarding MWE annotations, the **REVIEWS** corpus contains 87 single-word preposition types over 5,637 tokens.

[15] Improved coarse semantic categories for prepositions are the result of ongoing collaborations; they reflect the efforts of myself and others including Vivek Srikumar, Jena Hwang, Martha Palmer, Tim O'Gorman, Katie Conger, Archna Bhatia, Carlos Ramírez, Yulia Tsvetkov, Michael Mordowanec, Matt Gardner, Spencer Onuffer, and Nora Kazour, as well as helpful conversations with Ed Hovy, Lori Levin, Ken Litkowski, and Orin Hargraves.

[16] Unfortunately, because **S&R** used the original TPP dictionary, we were unable to benefit from Tratz's (2011) sense refinements (§5.1.4).

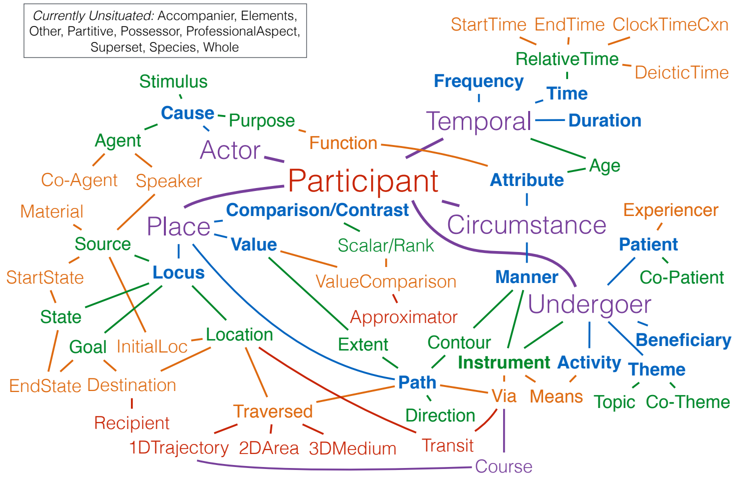


Figure 5.2: The full supersense hierarchy, extending radially outward from PARTICIPANT'. Colors indicate the level of the hierarchy.

5.4.2 Harmonization and Hierarchy

Two other semantic annotation schemes offer similarly sized inventories of roles/relations: VerbNet (Kipper et al., 2008) and AMR (Banarescu et al., 2013). Many of the categories in those schemes overlap (or nearly overlap) with **S&R** labels. Others characterize semantic categories that are absent from **S&R**, but plausibly apply to English prepositions. A comparison of the three inventories is given in table 5.1. The new hierarchy, comprising 70 preposition supersenses, appears in the middle column of the table, and also in figure 5.2.

S&R	Revised	VerbNet / AMR
	Participant	V
	· Actor	V
✓	· · Cause	V A
	· · · Stimulus	V
✓	· · · Purpose	A
	· · · · Function (^Attribute)	≈:meaning
✓	· · · Agent	V
≈Co-Participants	· · · · Co-Agent	V
	· · · · Speaker (^Source)	
	· Undergoer +1	V
✓	· · Activity +1	
✓	· · Beneficiary	V A
	· · Theme	V
≈Co-Participants	· · · Co-Theme	V
✓	· · · Topic	V A
	· · Patient	V
≈Co-Participants	· · · Co-Patient	V
	· · · Experiencer	V
	· Place	V
	· · Locus	
✓	· · · Location	V A
	· · · · InitialLocation (^Source)	V
✓	· · · · Destination (^Goal)	V A
✓	· · · · · Recipient	V ≈:beneficiary
	· · · · Traversed (^Path)	
	· · · · · 1DTrajectory	
Via	· · · · · · Course (^Via)	
	· · · · · 2DArea	
	· · · · · 3DMedium	≈:medium
Via	· · · · Transit (^Via)	

S&R	Revised	VerbNet / AMR
	· · · State +2	
✓	· · · Source +2	V A
	· · · · Material	V
✓	· · · · StartState (^State)	
	· · · Goal +1	V
✓	· · · · EndState (^State)	≈Result
	· · Path +3	:path
✓	· · · Direction	A
	· · · Contour (^Manner)	
≈Numeric	· · Value +1	V /Asset A /:cost
	· · · Extent (^Path)	V A
	· · Comparison/Contrast	:compared-to
	· · · Scalar/Rank	
	· · · · ValueComparison (^Value)	
	· · · · · Approximator	
✓	· Temporal	Time
	· · Frequency	V A
	· · Duration	V A
	· · Age (^Attribute)	A
	· · Time	A
	· · · RelativeTime	
	· · · · StartTime	Initial_time
	· · · · EndTime	Final_time
	· · · · DeicticTime	
	· · · · ClockTimeCxn	
	· Circumstance	V
✓	· · Attribute +2	V ≈:mod/:part
✓	· · Manner +1	V A
✓	· · · Instrument (^Undergoer)	V A
≈Instrument	· · · · Means (^Activity)	

S&R	Revised	VerbNet / AMR
MediumOfCom	││││ Via (ˆPath) +2	≈:medium
≈Partic/Accomp	Accompanier	$\boxed{\text{A}}$
	Elements	:ex/:subset
≈PartWhole	Partitive	≈:consist-of
✓	Possessor	:poss
✓	ProfessionalAspect	:employed-by/:role
✓	Species	
≈PartWhole	Whole	:part-of
	│ Superset	:superset
✓	Other	
Experiencer[17]	—	
ObjOfVerb	—	
Oppon/Contrast	—	
PhysicalSupport	—	
Separation	—	
	—	Pivot
	—	Product

Table 5.1: Harmonization of the **S&R** inventory, **VerbNet** thematic role hierarchy, and **AMR** non-core roles. In the middle column, categories with multiple parents indicate one of them in parentheses, and categories with n children listed under some other parent have a $+n$ designation. In the right column, role names starting with ":" are from AMR and others are from VerbNet. (Some of the above are new in VerbNet, having been added subsequent to the latest published guidelines. Several roles only in AMR are not shown.)

[17] This S&R category has a substantially different meaning from the one in VerbNet and the new scheme.

In designing our supersense label set, we decided to modify **S&R** where possible to be more closely compatible with the other schemes. On a descriptive level, this allows us to take advantage of the linguistic analyses and explanations motivating those categories. On a practical level, this will make it easier to combine resources (lexicons and annotated corpora enriched with semantic role labels).

Following VerbNet, our preposition supersense categories are organized into a hierarchical (multiple inheritance) taxonomy. Not only does this explicate some of the distinctions between related categories that were described textually in **S&R** (e.g., the relationship between StartState› and Source›), but it also provides a practical strategy for annotators who are unsure of how to apply a category—there is often a less specific label to fall back on.

The preposition label set proposed here is noticeably larger than the noun and verb supersenses. This might warrant concern that it will be too difficult for annotators to learn. However, there are arguments in favor of a larger set when it comes to prepositions:

- Because prepositions range from the lexical to the grammatical, they perhaps cover a wider/higher-dimensional semantic space than verbs or nouns. Thus, more categories might be needed for comparable descriptive adequacy.

- The hierarchy should help guide annotators to the right category or small set of related categories. They will not have to consider all of them one by one.

- The presence of more and less abstract categories gives annotators flexibility when they are uncertain.

- Because prepositions are closed-class, we envision that the annotation process will be guided (to a much greater extent than for nouns and verbs) by the word type. Consequences include:

- Having several dozen categories at multiple levels of granularity should mean that the number of prepositions associated with each category is small.
- For TPP prepositions (with fine-grained senses mapped to the new scheme), it will be possible to suggest a filtered list of supersenses to the annotator, and these should suffice for the vast majority of tokens.
- It may even be desirable to annotate a corpus by type rather than by token, so the annotator can focus on a few supersenses at a time.

Based on preliminary rounds of annotation—a mix of type-driven and token-driven—by several annotators, we are optimistic that the general approach will be successful. The preliminary annotation has also uncovered shortcomings in the annotation guidelines that have informed revisions to the categories and hierarchy. More extensive annotation practice with the current scheme is needed to ascertain its adequacy and usability. Should the size of the hierarchy prove too unwieldy, it will be possible to remove some of the finer-grained distinctions.

Below, we examine some of the areas of the hierarchy that are being overhauled.

5.4.3 Temporal Refinement

In **S&R**, all temporal preposition usages fall under a single label, TEMPORAL. VerbNet is slightly more discriminative, with an equivalent TIME supercategory whose daughters are INITIAL_TIME, FINAL_TIME, DURATION, and FREQUENCY.

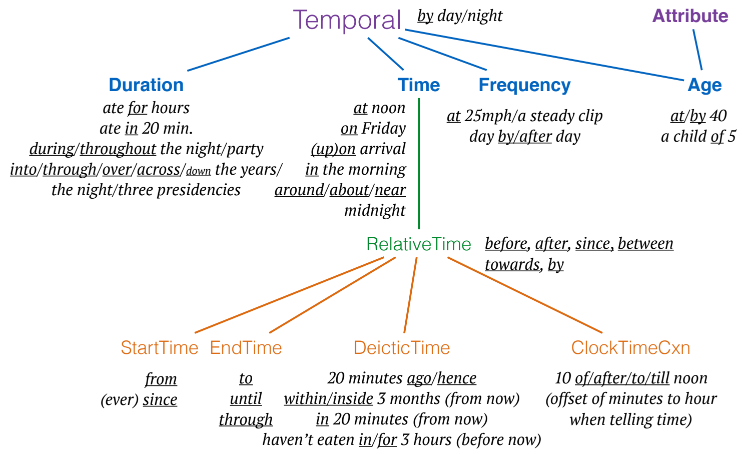


Figure 5.3: The TEMPORAL› subhierarchy, with example preposition usages associated with each supersense.

We have refined this further, as shown in figure 5.3, after coming to the conclusion that the major temporal prepositions cluster neatly into finer-grained subcategories. Relations that situate a time as before or after another time are under RELATIVETIME›; special cases are STARTTIME›, ENDTIME›, times implicitly situated relative to the present (DEICTICTIME›), and constructions for telling time that express an offset in minutes relative to the hour (CLOCKTIMECXN›). We also follow AMR's lead in creating a dedicated AGE› category, which inherits from both TEMPORAL› and ATTRIBUTE›.

Note that most of the prepositions in figure 5.3 are only associated with one or two temporal supersenses; only *in* and *at* are known to occur with three. Therefore we do not expect that the subcategories will impose too much of a burden on annotators.

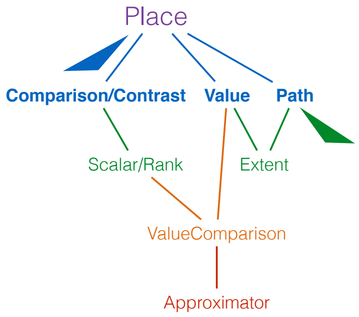


Figure 5.4: Portion of the preposition supersense hierarchy involving COMPARISON/CONTRAST›, VALUE›, and their subtypes. Triangles indicate where other subcategories are omitted.

5.4.4 Values and Comparisons

Many prepositions can be used to express a quantitative value (measuring attributes such as a quantity, distance, or cost), to compare to another value, or to compare to something qualitatively.[18] **S&R** define a broad category called NUMERIC for preposition senses that mark quantitative values and classify some qualitative comparison senses as OTHER. We have developed a finer-grained scheme, seen in figure 5.4. The main categories are COMPARISON/CONTRAST› and VALUE›.

COMPARISON/CONTRAST› applies to qualitative or quantitative analogies, comparisons, and differentiations: e.g., *he used to have a car **like** mine*; *he was screaming **like** a banshee*; *the club's nothing **to** what it once was*; *the benefits must be weighed **against** the costs*; *the difference **between** income and expenditure*; *these fees are quite distinct **from** expenses*. Where these are relative to a specific scale or ranking, the subcategory SCALAR/RANK› is used. Qualitative SCALAR/RANK› examples include: *the firm chose profit **above** car safety*; *a woman who placed duty **before** all else*; *at a level **above** the common people*; *warm weather **for** the time of year*.

VALUE› applies to points on a formal scale—e.g., *prices start **at** $10*; *the drunken yobbos who turned up **by** the cartload*; *my car only does ten miles **to** the gallon*. It also covers prepositions used as mathematical operators: *a map measuring 400 **by** 600 mm*; *she multiplied it **by** 89*; *three **into** twelve goes four*; *ten **to** the minus thirty-three*.

SCALAR/RANK› and VALUE› share a subtype, VALUECOMPARISON›, for comparisons/differentiations on a formal scale—e.g., *the hill was **above/below** sea level*. A special case of this, APPROXIMATOR›, is discussed in some detail below.

VALUE› and PATH› share a subtype, EXTENT›, which is described in §5.4.5.

5.4.4.1 Approximate Values

We propose a new category, APPROXIMATOR›, for cases such as:

(12) a. We have **about** 3 eggs left.
b. We have **around** 3 eggs left.
c. We have **in the vicinity of** 3 eggs left.
d. We have **over** 3 eggs left.
e. We have **between** 3 and 6 eggs left.

Reference sources take several different approaches to such expressions. Dictionaries disagree as to whether these senses of *about* and *around* should be considered prepositions or adverbs. Pullum and Huddleston (2002, p. 646) distinguish the syntactic behavior of *over* in "*She wrote [[over fifty] novels]*" vs. "*I spent [over [a year]] here.*"

[18] Temporal expressions, though sometimes considered values, are here treated separately, as described in the foregoing section.

Whatever the syntactic evidence, semantically these are all similar: they take a measurement, quantity, or range as an argument and "transform" it in some way into a new measurement, quantity, or range. Prepositional expressions *under*, *more than*, *less than*, *greater than*, *fewer than*, *at least*, and *at most* fit into this category as well. Note that these can all be paraphrased with mathematical operators: $\approx < > \leq \geq$.

Approximator› is a subtype of ValueComparison› in the hierarchy. It applies regardless of the semantic type of the thing measured (whether it is a spatial extent, temporal duration, monetary value, ordinal count, etc.). Thus Approximator› also applies to the highlighted prepositions in:

(13) a. It took **about**/**over** 1 year.
b. It took **about**/**over** a year.
c. We swam **about** half a lap. *(no explicit marker of* Extent›*)*
d. We swam for **about** a lap. *(for marks* Extent›*)*
e. I outpaced him by **over** a mile.
f. We ate in **under** 5 minutes.
g. I was there for **over** a year.
h. I heard from **over** a mile away.

We are only annotating preposition expressions, so words that are morphologically more like adverbs—*nearly*, *roughly*, *approximately*—are not included though they may bear the same semantics.

It should be noted that several spatiotemporal prepositions involve a semantics of imprecision. Approximator› is not intended to cover all imprecise preposition senses. As a test, we check which of *near* and *nearly* can be substituted:

(14) a. He lives somewhere **around**/**by**/**near**/*nearly where I used to live.: Location›

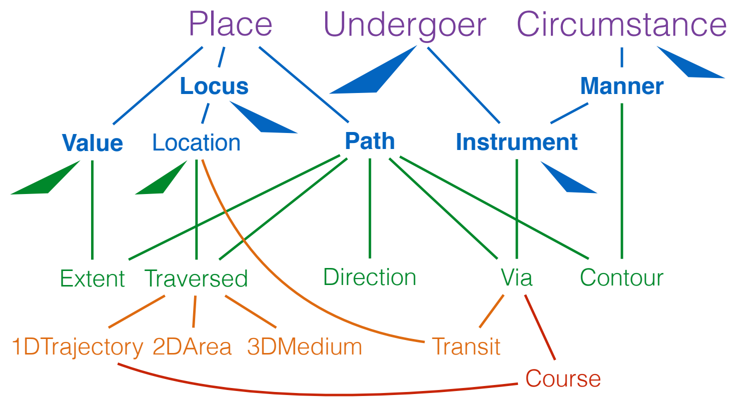


Figure 5.5: Portion of the preposition supersense hierarchy involving Path› and its subtypes. Triangles indicate where other subcategories are omitted.

b. He left **around**/**by**/**near**/*nearly midnight.: Time›
c. He left at **around**/***by**/***near**/nearly midnight.: Approximator›

5.4.5 Paths

Extensive discussion has gone into developing a section of the hierarchy for *paths*, which were not accounted for to our satisfaction in any of the existing schemes. Our analysis draws upon prior and concomitant studies of caused motion constructions in the context of improving their treatment in VerbNet. Those studies address the basic scenarios of Change of Location, Change of State, Transfer of Possession, Transfer of Information, and Change in Value on a Scale with regard to their syntactic and semantic argument structures (Hwang et al., 2014; Hwang, 2014, ch. 5). A proposed subhierarchy for paths—closely related to the approach adopted for VerbNet, but in some respects more detailed—is shown in figure 5.5. Taking PATH› to be the intermediate part of literal or abstract/metaphoric motion,[19] we distinguish the following subtypes:

- **TRAVERSED›:** A stretch of physical space that the figure[20] inhabits during the middle of motion (not necessarily where the event as a whole is located, which be marked with a simple LOCATION› preposition). This category is a subtype of LOCATION› as it describes the “where” of the intermediate phase of motion. It is further refined into:
  - **1DTRAJECTORY›:** A one-dimensional region of space that is traversed, such as by following a path or passing a landmark. Examples: *I walked **along** the river, **over** the bridge, and **past** the castle*
  - **2DAREA›:** The two-dimensional region of space that is “covered”, though there is less of a notion of completeness than with a one-dimensional trajectory. Examples: *I walked **about/through/around** the room*
  - **3DMEDIUM›:** Volumetric material that the figure moves through, and which may exert a facilitatory or opposing force on the figure. Examples: *I waded **through** the swamp*; *the fish swim **with** the current*

  It is expected to be rare that an event will have phrases expressing more than one of these dimension-specific subclasses.

[19] Our notion of path does not include the endpoints, which are captured by INITIALLOCATION› and DESTINATION› in the motion domain, STARTSTATE› and ENDSTATE› for changes of state, and SOURCE› and GOAL› in more abstract domains.

[20] **Figure** is a term for an entity that moves or is spatially situated with respect to another entity, the **ground**. Alternate terms in the literature are **trajector** and **landmark**, respectively. See (Evans and Green, 2006).

- **DIRECTION›:** This covers prepositions marking how the motion of the figure, or the figure itself, is aimed/oriented. This category contrasts with DESTINATION›, where the preposition expressly indicates an intended endpoint of motion. Examples: *walk **toward** the door*, *kick **at** the wall*, *toss the ball **up***, *Step **away from** the cookie jar!*.

- **CONTOUR›:** This describes the shape, but not the location, of a path; it is also a kind of MANNER›. Examples: *walk **in a** zigzag*

- **EXTENT›:** Also a subtype of VALUE›, this is the size of a path: the physical distance traversed or the amount of change on a scale. Examples: *ran **for** miles*, *the price shot up **by** 10%*

- **VIA›:** Prepositions in this category mark something that is used for translocation, transfer, or communication between two points/parties. It is a subtype of PATH› because it pertains to the intermediate phase of (literal or figurative[21]) motion, and also a subtype of INSTRUMENT› because it is something used in order to facilitate that motion. S&R used the label VIA for the spatial domain and MEDIUMOFCOMMUNICATION for communication devices; we instead use the VIA› supersense directly for cases that are *not* physical motion, e.g.: *talk **by** phone*; *talk **on/over** the phone*; *make an appearance **on** TV*; *order **by** credit card **via/on** the Internet*; *I got the word out **via** a friend*. Enablers expressed metaphorically as paths, e.g. *Hackers accessed the system **via** a security hole*, are included as well. There are two subcases:

[21] Communication is systematically framed as transfer of ideas; this is known as the **conduit metaphor** (Reddy, 1979).

– **TRANSIT›:** The vehicle/mode of conveyance that facilitates physical motion traversing a path. It is also a subtype of LOCATION› because it specifies where the figure was during the motion. The category helps distinguish the concrete cases of VIA› from non-concrete cases. Examples: *I went to Paris **by** plane*

– **COURSE›:** The roadway, route, or stopping point on a route that facilitates physical motion traversing a path. It is also a subtype of 1DTRAJECTORY› because it specifies a one-dimensional path for the figure's motion. The category, along with TRANSIT›, helps distinguish the concrete cases of VIA› from non-concrete cases. Examples: *I went to Paris **via** London*; *drive **via** back roads*; *connected **via** Mediterranean shipping routes*; *sent a letter **by** snail mail*

A heuristic for VIA› and its subtypes TRANSIT› and COURSE› is the ability to construct a paraphrase with the word *via*.

5.4.5.1 Fictive Motion

There are spatial usages of certain prepositions that portray static scenes as motion: these fall under the term **fictive motion** (Talmy, 1996). Our conventions are as follows:

- **With a figure whose shape/spatial extent is being described with respect to a landmark:**

  – 1DTRAJECTORY› for the extent of a one-dimensional shape: *a cable runs **above** the duct*; *the bridge [that goes] **across** the river*; *cars were parked **along** the grass verge*; *the tear runs all the way **down** my pants*; *the sun was streaming in **through** the window*, etc.[22]

  – 2DAREA› for the extent of a two-dimensional shape: *She wore her dark hair in plaits **about** her head*

  – INITIALLOCATION› for the "starting point": *There is a lovely road which runs **from** Ixopo into the hills*; *single wires leading **off** the main lines*

  – DESTINATION› for the "ending point": *There is a lovely road which runs from Ixopo **into** the hills*; *every driveway **to** the castle was crowded*

- **For the spatial orientation of a figure:** DIRECTION›: *the gun was aimed **at** his head*; *they faced **away from** each other*

- **Suggesting the spatial path that may be traversed to access a place starting from a reference point (such as the speaker's location):** LOCATION›: *in a little street **off** Whitehall*; *He must have parked **around** the front of the motel*; *the ambush occurred 50 metres **from** a checkpoint*; *they lived **across the street from** one another*; *the auditorium is **through** a set of double doors*; *he lives a dozen miles or so **down** the Thames*; ***over** the hill is a small village*

- **For a physical path of perception (line of sight, hearing, etc.):** 1DTRAJECTORY›: *Lily peeped **around** the open curtain*; *he looked **across** at me*; *glance **over** her shoulder*

- **For a perspective in perception or communication:** LOCATION›: *I can see Russia **from** my house*; *views **over** Hyde Park*; *she rang him **at** home **from** her hotel*

[22] Note that the spatial extent trajectories can be used with verbs like *goes* and *runs*.

5.4.6 Manner and Means

In our supersense hierarchy, we place MANNER› as a parent of INSTRUMENT› (see figure 5.5). We also propose to distinguish MEANS› for prepositions that mark an action that facilitates a goal (S&R include these under INSTRUMENT). We define MEANS› as a subtype of both INSTRUMENT› and ACTIVITY›. Contrast:

(15) a. He broke up the anthill **with** enthusiasm.: MANNER›
b. He broke up the anthill **with** a stick.: INSTRUMENT›
c. He broke up the anthill **by** hitting it with a stick.: MEANS›[23]

(16) a. We coordinated **over** Skype.:[24] VIA›
b. We coordinated **by** setting up a Skype call.: MEANS›

(17) a. The system was compromised by hackers **via** a security hole.: VIA›
b. The system was compromised **through** an exploitation of the security hole by hackers.: MEANS›

(18) a. I drove **in** a zigzag to avoid the traffic.: CONTOUR›
b. I avoided the traffic **by** driving in a zigzag.: MEANS›

In general, the MANNER› category is for prepositions that mark the "how" of an event. For all of the above examples, the PP would be a valid answer to a "how" question (*How did we coordinate? Over Skype. How did you drive? In a zigzag.*).

[23] Note that in our current annotation approach we would not mark prepositions with clausal complements, though we would mark prepositions with eventive NP complements (§5.3). The *with*, of course, marks an INSTRUMENT› as in (15b).

[24] In case anyone from the 20th or 22nd century is reading this, Skype is a service and application for video calling over the Internet.

5.4.7 Communication

Communication is a frequent an important domain in many genres of text, and English systematically invokes language of motion and transfer to describe communication (Reddy, 1979). S&R includes a specific MEDIUMOFCOMMUNICATION category, but its boundaries are not entirely clear. Similarly, AMR incorporates a :MEDIUM role, though this conflates communicative mediums with what we have called 3DMEDIUM› above. In the previous section, we have proposed using VIA› in a way that includes instruments of communication but is slightly more general.

There are also cases where the preposition marks an entity involved in communication, but that entity is not really framed as an intermediary between two parties:

(19) a. I got the scoop **from** a friend/the Internet. *(source of information)*
b. I uploaded the cat picture **to** icanhascheezburger.com. *(abstract destination of abstract information)*
c. I put the file **on** your computer. *(concrete destination of abstract information)*
d. I put it down **on** paper. *(destination of concretely encoded information)*
e. The answer is somewhere **in** this book/room. *(location of concretely encoded information)*
f. The rumor spread **around** the school. *(information metaphorically covering an area metonymically associated with a group of people)*

While it would be potentially useful to know that all of these involve communication, we want to avoid creating a proliferation of communication-specific categories where our current abstract categories—LOCUS›, SOURCE›, GOAL›—would suffice. The same goes for communication with a concrete component, such as writing, where we can use LOCATION›, INITIALLOCATION›, DESTINATION›, etc. Moreover, both nouns and verbs have a COMMUNICATION supersense, which should be enough to identify an associated preposition as functioning in the communication domain. Therefore, we will refrain from including any communication-specific preposition supersenses, though some (such as non-motion VIA›) will be primarily communication-oriented in practice.

A related set of cases involve a language or code of communication:

(20) a. "Shampooing" means "shampoo" **in** French.
b. I am writing a book **in** French.
c. I translated the book **from** English **to** French.

Again, rather than applying a communication-specific role, we can exploit existing categories: ATTRIBUTE›, STARTSTATE›, and ENDSTATE›.

5.4.8 Part-Whole and Set-Element Relations

S&R's PARTWHOLE category is broadly defined to include prepositions whose object is a whole or containing set relative to another entity, as well as for *of* in construction with a partitive, collective, or measure word.[25] We decided that it would be more straightforward to designate three distinct categories: WHOLE›, SUPERSET›, and PARTITIVE›. SUPERSET› is a subtype of WHOLE›.

Part-whole and set-element relations can also occur in reverse order: e.g., *the shower* ***of*** *the bath* vs. *the bath* ***with*** *a shower*. We do not create a special label for parts because the object can usually be interpreted as an attribute, such that it would be difficult to develop

[25]For some of these the first noun can be thought of as "light" or "transparent" in designating a familiar unit of some material; the object of the preposition is not necessarily a "whole" at all—e.g., *a loaf* ***of*** *bread*.

distinguishing criteria for the ATTRIBUTE› category. But we create ELEMENTS› for cases where the object of the preposition exemplifies a set (used with *like*/*such as*/*including*) or notes items excluded from it (*except (for)*/*excluding*).

5.4.9 States

S&R has categories STARTSTATE and ENDSTATE for changes of state, but no label for states in general. We create STATE› as a supertype of STARTSTATE› and ENDSTATE›, which accommodates usages such as ***in*** *love* (moved from **S&R** MANNER), ***on*** *morphine* (moved from OTHER), and ***off*** *work* (moved from SEPARATION). In general, STATE› prepositions can be paraphrased as "in a state of" or "in a state induced by": ***in*** *love* → *in a state of love*, ***on*** *morphine* → *in a state induced by morphine*, etc.

5.4.10 Accompaniment vs. Joint Participation

The preposition *with* is frustratingly promiscuous. It often marks an entity that is associated with a main entity or event; what is frustrating is that the nature of the association seems to lie on a continuum from physical copresence to active counterpart in an event:[26]

(21) a. Tim prefers [tea **with** crumpets].
b. Tim sat **with** his computer.
c. Tim walked **with** Lori.
d. Tim had dinner **with** Lori.
e. Tim talked **to** Lori.
f. Tim talked **with** Lori.

[26]We exclude cases like *Tim walked in with [his hat on]*, where *with* serves as a subordinator of a verbless clause. See §5.3.1.2.

g. Tim argued **with** Lori.
h. Tim fought **with** Lori.
i. Tim fought **against** Lori.
j. Tim fought **against**/#**with** the idea.

S&R provides two relevant categories: Participant/Accompanier and Opponent/Contrast. The former includes cases of physical copresence (as well as attachment, for *onto*); the latter includes several senses of *against* and TPP sense 6(4) of *with*, defined as "in opposition to." But neither S&R nor TPP (on which it is based) provides an obvious home for the (quite frequent) use of *with* as in *talk with Lori*, which implies that Lori is engaged in a conversation (viz.: *#I talked with Lori, but she didn't say anything*).[27]

VerbNet does not provide a role for physical copresence, which would be considered non-core. On the other hand, it has roles Co-Agent, Co-Theme, and Co-Patient for "events with symmetrical participants": Co-Agent is defined as "Agent who is acting in coordination or reciprocally with another agent while participating in the same event" (VerbNet Annotation Guidelines, p. 20), and applies for *talk to/with someone* (the talk-37.5 class) and *fight with/against someone* (meet-36.3-2). However, VerbNet has no entry mapped to the WordNet sense of *fight* where the enemy can be an idea occurring as the direct object or marked with *against* ("*The senator said he would oppose the bill*"—*oppose* is in the same synset as sense 2 of *fight*).

Thus, the S&R's Opponent/Contrast category emphasizes the commonalities between *argue with*, *fight with*, and *fight against*, while ignoring the similarity between *talk with* and *argue with*;

[27] There seems to be a reading of *talk to* where the talking is unidirectional—*I talked to Lori, but she didn't say anything* is acceptable—but the more common case is probably no different from *talk with*. Note that *speak to/with* is similar, but *tell/say to/*with* are strictly unidirectional.

VerbNet instead groups those together under Co-Agent when the second party is a person, but would likely distinguish fighting against a person from fighting against an idea.[28] On balance, something closer to VerbNet's strategy is probably preferable for compatibility with existing parts of the hierarchy.

We therefore propose:

- Co-Agent›, Co-Patient›, and Co-Theme›, following VerbNet, where both participants are engaged in the same event in the same basic capacity, as in (21e–21i);

- Theme› for (21j), where the thing being fought is not fighting back; and

- Accompanier› for (21a–21d), where the two participants are physically colocated or performing the same action in separate (but possibly inferentially related) events. The inclusion of *together* seems more natural for these: *Tim walked/?talked together with Lori.*

[28] This is an interesting case where there are seemingly two dimensions of meaning captured in the prepositions, and the previous schemes encode different ones in their categorizations. VerbNet encodes thematic roles relating the object of the preposition to an event, whereas S&R's Opponent/Contrast is reminiscent of an **image schema** (Johnson, 1987) in the parlance of cognitive linguistics, and also falls within the scope of Talmy's (1988) **force dynamics**. That is, the "opponent" part of Opponent/Contrast can be understood as schematically encoding situations where two forces come into opposition, whatever their roles (agent, cause, theme, ...) in the event framing that opposition. The notion of "attachment" covered by Participant/Accompanier is another example of an image schema. For in-depth analyses of prepositions and spatial and causal categories in cognitive linguistics, see Brugman (1981); Lakoff (1987, pp. 416–461); Tyler and Evans (2003); Lindstromberg (2010); Croft (2012); and the survey in Evans and Green (2006, ch. 10).

5.4.11 Abandoned or Modified S&R Categories

We list the examples from Srikumar and Roth (2013a) for the categories that have been removed or undergone major restructuring in our supersense hierarchy:

Direction has been narrowed due to the creation of other Path› subcategories (§5.4.5).

- driving **along** the road → 1DTrajectory›
- drive **by** the house → 1DTrajectory›
- tears streaming **down** her face → 1DTrajectory›
- wander **down** the road → 1DTrajectory›
- roll **off** the bed → InitialLocation›
- swim **with** the current → 3DMedium›

Experiencer in S&R has a different meaning than the category by the same name in VerbNet. We therefore remove it.

- focus attention **on** her → Goal›
- he blamed it **on** her → Beneficiary›
- he was warm **toward** her → Beneficiary›
- felt angry **towards** him → Stimulus›

Manner has been narrowed due to new categories:

- to be **in** love → State›
- a woman **in** her thirties → Age›
- planets move **in** ellipses around the sun → Contour›
- plummet **like** a dive-bomber → Comparison/Contrast›
- obtained **through** fraudulent means → Means›
- freedom of expression **through** words → Via›

Instrument has been narrowed due to new categories including Means›.

- provide capital **by** borrowing → Means›
- banged his head **on** the beam → Location›
- voice **over** the loudspeaker → Via›
- heard **through** the grapevine → Via›
- fill the bowl **with** water → Theme›

MediumOfCommunication has been removed in favor of more abstract categories such as Via› and Attribute› (§5.4.5).

- say it **in** French → Attribute›
- put your idea down **on** paper → Destination›
- saw the new series **on** TV → Via›

Numeric has been largely renamed to Value›, per VerbNet, and reinterpreted as described in §5.4.4. Some senses are reassigned to Extent› or the new Temporal› subcategories as appropriate.

- driving **at** 50mph → Frequency›
- missed the shot **by** miles → Extent›
- crawled **for** 300 yards → Extent›
- a boy **of** 15 → Age›

ObjectOfVerb was awkwardly defined and has been removed; other categories (or MWEs) should accommodate its senses.

- inquired **after** him → *inquire after* as MWE
- chase **after** something → *chase after* (and *go after*) as MWE, following WordNet
- sipped **at** his coffee → Direction›
- considerations **for** the future → Topic›, following FrameNet

- saved **from** death → ACTIVITY›
- the wedding **of** his daughter → AGENT›
- it was kind **of** you → POSSESSOR›
- she tells **of** her marriage → TOPIC›
- presided **over** the meeting → *preside over* as MWE
- scan **through** document → 2DAREA›
- a grant **towards** the cost → PURPOSE›
- a threat **to** world peace → THEME›
- cross **with** her → STIMULUS›

OPPONENT/CONTRAST is removed in favor of VerbNet-inspired categories CO-AGENT›, THEME›, CO-THEME›, etc.; see §5.4.10.

- fight **against** crime → THEME›
- gave evidence **against** him → BENEFICIARY› *(maleficiary)*
- the match **against** Somerset → CO-AGENT›
- fought **with** another man → CO-AGENT›
- the wars **between** Russia and Poland → AGENT›
- fees are distinct **from** expenses → SEPARATION›
- turned up his collar **against** the wind → DIRECTION›

OTHER is retained for truly miscellaneous senses, such as:

- drinks are **on** me *(responsibility—cf. (7e))*

However, we note that many of the original examples can be relocated to another category or solved by treating the preposition as part of an MWE:

- at a level **above** the people, married **above** her, the director is **over** him, he was rather **beneath** the princess → SCALAR/RANK›
- health comes **after** housing, placed duty **before** everything → SCALAR/RANK›

- heard **above** the din → STIMULUS›
- felt clumsy **beside** [=compared to] her → COMPARISON/CONTRAST›
- a drawing **after** Millet's The Reapers, named her Pauline **after** her mother → COMPARISON/CONTRAST›
- married for **over** a year → APPROXIMATOR›
- he is **on** morphine → STATE›
- he smiled **to** her astonishment → ENDSTATE›
- leave it **with** me → LOCATION›
- swap **for** that → CO-THEME›
- 'F' is **for** fascinating → FUNCTION›
- tall **for** her age → SCALAR/RANK›
- works **like** [=such as] Animal Farm → CONTENTS›
- picked up tips **along** the way → *along the way* as MWE marking PATH›
- swear **by** God → *swear by* as MWE

PARTICIPANT/ACCOMPANIER seemed to conflate attachment, co-presence, and co-participation; the new ACCOMPANIER› category has a narrower meaning (see §5.4.10).

- a nice steak **with** a bottle of red wine → ACCOMPANIER›
- his marriage **with** Emma → ACCOMPANIER›
- he is married **to** Emma → *married to* (and *wedded to*) as MWEs
- he pinned the map **to** the wall → CO-PATIENT›
- a map pinned **to** the wall → LOCATION›
- stick the drawings **onto** a large map → DESTINATION›

CO-PARTICIPANTS has been removed.

- drop in tooth decay **among** children → LOCUS›
- divide his kingdom **among** them → RECIPIENT›
- links **between** science and industry → CO-THEME›

- the difference **between** income and expenditure → COMPARISON/CONTRAST›
- choose **between** two options → COMPARISON/CONTRAST›

PARTWHOLE has been removed in favor of the narrower categories WHOLE›, SUPERSET›, and PARTITIVE› (§5.4.8).

- sleeve **of** the coat → WHOLE›
- see a friend **among** them → SUPERSET›
- a slice **of** cake → PARTITIVE›
- cup **of** soup → PARTITIVE›

PHYSICALSUPPORT has been removed in favor of LOCATION› on the grounds that it is too narrow.

- stood with her back **against** the wall → LOCATION›
- a water jug **on** the table → LOCATION›

SEPARATION has been removed.

- the party was ousted **from** power → STARTSTATE›
- tear the door **off** its hinges → INITIALLOCATION›
- burden **off** my shoulders → STARTSTATE›
- I stay **off** alcohol → STATE›
- part **with** possessions → *part with* as MWE

SOURCE has been narrowed slightly due to INITIALLOCATION›.

- I am **from** Hackeney → INITIALLOCATION›

VIA has been made more abstract; the new subcategories TRANSIT› and COURSE› cover most previous VIA cases (§5.4.5).

- traveling **by** bus → TRANSIT›
- he is **on** his way → COURSE›
- sleep **on** the plane → LOCATION› (the plane does not represent a path of sleeping)
- got **on** the train → DESTINATION›
- go **through** the tube → 1DTRAJECTORY›

5.5 Conclusion

English prepositions are a challenging class, given that there are so many of them and they are put to so many uses. As Orin Hargraves put it to me: “They are without a doubt the most chameleonlike of all parts of speech.” In the interest of uncovering the chameleon’s palette, we have built on prior work to propose a new hierarchical taxonomy of preposition supersenses, so that (like nouns and verbs) their semantics can be modeled in a coarse WSD framework. The taxonomy will hopefully port well to adpositions and case markers in other languages, though we have not investigated that yet. Our annotation scheme is, to our knowledge, the first to engage deeply with multiword expressions, and intends to capture a broader selection of preposition types than the most similar previous approach (Srikumar and Roth, 2013a). Having piloted preposition annotations for sentences in the **REVIEWS** corpus, the next step will be full-fledged annotation.

Part II

Automation

To curry favor, favor curry.

P.D.Q. Bach, *The Seasonings*

Q. What about "yore?"
A. That refers to "the days of yore," when there was a lot of yore lying around, as a result of pigs.

Dave Barry, "Mr. Language Person on nitches, yores and defective sea lions" (Dec. 5, 1999)

CHAPTER 6

Multiword Expression Identification

Ch. 3 introduced a representation, annotation scheme, and comprehensive corpus of MWEs. This chapter:

- Shows how lexical semantic segmentations (allowing for gaps and a strength distinction) can be encoded with word-level tags
- Describes a supervised model of MWE identification
- Introduces an evaluation measure for the MWE identification task
- Analyzes the model's performance on held-out data from our corpus, and on a corpus in another domain
- Measures the impact of features that use external resources (lexicons, clusters)
- Compares against a simpler baseline consisting of heuristic lexicon lookup

6.1 Introduction

Ch. 3 presented our *comprehensive annotation approach* for MWEs: unlike most existing MWE corpora, it neither targets specific varieties of MWEs nor relies upon any preexisting lexical resource. The annotations are *shallow*, not relying explicitly on syntax (though in principle they could be mapped onto the parses in the Web Treebank). In this chapter we use that corpus (version 1.0) to train and evaluate statistical MWE identification models. This reprises work that appeared as Schneider et al. (2014a). Additionally, we conduct an out-of-domain evaluation on the **WIKI50** corpus (Vincze et al., 2011), which was likewise annotated for named entities and several kinds of idiomatic MWEs.

6.2 Evaluation

6.2.1 Matching Criteria

Given that most tokens do not belong to an MWE, to evaluate MWE identification we adopt a precision/recall-based measure similar to one in the coreference resolution literature. The MUC criterion (Vilain et al., 1995) measures precision and recall of links in terms of groups (units) implied by the transitive closure over those links.[1] Our measure can be defined as follows.

Let $a - b$ denote a link (undirected) between two elements in the gold standard, and let $a \hat{-} b$ denote a link in the system prediction. Let the $*$ operator denote the transitive closure over all links, such that $[\![a -^* b]\!]$ is 1 if a and b belong to the same (gold) set, and

[1]As a criterion for coreference resolution, the MUC measure has perceived shortcomings which have prompted several other measures (see Recasens and Hovy, 2011 for a review). It is not clear, however, whether any of these criticisms are relevant to MWE identification.

Precision: The proportion of predicted links whose words both belong to the same expression in the gold standard.
Recall: Same as precision, but swapping the predicted and gold annotations.
Strength Averaging: A weak link is treated as intermediate between a strong link and no link at all: precision, recall, and F_1 computed on strong links only are averaged with the respective calculations computed on all links without regard to strength.

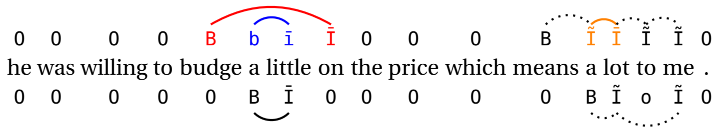


Figure 6.1: A sentence with two hypothetical MWE annotations. Strong links are depicted with solid arcs, and weak links with dotted arcs. Precision of the top annotation relative to the bottom one is 1/3 with weak links removed and 2/6 with weak links strengthened to strong links (note that a link between words w_1 and w_2 is "matched" if, in the other annotation, there is a path between w_1 and w_2). The respective recall values are 1/1 and 3/3. Overall F_1 is computed as the average of two F_1-scores: $\frac{1}{3} \cdot \frac{1}{1} / (\frac{1}{3} + \frac{1}{1}) + \frac{2}{6} \cdot \frac{3}{3} / (\frac{2}{6} + \frac{3}{3}) = 0.50$.

0 otherwise. Assuming there are no redundant[2] links within any annotation (which in our case is guaranteed by linking consecutive words in each MWE), we approximate the MUC precision and recall

[2]A link between a and b is redundant if the other links already imply that a and b belong to the same set. A set of N elements is expressed non-redundantly with exactly $N-1$ links.

measures as:[3]

$$P = \frac{\sum_{a,b:\, a \mathrel{\hat{-}} b} [\![a \mathrel{-^*} b]\!]}{\sum_{a,b:\, a \mathrel{\hat{-}} b} 1} \qquad R = \frac{\sum_{a,b:\, a - b} [\![a \mathrel{\hat{-}^*} b]\!]}{\sum_{a,b:\, a - b} 1}$$

This awards partial credit when predicted and gold expressions overlap in part. Requiring full MWEs to match exactly would arguably be too stringent, overpenalizing larger MWEs for minor disagreements. We combine precision and recall using the standard F_1 measure of their harmonic mean. This is the **link-based** evaluation used for most of our experiments. Figure 6.1 presents a worked example. For comparison, we also report some results with a more stringent **exact match** evaluation where the span of the predicted MWE must be identical to the span of the gold MWE for it to count as correct.

6.2.2 Strength Averaging

Recall that the 2-level scheme (§3.5.1) distinguishes *strong* vs. *weak* links/groups, where the latter category applies to reasonably compositional collocations as well as ambiguous or difficult cases. We want to penalize weak-link-vs.-no-link and weak-link-vs.-strong-link disagreements less than strong-link-vs.-no-link disagreements. To accommodate the 2-level scheme, we therefore average $F_1^{\uparrow}$, in which all weak links have been converted to strong links, and $F_1^{\downarrow}$, in

[3]This is actually a slight simplification of the MUC measure, which is defined directly over clusters. Consider the three-word sequence a, b, c. Suppose the three words constitute an MWE in the prediction, forming a cluster $\{a, b, c\}$, but the gold standard has a gappy expression, and so describes two clusters, $\{a, c\}$ and $\{b\}$. By our measure, precision examines the $a - b$ and $b - c$ links, neither of which is compatible with the gold clustering, so the score is 0. By the MUC measure, the cardinality-3 predicted cluster maps to two gold clusters, so the precision is $(3-2)/(3-1) = 0.5$. Both measures will arrive at 1.0 for recall. Because we are microaveraging across the entire corpus, and because of the relative rarity of gaps, there should be very little difference between the two measures in terms of the overall scores.

he	was	willing	to	budge	a	little	on	the	price	which	means	a	lot	to	me	.	no gaps, 1-level
O	O	O	O	O	B	I	O	O	O	O	B	I	I	I	I	O	

he	was	willing	to	budge	a	little	on	the	price	which	means	a	lot	to	me	.	no gaps, 2-level
O	O	O	O	O	B	Ī	O	O	O	O	B	Ĩ	Ī	Ĩ	Ĩ	O	

he	was	willing	to	budge	a	little	on	the	price	which	means	a	lot	to	me	.	gappy, 1-level
O	O	O	O	B	b	i	I	O	O	O	B	I	I	I	I	O	

he	was	willing	to	budge	a	little	on	the	price	which	means	a	lot	to	me	.	gappy, 2-level
O	O	O	O	B	b	ī	Ī	O	O	O	B	Ĩ	Ī	Ĩ	Ĩ	O	

Figure 6.2: Examples for the 4 tagging schemes. Strong links are depicted with solid arcs, and weak links with dotted arcs. The bottom analysis was provided by an annotator; the ones above are simplifications.

which they have been removed: $F_1 = \frac{1}{2}(F_1^{\uparrow} + F_1^{\downarrow})$.[4] If neither annotation contains any weak links, then $F_1 = F_1^{\uparrow} = F_1^{\downarrow}$. This method applies to both the link-based and exact match evaluation criteria.

6.3 Tagging Schemes

Following (Ramshaw and Marcus, 1995), shallow analysis is often modeled as a sequence-chunking task, with tags containing chunk-positional information. The BIO scheme and variants (e.g., BILOU;

[4]Overall precision and recall are likewise computed by averaging "strengthened" and "weakened" measurements.

	no gaps	gappy
1-level	$(O \mid BI^{+})^{+}$	$(O \mid B(o \mid bi^{+} \mid I)^{*} I^{+})^{+}$
2-level	$(O \mid B[\bar{I}\tilde{I}]^{+})^{+}$	$(O \mid B(o \mid b[\bar{\imath}\tilde{\imath}]^{+} \mid [\bar{I}\tilde{I}])^{*} [\bar{I}\tilde{I}]^{+})^{+}$

Figure 6.3: Regular expressions for the 4 tagging schemes.

Ratinov and Roth, 2009) are standard for tasks like named entity recognition, supersense tagging, and shallow parsing.

The language of derivations licensed by the grammars in §3.5 allows for a tag-based encoding of MWE analyses with only bigram constraints. We describe 4 tagging schemes for MWE identification, starting with BIO and working up to more expressive variants. They are depicted in figure 6.2, and regular expressions defining valid tag sequences appear in figure 6.3.

No gaps, 1-level (3 tags). This is the standard contiguous chunking representation from Ramshaw and Marcus (1995) using the tags {O B I} (introduced in §2.4.2 above). O is for tokens **o**utside any chunk; B marks tokens **b**eginning a chunk; and I marks other tokens **i**nside a chunk. Multiword chunks will thus start with B and then I. B must always be followed by I; I is not allowed at the beginning of the sentence or following O.

No gaps, 2-level (4 tags). We can distinguish strength levels by splitting I into two tags: Ī for strong expressions and Ĩ for weak expressions. To express strong and weak contiguous chunks requires 4 tags: {O B Ī Ĩ}. (Marking B with a strength as well would be redundant because MWEs are never length-one chunks.) The constraints on Ī and Ĩ are the same as the constraints on I in previous schemes. If Ī and Ĩ occur next to each other, the strong attachment will receive higher precedence, resulting in analysis of strong MWEs as nested

within weak MWEs.

Gappy, 1-level (6 tags). Because gaps cannot themselves contain gappy expressions (we do not support full recursivity), a finite number of additional tags are sufficient to encode gappy chunks. We therefore add lowercase tag variants representing tokens *within a gap*: {O o B b I i}. In addition to the constraints stated above, no within-gap tag may occur at the beginning or end of the sentence or immediately following or preceding O. Within a gap, b, i, and o behave like their out-of-gap counterparts.

Gappy, 2-level (8 tags). 8 tags are required to encode the 2-level scheme with gaps: {O o B b Ī ī Ĩ ĩ}. Variants of the inside tag are marked for strength of the incoming link—this applies gap-externally (capitalized tags) and gap-internally (lowercase tags). If Ī or Ĩ immediately follows a gap, its diacritic reflects the strength of the gappy expression, not the gap's contents.

6.4 Model

With the above representations we model MWE identification as sequence tagging, one of the paradigms that has been used previously for identifying *contiguous* MWEs (Constant and Sigogne, 2011, see §6.6).[5] Constraints on legal tag bigrams are sufficient to ensure the full tagging is well-formed subject to the regular expressions in figure 6.2; we enforce these constraints in our experiments.[6] For learning we use the framework of the cost-augmented structured

[5] Hierarchical modeling based on our representations is left to future work.

[6] The 8-tag scheme licenses 42 tag bigrams: sequences such as B O and o ī are prohibited. There are also constraints on the allowed tags at the beginning and end of the sequence.

perceptron, reviewed in §2.4.3 and §2.4.4. Below we detail our cost function, features, and experimental setup.

6.4.1 Cost Function

To better align the learning algorithm with our F-score–based MWE evaluation (§6.2), we use a cost-augmented version of the structured perceptron that is sensitive to different kinds of errors during training (§2.4.4). When recall is the bigger obstacle, we can adopt the following cost function: given a sentence $\mathbf{x}$, its gold tagging $\mathbf{y}^*$, and a candidate tagging $\mathbf{y}'$,

$$cost(\mathbf{y}^*, \mathbf{y}', \mathbf{x}) = \sum_{j=1}^{|\mathbf{y}^*|} c(y_j^*, y_j') \quad \text{where}$$
$$c(y^*, y') = [\![y^* \neq y']\!] + \rho [\![y^* \in \{\mathtt{B}, \mathtt{b}\} \wedge y' \in \{\mathtt{O}, \mathtt{o}\}]\!]$$

A single nonnegative hyperparameter, ρ, controls the tradeoff between recall and accuracy; higher ρ biases the model in favor of recall (possibly hurting accuracy and precision). This is a slight variant of the recall-oriented cost function of Mohit et al. (2012). The difference is that we only penalize *beginning-of-expression* recall errors. Preliminary experiments showed that a cost function penalizing all recall errors—i.e., with $\rho [\![y^* \neq \mathtt{O} \wedge y' = \mathtt{O}]\!]$ as the second term, as in Mohit et al.—tended to append additional tokens to high-confidence MWEs (such as proper names) rather than encourage new MWEs, which would require positing at least two new non-outside tags.

6.4.2 Features

6.4.2.1 Basic Features

These are largely based on the sequence model features of Constant and Sigogne (2011); Constant et al. (2012): they look at word unigrams and bigrams, character prefixes and suffixes, and POS tags, as well as lexicon entries that match lemmas[7] of multiple words in the sentence.[8]

Some of the basic features make use of *lexicons*. We use or construct 10 lists of English MWEs: all multiword entries in **WordNet** (Fellbaum, 1998); all multiword chunks in **SEMCOR** (Miller et al., 1993); all multiword entries in English **Wiktionary**;[9] the **WikiMwe** dataset mined from English Wikipedia (Hartmann et al., 2012); the **SAID** database of phrasal lexical idioms (Kuiper et al., 2003); the named entities and other MWEs in the WSJ corpus on the English side of the **PCEDT** (Hajič et al., 2012); the **verb-particle constructions** (VPCs) dataset of Baldwin (2008); a list of **light verb constructions** (LVCs) provided by Claire Bonial; and two idioms websites.[10] After preprocessing, each lexical entry consists of an ordered sequence of word lemmas, some of which may be variables like <*something*>.

Given a sentence and one or more of the lexicons, lookup for the lexicon features proceeds as follows: we enumerate entries whose lemma sequences match a sequence of lemmatized tokens, and

[7] The WordNet API in NLTK (Bird et al., 2009) was used for lemmatization.

[8] Our MWE system, like the sequence model (but unlike the reranking model) of Constant et al. (2012), does not include any features derived from the output of a syntactic parser, as explained in §8.1.3.

[9] `http://en.wiktionary.org`; data obtained from `https://toolserver.org/~enwikt/definitions/enwikt-defs-20130814-en.tsv.gz`

[10] `http://www.phrases.net/` and `http://home.postech.ac.kr/~oyz/doc/idiom.html`

build a lattice of possible analyses over the sentence. We find the shortest path (i.e., using as few expressions as possible) with dynamic programming, allowing gaps of up to length 2.[11]

In detail, the basic features are:

All are conjoined with the current tag, y_i.

Tag Features

1. previous tag (the only first-order feature)

Token Features

Original token

2. $i = \{1, 2\}$
3. $i = |\mathbf{w}| - \{0, 1\}$
4. capitalized $\wedge$ $[\![i = 0]\!]$
5. word shape

Lowercased token

6. prefix: $[w_i]_1^k \big|_{k=1}^4$
7. suffix: $[w_i]_j^{|w|} \big|_{j=|w|-3}^{|w|}$
8. has digit
9. has non-alphanumeric c
10. context word: $w_j \big|_{j=i-2}^{i+2}$
11. context word bigram: $\mathbf{w}_j^{j+1} \big|_{j=i-2}^{i+1}$

Lemma Features

12. lemma + context lemma if one of them is a verb and the other is a noun, verb, adjective, adverb, preposition, or particle: $\lambda_i \wedge \lambda_j \big|_{j=i-2}^{i+2}$

[11] Each top-level lexical expression (single- or multiword) incurs a cost of 1; each expression within a gap has cost 1.25.

Part-of-speech Features

13. context POS: $pos_j \big|_{j=i-2}^{i+2}$
14. context POS bigram: $\mathbf{pos}_j^{j+1} \big|_{j=i-2}^{i+1}$
15. word + context POS: $w_i \wedge pos_{i\pm1}$
16. context word + POS: $w_{i\pm1} \wedge pos_i$

Lexicon Features (unlexicalized)

WordNet only

17. OOV: λ_i is not in WordNet as a unigram lemma $\wedge$ pos_i
18. compound: non-punctuation lemma λ_i and the {previous, next} lemma in the sentence (if it is non-punctuation; an intervening hyphen is allowed) form an entry in WordNet, possibly separated by a hyphen or space
19. compound-hyphen: pos_i = `HYPH` $\wedge$ previous and next tokens form an entry in WordNet, possibly separated by a hyphen or space
20. ambiguity class: if content word unigram λ_i is in WordNet, the set of POS categories it can belong to; else pos_i if not a content POS $\wedge$ the POS of the longest MW match to which λ_i belongs (if any) $\wedge$ the position in that match (B or I)

For each multiword lexicon

21. lexicon name $\wedge$ status of token i in the shortest path segmentation (O, B, or I) $\wedge$ subcategory of lexical entry whose match includes token i, if matched $\wedge$ whether the match is gappy
22. the above $\wedge$ POS tags of the first and last matched tokens in the expression

Over all multiword lexicons

23. at least k lexicons contain a match that includes this token (if $n \geq 1$ matches, n active features)
24. at least k lexicons contain a match that includes this token, starts with a given POS, and ends with a given POS

6.4.2.2 Unsupervised Word Clusters

Distributional clustering on large (unlabeled) corpora can produce lexical generalizations that are useful for syntactic and semantic analysis tasks (e.g.: Miller et al., 2004; Koo et al., 2008; Turian et al., 2010; Owoputi et al., 2013; Grave et al., 2013). We were interested to see whether a similar pattern would hold for MWE identification, given that MWEs are concerned with what is lexically *idiosyncratic*—i.e., backing off from specific lexemes to word classes may lose the MWE-relevant information. Brown clustering[12] (Brown et al., 1992) on the 21-million-word Yelp Academic Dataset[13] (which is similar in genre to the annotated web reviews data) gives us a hard clustering of word types. To our tagger, we add features mapping the previous, current, and next token to Brown cluster IDs. The feature for the current token conjoins the word lemma with the cluster ID.

6.4.2.3 Part-of-Speech Tags

We compared three PTB-style POS taggers on the full **REVIEWS** subcorpus (**train**+**test**). The Stanford CoreNLP tagger[14] (Toutanova et al., 2003) yields an accuracy of 90.4%. The ARK TweetNLP tagger v. 0.3.2 (Owoputi et al., 2013) achieves 90.1% with the model[15] trained on the Twitter corpus of Ritter et al. (2011), and 94.9% when trained on the **ANSWERS**, **EMAIL**, **NEWSGROUP**, and **WEBLOG** subcorpora of WTB. We use this third configuration to produce automatic POS tags for training and testing our MWE tagger. (A comparison condition in §6.5.3 uses oracle POS tags.)

[12]With Liang's (2005) implementation: `https://github.com/percyliang/brown-cluster`. We obtain 1,000 clusters from words appearing at least 25 times.

[13]`https://www.yelp.com/academic_dataset`

[14]v. 3.2.0, with the `english-bidirectional-distsim` model

[15]`http://www.ark.cs.cmu.edu/TweetNLP/model.ritter_ptb_alldata_fixed.20130723`

6.4.3 Experimental Setup

The corpus of web reviews described in §3.8 is used for training and evaluation. 101 arbitrarily chosen documents (500 sentences, 7,171 words) were held out as a final **test** set. This left 3,312 sentences/48,408 words for training/development (**train**). Feature engineering and hyperparameter tuning were conducted with 8-fold cross-validation on **train**. The 8-tag scheme is used except where otherwise noted.

In learning with the structured perceptron (§2.4.3: algorithm 1), we employ two well-known techniques that can both be viewed as regularization. First, we use the average of parameters over all timesteps of learning. Second, within each cross-validation fold, we determine the number of training iterations (epochs) M by early stopping—that is, after each iteration, we use the model to decode the held-out data, and when that accuracy ceases to improve, use the previous model. The two hyperparameters are the number of iterations and the value of the recall cost hyperparameter (ρ). Both are tuned via cross-validation on **train**; we use the multiple of 50 that maximizes average link-based F_1. The chosen values are shown in table 6.3. Experiments were managed with the `ducttape` tool.[16]

6.5 Results

We experimentally address the following questions to probe and justify our modeling approach.[17]

[16]`https://github.com/jhclark/ducttape/`

[17]But first, if our calculations are correct, it has been approximately 800 pages since the last diversion, and is therefore time for a Strategic Jocular MWE-Themed Footnote (SJMWETF) courtesy of Mister Language Person:

> Q. Please explain the correct usage of "exact same."

6.5.1 Is supervised learning necessary?

Previous MWE identification studies have found benefit to statistical learning over heuristic lexicon lookup (Constant and Sigogne, 2011; Green et al., 2012). Our first experiment tests whether this holds for comprehensive MWE identification: it compares our supervised tagging approach with baselines of heuristic lookup on preexisting lexicons. The baselines construct a lattice for each sentence using the same method as lexicon-based model features (§6.4.2). If multiple lexicons are used, the union of their entries is used to construct the lattice. The resulting segmentation—which does not encode a strength distinction—is evaluated against the gold standard.

Results are shown in tables 6.1 and 6.2. Even with just the labeled training set as input, the supervised approach beats the strongest heuristic baseline (that incorporates in-domain lexicon entries extracted from the training data) by 30 precision points, while achieving comparable recall. For example, the baseline (but not the statistical model) incorrectly predicts an MWE in *places to **eat in** Baltimore* (because *eat in*, meaning 'eat at home,' is listed in WordNet). The supervised approach has learned not to trust WordNet too much due to this sort of ambiguity. Downstream applications that currently use lexicon matching for MWE identification (e.g., Ghoneim and Diab, 2013) likely stand to benefit from our statistical approach.

A. "Exact same" is a corpuscular phrase that should be used only when something is exactly the same as something. It is the opposite (or "antibody") of "a whole nother." EXAMPLE: "This is the exact same restaurant where Alma found weevils in her pie. They gave her a whole nother slice."

(Dave Barry, *Dave Barry Is Not Making This Up*: "Punctuation 'R Easy")

	LOOKUP				SUPERVISED MODEL			
preexising lexicons	$\overline{\text{P}}$	$\overline{\text{R}}$	$\overline{F_1}$	σ	$\overline{\text{P}}$	$\overline{\text{R}}$	$\overline{F_1}$	σ
none					74.39	44.43	55.57	2.19
WN + SemCor (71k)	**46.15**	28.41	35.10	2.44	74.51	45.79	56.64	1.90
6 lexicons (420k)	35.05	46.76	**40.00**	2.88	**76.08**	**52.39**	**61.95**	1.67
10 lexicons (437k)	33.98	**47.29**	39.48	2.88	75.95	51.39	61.17	2.30

Table 6.1: Use of preexisting lexicons for lookup-based vs. statistical segmentation. Supervised learning used only basic features and the structured perceptron, with the 8-tag scheme. Results are with the link-based matching criterion for evaluation.

"6 lexicons" refers to WordNet and SemCor plus SAID, WikiMwe, Phrases.net, and English Wiktionary; "10 lexicons" adds MWEs from CEDT, VNC, LVC, and Oyz. (In these lookup-based configurations, allowing gappy MWEs never helps performance.)

All precision, recall, and F_1 percentages are averaged across 8 folds of cross-validation on **train**; standard deviations are shown for the F_1 score. The highest overall value in each column is bolded. The boxed row indicates the configuration used as the basis for subsequent experiments.

6.5.2 How best to exploit MWE lexicons (type-level information)?

For statistical tagging (right portion of table 6.1), using more *preexisting* (out-of-domain) lexicons generally improves recall; precision also improves a bit.

A lexicon of MWEs occurring in the non-held-out training data *at least twice*[18] (table 6.2, bottom row) is marginally worse (better precision/worse recall) than the best result using only preexisting lexicons.

[18]If we train with access to the full lexicon of *training set* MWEs, the learner credulously overfits to relying on that lexicon—after all, it has perfect coverage of the training data!—which proves fatal for the model at test time.

	$\bar{\text{P}}$	$\bar{\text{R}}$	$\overline{F_1}$	σ
LOOKUP: 2 lexicons + *MWtypes*(train)$_{\geq 1}$, max gap length=1	**46.66**	**47.90**	**47.18**	2.31
SUPERVISED MODEL: 6 lexicons + *MWtypes*(train)$_{\geq 2}$	**76.64**	51.91	61.84	1.65

Table 6.2: Best lookup-based and supervised configurations using an in-domain lexicon. These are cross-validation averages. The in-domain lexicon is derived from MWEs annotated in the training portion of each cross-validation fold at least once (lookup) or twice (model). Results that are superior to analogous configurations without an in-domain lexicon (table 6.1) are bolded. Because the best average F_1 score for the supervised model is slightly lower, we do not use the in-domain lexicon in subsequent experiments.

				LINK-BASED			EXACT		
configuration	M	ρ	$\lvert\boldsymbol{\theta}\rvert$	P	R	F_1	P	R	F_1
base model	5	—	1,765k	69.27	50.49	58.35	60.99	48.27	53.85
+ recall cost	4	150	1,765k	61.09	57.94	59.41	53.09	55.38	54.17
+ clusters	3	100	2,146k	63.98	55.51	59.39	56.34	53.24	54.70
+ oracle POS	4	100	2,145k	66.19	59.35	62.53	58.51	57.00	57.71

Table 6.3: Comparison of supervised models on **test** (using the 8-tag scheme). The base model corresponds to the boxed result in table table 6.1, but here evaluated on **test**. For each configuration, the number of training iterations M and (except for the base model) the recall-oriented hyperparameter ρ were tuned by cross-validation on **train**.

6.5.3 Variations on the base model

We experiment with some of the modeling alternatives discussed in §6.4. Results appear in table 6.3 under both the link-based and exact match evaluation criteria. We note that the exact match scores are (as expected) several points lower.

Recall-oriented cost. The recall-oriented cost adds about 1 link-based F_1 point, sacrificing precision in favor of recall.

Unsupervised word clusters. When combined with the recall-oriented cost, these produce a slight improvement to precision/degradation to recall, improving exact match F_1 but not affecting link-based F_1. Only a few clusters receive high positive weight; one of these consists of *matter*, *joke*, *biggie*, *pun*, *avail*, *clue*, *corkage*, *frills*, *worries*, etc. These words are diverse semantically, but all occur in collocations with *no*, which is what makes the cluster coherent and useful to the MWE model.

Oracle part-of-speech tags. Using human-annotated rather than automatic POS tags improves MWE identification by about 3 F_1 points on **test** (similar differences were observed in development).

6.5.4 What are the highest-weighted features?

An advantage of the linear modeling framework is that we can examine learned feature weights to gain some insight into the model's behavior.

In general, the highest-weighted features are the lexicon matching features and features indicative of proper names (POS tag of proper noun, capitalized word not at the beginning of the sentence, etc.).

Despite the occasional cluster capturing collocational or idiomatic groupings, as described in the previous section, the clusters appear to be mostly useful for identifying words that tend to belong (or not) to proper names. For example, the cluster with *street*, *road*, *freeway*, *highway*, *airport*, etc., as well as words outside of the cluster vocabulary, weigh in favor of an MWE. A cluster with every-

POS pattern	#	examples (lowercased lemmas)
NOUN NOUN	53	*customer service, oil change*
VERB PREP	36	*work with, deal with, yell at*
PROPN PROPN	29	*eagle transmission, comfort zone*
ADJ NOUN	21	*major award, top notch, mental health*
VERB PART	20	*move out, end up, pick up, pass up*
VERB ADV	17	*come back, come in, come by, stay away*
PREP NOUN	12	*on time, in fact, in cash, for instance*
VERB NOUN	10	*take care, make money, give crap*
VERB PRON	10	*thank you, get it*
PREP PREP	8	*out of, due to, out ta, in between*
ADV ADV	6	*no matter, up front, at all, early on*
DET NOUN	6	*a lot, a little, a bit, a deal*
VERB DET NOUN	6	*answer the phone, take a chance*
NOUN PREP	5	*kind of, care for, tip on, answer to*

Table 6.4: Top predicted POS patterns and their frequencies.

day destinations (*neighborhood*, *doctor*, *hotel*, *bank*, *dentist*) prefers non-MWEs, presumably because these words are not typically part of proper names in this corpus. This was from the best model using non-oracle POS tags, so the clusters are perhaps useful in correcting for proper nouns that were mistakenly tagged as common nouns. One caveat, though, is that it is hard to discern the impact of these specific features where others may be capturing essentially the same information.

6.5.5 How heterogeneous are learned MWEs?

On **test**, the final model (with automatic POS tags) predicts 365 MWE instances (31 are gappy; 23 are weak). There are 298 unique MWE types.

Organizing the predicted MWEs by their coarse POS sequence

reveals that the model is not too prejudiced in the kinds of expressions it recognizes: the 298 types fall under 89 unique POS+strength patterns. Table 6.4 shows the 14 POS sequences predicted 5 or more times as strong MWEs. Some of the examples (*major award*, *a deal*, *tip on*) are false positives, but most are correct. Singleton patterns include PROPN VERB (*god forbid*), PREP DET (*at that*), ADJ PRON (*worth it*), and PREP VERB PREP (*to die for*), all of which were matched in at least 2 lexicons.

True positive MWEs mostly consist of (a) named entities, and (b) lexical idioms seen in training and/or listed in one of the lexicons. Occasionally the system correctly guesses an unseen and OOV idiom based on features such as hyphenation (*walk - in*) and capitalization/OOV words (*Chili Relleno*, *BIG MISTAKE*). On **test**, 244 gold MWE types were unseen in training; the system found 93 true positives (where the type was predicted at least once), 109 false positives, and 151 false negatives—an unseen type recall rate of 38%. Removing types that occurred in lexicons leaves 35 true positives, 61 false positives, and 111 false negatives—a unseen and OOV type recall rate of 24%.

6.5.6 What kinds of mismatches occur?

Inspection of the output turns up false positives due to ambiguity (e.g., *Spongy and* ***sweet bread***); false negatives (*top to bottom*); and overlap (***get high*** *quality service*, gold *get* ***high quality*** *service*; ***live up*** *to*, gold ***live up to***). A number of the mismatches turn out to be problems with the gold standard, like *having our water* ***shut off*** (gold *having our* ***water shut off***). This suggests that even noisy automatic taggers might help identify annotation inconsistencies and errors for manual correction.

scheme	$\lvert\mathcal{Y}\rvert$	ρ	$\overline{M}$	$\overline{\lvert\boldsymbol{\theta}\rvert}$	$\bar{\text{P}}$	$\bar{\text{R}}$	$\overline{F_1}$
no gaps, 1-level	3	100	2.1	733k	73.33	55.72	63.20
no gaps, 2-level	4	150	3.3	977k	72.60	59.11	65.09
gappy, 1-level	6	200	1.6	1,466k	66.48	61.26	63.65
gappy, 2-level	8	100	3.5	1,954k	73.27	60.44	66.15

Table 6.5: Training with different tagging schemes. Results are cross-validation averages on **train**. All schemes are evaluated against the full gold standard (8 tags).

6.5.7 Are gappiness and the strength distinction learned in practice?

Three quarters of MWEs are strong and contain no gaps. To see whether our model is actually sensitive to the phenomena of gappiness and strength, we train on data simplified to remove one or both distinctions—as in the first 3 taggings in figure 6.2—and evaluate against the full 8-tag scheme. For the model with the recall cost, clusters, and oracle POS tags, we evaluate each of these simplifications of the training data in table 6.5. The gold standard for evaluation remains the same across all conditions.

If the model was unable to recover gappy expressions or the strong/weak distinction, we would expect it to do no better when trained with the full tagset than with the simplified tagset. However, there is some loss in performance as the tagset for learning is simplified, which suggests that gappiness and strength are being learned to an extent.

6.5.8 How does the model fare out of domain, and on particular MWE classes?

As detailed in §3.3.2, the **Wiki50** corpus similarly annotates sentences of running text for several kinds of MWEs (plus named entities). There are two major differences between **Wiki50** and our corpus. First, the domains are different: **Wiki50** contains Wikipedia articles written in a scholarly style, whereas our corpus is of online reviews written in a conversational style. We observe that the former tends to contain long sentences (table 3.1) and advanced terminology/jargon, whereas the latter contains short sentences and a great number of colloquialisms. Second, **Wiki50** treats a more limited inventory of MWE classes, but explicitly categorizes each annotated instance, which allows us to quantify system *recall* by MWE class.

Testing our REVIEWS-trained model on the full **Wiki50** dataset (the “distilled” version) gives the results in table 6.6. The system’s recall is worst on compounds, light verb constructions, and miscellaneous named entities, likely because these are more frequent in the Wikipedia genre than in web reviews. It is important to note that annotation conventions between the two corpora differed in many subtle ways, so the domain difference does not fully account for the measured differences between the two datasets. A comparison of example predictions vs. the gold standard for selected sentences, figure 6.4, reflects several of the differences in annotation conventions.

6.6 Related Work

In terms of modeling, the use of machine learning classification (Hashimoto and Kawahara, 2008; Shigeto et al., 2013) and specifically BIO sequence tagging (Diab and Bhutada, 2009; Constant and Sigogne, 2011; Constant et al., 2012; Vincze et al., 2013a; Le Roux long gap is not detected

tagger includes titles within NEs

single-word NEs are not predicted

(22) Even though H.C.P._Bell did_a very careful and thorough research
NE:PER LVC
on the Maldivian documents , Prime_Minister _Ibrahim_Nasir _'s
NE:MISC COMPOUND_NOUN NE:PER
intention was to have a book on the ancient script of the Maldives
NE:MISC
written by a Maldivian .
NE:MISC

short gap is detected; the system also includes the preposition

(23) He does , however , have_ an _affair _with Clotho , the youngest
LVC NE:PER
aspect of Fate .
NE:MISC

prepositional verbs, miscellaneous MWE subtypes

(24) A_few months later , he was served_with divorce papers by
COMPOUND_NOUN
his new wife .

phrasal idiom is detected

compound is not detected

system detects nested parts of long name

(25) In 1976 , Mixner began the process of coming_out_of_the_closet ,
NE:PER IDIOM
and soon thereafter was a founding member
COMPOUND_NOUN
of the Municipal_Elections_Committee of Los_Angeles (MECLA) ,
NE:ORG NE:ORG
the nation 's first gay and lesbian Political_Action_Committee .
NE:ORG

spurious annotation?

technical term

(26) The common feature of all these routine screening procedures
COMPOUND_NOUN
is that the primary analysis is for indicator organisms rather than
COMPOUND_NOUN
the pathogens that might cause concern .

system detects nested compound

(27) Edwards picked_on nitric_oxide synthase inhibition
NE:PER VPC COMPOUND_NOUN
which was also a failure .

Figure 6.4: **Wiki50** gold standard annotations (Vincze et al., 2011; shown with underlining and category labels) versus our model's predictions (shown by color and underscores).

	R
MW NEs	**7,822/ 9,554=.82**
PER	2,824/ 2,949=.96
ORG	2,020/ 2,510=.80
LOC	1,242/ 1,332=.93
MISC	1,736/ 2,763=.63
other MWEs	**3,893/ 8,280=.47**
COMPOUND_ADJ	52/ 158=.33
COMPOUND_NOUN	2,775/ 6,436=.43
IDIOM	46/ 64=.72
LVC	294/ 740=.40
VPC	693/ 837=.83
OTHER	33/ 45=.73
overall	**13,050/23,175=.56**

Table 6.6: Tag-level recall of the MWE identification model on the full **Wiki50** corpus. The denominator is the number of tokens with gold tags other than O or o, and the numerator is the number of those tokens that also have a predicted tag other than O or o. (No **Wiki50** data was used for training. Excludes single-word NEs and sentences of ≥100 words.)

et al., 2014) for contextual recognition of MWEs is not new. Lexical semantic classification tasks like named entity recognition (e.g., Ratinov and Roth, 2009), supersense tagging (Ciaramita and Altun, 2006; Paaß and Reichartz, 2009), and index term identification (Newman et al., 2012) also involve chunking of certain MWEs. But our discriminative models, facilitated by the new corpus, *broaden the scope of the MWE identification task* to include many varieties of MWEs at once, including explicit marking of gaps and a strength distinction. By contrast, the aforementioned identification systems have been restricted to contiguous MWEs. Of shallow approaches to gappy MWEs: Blunsom and Baldwin (2006) present a sequence model for HPSG supertagging, and evaluate performance on discontinuous

MWEs, though the sequence model treats the non-adjacent component supertags like other tags—it cannot enforce that they mutually require one another, as we do via the gappy tagging scheme (§3.5.1). Gimpel and Smith's (2011) shallow, gappy language model allows arbitrary token groupings within a sentence, whereas our model imposes projectivity and nesting constraints (§3.5).

There are syntax-based approaches that do seek to identify gappy MWEs. Some MWE-enhanced syntactic parsers (but not others: e.g., Green et al., 2012; Candito and Constant, 2014) allow gaps to be described as constituents (Green et al., 2011) or skipped over by MWE dependency links (Vincze et al., 2013b). Techniques based on lexicon lookup and/or syntactic pattern matching can, in some cases, also match gappy MWEs; heuristic lookup may be followed by a statistical idiomatic vs. literal classification step (e.g., Kim and Baldwin, 2010; Fothergill and Baldwin, 2012) or face a high error rate (e.g., Bejček et al., 2013). Unlike these, our approach does not depend on syntactic parsing (see further discussion in §8.1.3).

Another major thread of research has pursued *unsupervised* discovery of multiword types from raw corpora, such as with statistical association measures (Church et al., 1991; Pecina, 2010; Ramisch et al., 2012; Brooke et al., 2014, *inter alia)*, parallel corpora (Melamed, 1997; Moirón and Tiedemann, 2006; Tsvetkov and Wintner, 2010), or a combination thereof (Tsvetkov and Wintner, 2011; Pichotta and DeNero, 2013; Salehi et al., 2014—the first of these uses almost no supervision, while the other two involve both unsupervised and supervised steps). This may be followed by a lookup-and-classify approach to contextual identification (Ramisch et al., 2010). Though preliminary experiments with our models did not show benefit to incorporating such automatically constructed lexicons, we hope these two perspectives can be brought together in future work.

6.7 Conclusion

This chapter has presented the first supervised model for identifying broadly heterogeneous multiword expressions in English text. Our feature-rich discriminative sequence tagger performs shallow chunking with a novel scheme that allows for MWEs containing gaps, and includes a strength distinction to separate highly idiomatic expressions from collocations. It is trained and evaluated on a corpus of English web reviews that are comprehensively annotated for multiword expressions. Beyond the training data, its features incorporate evidence from external resources—several lexicons as well as unsupervised word clusters; we show experimentally that this statistical approach is far superior to identifying MWEs by heuristic lexicon lookup alone. In the next chapter, ch. 7, we enhance the lexical representation with semantic tags. Future extensions might integrate additional features (e.g., exploiting statistical association measures computed over large corpora), improve the expressiveness of the model (e.g., with higher-order features and inference), or integrate the model with other tasks (such as parsing and translation).

Four days later saw me standing at the gates of Castle Dracula, weary and travel-stained. Prudence had demanded that I leave her behind, so I was alone. Night was just falling as I knocked... and Count Dracula's manservant stood before me. Of all the hideously disfigured spectacles I have ever beheld, those perched on the end of this man's nose remain forever pasted into the album of my memory.

Stephen Fry, "The Letter", in *Cambridge Footlights Revue: The Cellar Tapes* (1982)

CHAPTER 7

Full Supersense Tagging

This chapter:

- Shows that the integrated lexical semantic representation set forth in Part I can be mapped to a tagging-chunking representation
- Trains a statistical model that subsumes the traditional supersense tagging task, but with a broader view of multiword expressions
- Evaluates the impact of features that generalize beyond individual word types
- Examines the model's ability to cope with supersense-ambiguous nouns and verbs

Having annotated English sentences with lexical semantic analyses consisting of a segmentation component (multiword expressions) and a categorization component (supersense labels), we now turn to automating this task in a single statistical model.

7.1 Background: English Supersense Tagging with a Discriminative Sequence Model

The model of Ciaramita and Altun (2006) represents the state of the art for full[1] English supersense tagging on the standard SemCor test set, achieving an F_1 score of 77%. It is a feature-based discriminative **tagging-chunking** sequence model learned in a supervised fashion with the structured perceptron, as described in §2.4.3, much like the model deployed in ch. 6 for multiword expressions.

For Ciaramita and Altun (2006) and hereafter, sequences correspond to sentences, with each sentence pre-segmented into words according to some tokenization.[2] Figure 7.1 shows how token-level tags combine BIO flags with supersense class labels to represent the segmentation and supersense labeling of a sentence. These tags are observed during training, predicted at test time, and compared against the gold standard tagging of the test data.

Ciaramita and Altun's (2006) model uses a simple feature set capturing the lemmas, word shapes, and parts of speech of tokens in a small context window, as well as the supersense category of

[1] Paaß and Reichartz (2009) train a similar sequence model for classifying noun and verb supersenses, but treat multiword phrases as single words. Their model is trained as a CRF rather than a structured perceptron, and adds LDA word cluster features, but the effects of these two changes are not separated in the experiments. They also find benefit from constraining the label space according to WordNet for in-vocabulary words (with what they call "lumped labels").

[2] Any ordering or grouping of sentences (e.g., into documents) is disregarded by our models.

United	States	financier	and	philanthropist	(	1855	-	1937	)
$_B$LOC⌢	$_I$LOC⌢	$_B$PERSON⌢	O	$_B$PERSON⌢	O	$_B$TIME⌢	O	$_B$TIME⌢	O

Figure 7.1: A supersense tagging shown with per-token BIO tags in the style of Ciaramita and Altun (2006).

the first WordNet sense of the current word. (WordNet senses are ordered roughly by frequency.) On SemCor data, the model achieves a 10% absolute improvement in F_1 over the first sense baseline.

Though our focus in this chapter is on English, supersense tagging has also been explored in Italian (Picca et al., 2008, 2009; Attardi et al., 2010, 2013; Rossi et al., 2013), Chinese (Qiu et al., 2011), and Arabic (Schneider et al., 2013).

7.2 Piggybacking off of the MWE Tagger

I hope I will not cause any fragile readers to fall off their chairs in surprise by announcing that the methodologies employed and resources created in the foregoing chapters shall be brought to bear on the supersense tagging problem, now somewhat enhanced thanks to a broader and more sophisticated treatment of multiword expressions. The corpus developed in ch. 3 and used to build a lexical semantic segmenter in ch. 6 has since been enriched with semantic class labels for nouns and verbs (ch. 4) to the point that we can build a lexical semantic analyzer in the manner of Ciaramita and Altun (2006). This analyzer has the advantage of being able to represent a more comprehensive assortment of MWEs, including those with gaps, and unlike the classical supersense tagging task is not limited to noun and verb MWEs (though for now, those are the only ones that receive a semantic category label). Despite the fact that many of the annotated expressions in existing supersense datasets contain multiple words, the relationship between MWEs and supersenses has not received much attention (though Piao et al. (2003, 2005) investigated MWEs in the context of a lexical tagger employing a finer-grained taxonomy of semantic classes).

7.3 Experiments: MWEs + Noun and Verb Supersenses

The **STREUSLE** 2.0 dataset, as described in §4.4, is annotated for multiword expressions as well as noun and verb supersenses and auxiliary verbs. We use this dataset for training and testing an integrated lexical semantic analyzer. The experimental setup mostly follows that of ch. 6, which used the **CMWE** 1.0 dataset—i.e., the same **REVIEWS** sentences, but annotated only for MWEs.[3] For simplicity, we use oracle POS tags and learn without the recall-oriented cost function.

7.3.1 Tagset

In the **STREUSLE** dataset, supersense labels apply to *strong* noun and verb expressions—i.e., singleton nouns/verbs as well as strong nominal/verbal MWEs. Weak MWEs are present in the dataset, but not as a unit labeled with a supersense. To convert to token-level tags, we use the 8-way scheme from §6.3 for positional flags to mark the lexical segmentation, and decorate beginners of strong lexical expressions—everything but Ī and ī—with supersense labels. This is illustrated in figure 7.2. Under this formulation, bigram constraints

[3] Here we use the same splits (**train**/**test**, and 8 cross-validation folds within **test** for tuning the number of training iterations M). A handful of the MWE analyses changed between versions of the data.

The white pages allowed me to get in touch with
$_{\text{B}}$COMMUNICATION⌢ Ī $_{\text{O}}$COGNITION˅ O O $_{\text{B}}$SOCIAL˅ Ĩ Ī Ĩ
parents of my high school friends so that I could
$_{\text{O}}$PERSON⌢ O O $_{\text{B}}$GROUP⌢ Ī $_{\text{O}}$PERSON⌢ O O O O
track people down one by one
$_{\text{B}}$SOCIAL˅ $_{\text{o}}$PERSON⌢ Ī B Ī Ī

Figure 7.2: Tagging for part of the lexical semantic analysis depicted in figure 4.5. Note that for nominal and verbal MWEs, the supersense label is only attached to the first tag of the expression.

are sufficient to ensure a globally consistent tagging of the sentence.[4]

Recall from ch. 4 that there are $|\mathcal{N}| = 26$ noun supersense classes[5] and $|\mathcal{V}| = 16$ verb classes (including the auxiliary verb class, abbreviated `a). In principle, then, there are

$$\underbrace{|\{\mathtt{O}\ \mathtt{o}\ \mathtt{B}\ \mathtt{b}\ \tilde{\mathtt{I}}\ \tilde{\mathtt{i}}\}|}_{6} \times \underbrace{(1 + |\mathcal{N}| + |\mathcal{V}|)}_{43} + \underbrace{|\{\bar{\mathtt{I}}\ \bar{\mathtt{i}}\}|}_{2} = 260$$

possible tags encoding chunk and class information, allowing for chunks with no class because they are neither nominal nor verbal expressions. In practice, though, many of these combinations are nonexistent in our data; for experiments we only consider tags occurring in **train**, yielding $|\mathcal{Y}| = 146$.

For comparison, we also run a condition where the substantive supersenses are collapsed to a coarse POS category—i.e., $\mathcal{N}$

[4] Unlike prior work, we do not include the class in strong continuation tags though the class label should be interpreted as extending across the entire expression. This is for a technical reason: as our scheme allows for gaps, the classes of the tags flanking a gap in a strong MWE would be required to match for the analysis to be consistent. To enforce this in a bigram tagger, the within-gap tags would have to encode the gappy expression's class as well as their own, leading to an undesirable blowup in the size of the state space.

[5] including OTHER⌢

in the above formula is replaced with $\{\text{NOUN}\}$ and $\mathcal{V}$ is replaced with $\{\text{VERB}, \grave{}\texttt{a}\}$, yielding 26 tags in principle of which 22 are seen in training;[6] and a condition where the supersense refinements are collapsed entirely, i.e. $\mathcal{Y}$ consists of the 8 MWE tags.

7.3.2 Features

We constrast three feature sets for full supersense tagging: (a) the basic MWE features (§6.4.2.1); (b) the basic MWE features plus Brown clusters (§6.4.2.2); and (c) the basic MWE features, Brown clusters, plus several new features shown below. Chiefly, these new features consult the supersenses of WordNet synsets associated with words in the sentence; there is also a feature aimed at distinguishing auxiliary verbs from main verbs, and new capitalization features take into account the capitalization of the first word in the sentence and the majority of words in the sentence. As with the MWE-only model, we refrain from including any features that depend on a syntactic parser (see §8.1.3 for an explanation).

These are in addition to the basic MWE features (§6.4.2.1). They are conjoined with the current tag, y_i.

New Capitalization Features

25. capitalized $\wedge$ $[\![i = 0]\!]$ $\wedge$ $[\![$majority of tokens in the sentence are capitalized$]\!]$
26. capitalized $\wedge\ i > 0 \wedge w_0$ is lowercase

Auxiliary Verb vs. Main Verb Feature

27. pos_i is a verb $\wedge$ $[\![pos_{i+1}$ is a verb $\vee$ $(pos_{i+1}$ is an adverb $\wedge$ pos_{i+2} is a verb$)]\!]$

^{6}The 4 unattested tags in this condition are $\tilde{\texttt{I}}$, $_{\texttt{b}}$VERB, $_{\texttt{b}}\grave{}\texttt{a}$, and $_{\tilde{\texttt{I}}}\grave{}\texttt{a}$.

WordNet Supersense Features (unlexicalized)

Let $cpos_i$ denote the coarse part-of-speech of token i: common noun, proper noun, pronoun, verb, adjective, adverb, etc. This feature aims primarily to inform the supersense label on the first token of nominal compounds and light verb constructions, where the "semantic head" is usually a common noun subsequent to the beginning of the expression:

28. subsequent noun's 1st supersense: where $cpos_i$ is a common noun, verb, or adjective, $cpos_i$ $\wedge$ for the smallest $k > i$ such that pos_k is a common noun, the supersense of the first WordNet synset for lemma λ_k—provided there is no intervening verb (j such that $cpos_j$ is a verb and $i < j < k$)

The following two feature templates depend on the tag y_i. Let $flag(y_i)$ denote the positional flag part of the tag (O, B, etc.) and $sst(y_i)$ denote the supersense class label:

29. 1st supersense:
    - if $flag(y_i) \in \{\texttt{O}, \texttt{o}\}$: the supersense of the first WordNet synset for lemma λ_i
    - else if $cpos_i$ is a verb and there is a subsequent verb particle at position $k > i$ with no intervening verb: the supersense of the first synset for the compound lemma $\langle \lambda_i, \lambda_k \rangle$ (provided that the particle verb is found in WordNet)
    - otherwise: the supersense of the first WordNet synset for the longest contiguous lemma starting at position i that is present in WordNet: $\langle \lambda_i, \lambda_{i+1}, \ldots, \lambda_j \rangle$ $(j \geq i)$
30. has supersense: same cases as the above, but instead of encoding the highest-ranking synset's supersense, encodes whether $sst(y_i)$ is represented in any of the matched synsets for the given lemma

Most of the model's percepts (binary or real-valued functions of the input[7]) can be conjoined with any tag $y \in \mathcal{Y}$ to form a feature

^{7}We use the term **percept** rather than "feature" here to emphasize that we are talking about functions of the input only, rather than input–output combinations that each receive a weight during learning.

that receives its own weight (parameter). To avoid having to learn a model with tens of millions of features, we impose a percept cutoff during learning: only those zero-order percepts that are active at least 5 times in the training data (with any tag) are retained in the model (with features for all tags). There is no minimum threshold for first-order percepts.[8] The resulting models are of a manageable size: 3–4 million parameters.

7.3.3 Results

Table 7.1 shows full supersense tagging results, separating the MWE identification performance (measured by link-based precision, recall, and F_1; see §6.2) from the precision, recall, and F_1 of class labels on the first token of each expression[9] (segments with no class label are ignored). Exact tagging accuracy is also shown—this number is higher because it gives credit for true negatives, i.e. single-word segments with no nominal or verbal class label (the `O` and `o` tags).

The sequence tagging framework makes it simple to model MWE identification jointly with supersense tagging: this is accomplished by packing information about both kinds of output into the tags. But there is a risk that the larger tag space would impair the model's ability to generalize. By comparing the top and bottom sections of the results, we can see that jointly modeling supersenses along with multiword expressions results in only a minor decrease in MWE identification performance. Thus, we conclude that it is empirically reasonable to model these lexical semantic phenomena together.

[8] Zero-order percepts are those percepts which are to be conjoined with only the present tag to form zero-order features. First-order percepts are to be conjoined with the present and previous tags.

[9] We count the class label only once for MWEs—otherwise this measure would be strongly dependent on segmentation performance. However, the MWE predictions do have an effect when the prediction and gold standard disagree on which token begins a strong nominal or verbal expression.

			MWE ID			Class labeling			Tag
Feature Set	$\|\mathcal{Y}\|$	$\|\boldsymbol{\theta}\|$	P	R	F_1	P	R	F_1	Acc
MWE	8	1,937k	72.97	**55.55**	**63.01**	—	—	—	—
MWE	22	5,330k	**73.26**	54.85	62.68	—	—	—	—
MWE	146	3,555k	67.77	55.76	61.14	64.68	66.78	65.71	80.73
MWE+Brown	146	4,371k	68.55	**56.73**	62.04	65.69	67.76	66.71	81.20
MWE+Brown+SST	146	4,388k	**71.05**	56.24	**62.74**	**69.47**	**71.90**	**70.67**	**82.49**

Table 7.1: Results on **test** for lexical semantic analysis of noun and verb supersenses and MWEs. All of these results use a percept cutoff of 5 and no recall-oriented cost. The first two result rows use a collapsed tagset (just the MWE status, or MWE status conjoined with coarse POS) rather than predicting full supersense labels, as described in §7.3.1. The best result in each column and section is bolded.

Comparing the bottom three rows in the table suggests that features that generalize beyond lexical items lead to better supersense labeling. The best model has access to supersense information in the WordNet lexicon; it is 3 F_1 points better at choosing the correct class label than its nearest competitor, which relies on word clusters to abstract away from individual lexical items.

To better understand the model's behavior, it behooves us to inspect its learned parameters.[10] The highest-weighted parameters suggest that the best model relies heavily on the supersense lookup features (table 7.2), whereas the second-best model—lacking the supersense lookup features—in large part relies on Brown clusters (cf. Grave et al., 2013). The auxiliary verb vs. main verb feature in

[10] Incidentally, this sentence provides an alternative solution to a challenge once posed to Mister Language Person (*Q. Like most people, I would like to use the words "parameters" and "behoove" in the same sentence, but I am not sure how. A. According to the Oxford English Cambridge Dictionary Of Big Words, the proper usage is: "Darlene, it frankly does not behoove a woman of your parameters to wear them stretch pants."* Dave Barry, "Mister Language Person Is Ready To Take Your Calls", Jan. 15, 1996).

y	**Feature Name**	**Weight**
$_{\text{O}}$FOOD^	WN_has_supersense(y)	37.1
Ī	pos$_{-1}^{+0}$: ⟨NNP, NNP⟩	35.2
$_{\text{O}}$`a	auxverb	31.3
$_{\text{O}}$COMMUNICATIONˇ	WN_1st_supersense: COMMUNICATIONˇ	30.3
$_{\text{O}}$PERSON^	WN_1st_supersense: PERSON^	29.0
O	suffix$_4$: `hing`	26.4
$_{\text{O}}$MOTIONˇ	WN_1st_supersense: MOTIONˇ	25.7
$_{\text{O}}$TIME^	WN_has_supersense(y)	24.6
$_{\text{O}}$STATIVEˇ	mainverb	24.2
$_{\text{O}}$GROUP^	WN_1st_supersense: GROUP^	23.6
$_{\text{O}}$EMOTIONˇ	WN_1st_supersense: EMOTIONˇ	23.3
$_{\text{O}}$ARTIFACT^	WN_has_supersense(y)	23.2
Ī	`I` position of non-gappy Wiktionary match	22.5
$_{\text{O}}$PERCEPTIONˇ	WN_1st_supersense: PERCEPTIONˇ	22.0
$_{\text{O}}$POSSESSION^	WN_has_supersense(y)	20.5
O	pos$_{+0}$: IN	20.4
$_{\text{O}}$ARTIFACT^	WN_1st_supersense: ARTIFACT^	20.4
$_{\text{O}}$ACT^	WN_1st_supersense: ACT^	20.3
$_{\text{O}}$COGNITIONˇ	WN_1st_supersense: COGNITIONˇ	19.5
$_{\text{O}}$POSSESSIONˇ	WN_1st_supersense: POSSESSIONˇ	19.5

Table 7.2: Highest positively-weighted features in the best supersense tagging model.

the best model is highly weighted as well, helping to distinguish between `a and STATIVEˇ. Table 7.3 shows the top-weighted features that pertain to the nominal and verbal *communication* categories: we see a mixture of cues in these features, including known WordNet supersenses associated with the current word, the noun subsequent to a verb (linking the verbal and nominal varieties of communication), character prefixes and suffixes, word clusters, and matches against the Wiktionary-derived MWE lexicon.

We have motivated the task of supersense tagging in part as a

y	**Feature Name**
$_{\text{O}}$COMMUNICATIONˇ	WN_1st_supersense: COMMUNICATIONˇ
$_{\text{O}}$COMMUNICATION^	WN_1st_supersense: COMMUNICATION^
$_{\text{O}}$COMMUNICATION^	WN_has_supersense(y)
$_{\text{B}}$COMMUNICATIONˇ	cpos: V, WN_next_N_1st_supersense: COMMUNICATION^
$_{\text{B}}$COMMUNICATIONˇ	WN_has_supersense(y)
$_{\text{O}}$COMMUNICATIONˇ	WN_has_supersense(y)
$_{\text{B}}$COMMUNICATIONˇ	WN_1st_supersense, COMMUNICATIONˇ
$_{\text{O}}$POSSESSION^	WN_1st_supersense, COMMUNICATION^
$_{\text{O}}$COMMUNICATIONˇ	cpos: V, WN_next_N_1st_supersense: COMMUNICATION^
$_{\tilde{\text{I}}}$COMMUNICATION^	WN_1st_supersense: COMMUNICATION^
$_{\text{O}}$COMMUNICATIONˇ	mainverb
$_{\text{O}}$ARTIFACT^	WN_1st_supersense: COMMUNICATION^
Ī	cpos: J, WN_next_N_1st_supersense: COMMUNICATION^
$_{\text{O}}$COMMUNICATION^	suffix$_2$: `te`
$_{\text{O}}$COMMUNICATION^	prefix$_3$: `com`
$_{\text{O}}$COMMUNICATIONˇ	prefix$_4$: `read`
$_{\text{B}}$COMMUNICATIONˇ	`B` position of gappy Wiktionary match
$_{\text{O}}$COMMUNICATIONˇ	cluster$_{+0}$: `1101101000` (‘replied’, ‘said’, ‘claimed’, …)
$_{\text{B}}$COMMUNICATIONˇ	cpos: V, WN_next_N_1st_supersense: COGNITION^
$_{\text{O}}$COMMUNICATIONˇ	prefix$_4$: `call`

Table 7.3: Highest positively-weighted features involving COMMUNICATION^ or COMMUNICATIONˇ. Weights are omitted, but range from 30.3 to 7.7.

coarse form of word sense disambiguation. Therefore, it is worth investigating the extent to which the learned model in fact succeeds at choosing the correct supersense for nouns and verbs that are ambiguous in the data. A handful of lemmas in **test** have at least two different supersenses predicted several times; an examination of four such lemmas in table 7.4 shows that for three of them the tagging accuracy exceeds the majority baseline. In the case of *look*, the model is clearly able to distinguish between COGNITIONˇ (as in

lemma	gold supersense distribution	uniq. gold SSTs	uniq. pred. SSTs	majority baseline	accuracy
get	STATIVEˇ 12, SOCIALˇ 5, COGNITIONˇ 3, POSSESSIONˇ 3, BODYˇ 2, MOTIONˇ 2, COMMUNICATIONˇ 1	7	8	**12/28**	6/28
look	PERCEPTIONˇ 8, COGNITIONˇ 5	2	3	8/13	**12/13**
take	SOCIALˇ 8, MOTIONˇ 7, POSSESSIONˇ 1, STATIVEˇ 4, EMOTIONˇ 1	5	5	8/21	**11/21**
time(s)	TIME^ 8, EVENT^ 5, COGNITION^ 1	3	2	8/14	**9/14**

Table 7.4: Four lemmas and counts of their gold vs. predicted supersenses in **test** (limited to cases where both the gold standard tag and the predicted tag included a supersense).

looking for a company with decent rates) and PERCEPTIONˇ (as in *sometimes the broccoli looks browned around the edges*).

7.4 Conclusion

We have integrated the lexical semantic segmentation task formulated in ch. 6 with the supersense tagging task of Ciaramita and Altun (2006), and applied the annotated English dataset of §4.4 to learn and evaluate a discriminative lexical semantic analyzer. Aside from experimenting with new features, richer models, and indirect forms of supervision (cf. Grave et al., 2013; Johannsen et al., 2014) for this task, the time will soon be ripe for broadening it to include preposition supersenses (ch. 5). Once the preposition supersense

annotations are complete for the corpus, retraining the model described in this chapter should provide a strong baseline for future studies of coarse lexical semantics in context.

PART III

Wrapping Up

Well, that covers a lot of ground. Say, you cover a lot of ground yourself.

Firefly in *Duck Soup*

Words are, in my not-so-humble opinion, our most inexhaustible source of magic.

Dumbledore in *Harry Potter and the Deathly Hallows: Part 2*

CHAPTER 8

Conclusion

8.1 Lessons, Limitations, and Possible Extensions

8.1.1 Summary of Contributions

This thesis has provided a framework for describing the lexical units and semantic classes within text sentences, manually and automatically, with broad coverage. Because the general framework does not depend on any preexisting lexical resource, it is expected to be suitable for a wide range of text domains and languages. The foregoing chapters have motivated and detailed approaches to the *representation* of lexical semantics, a practical approach to human *annotation* of corpora, and statistical techniques for the *automation* of the analysis using said corpora. The primary case study concerning sentences from English web reviews allowed for each of these steps to be understood and documented qualitatively and quantitatively. It has also produced an annotated corpus resource and analysis software, both of which will be released to facilitate further linguistic investigation, computational modeling, and application to other tasks.

The main specific methodological contributions are:

- a shallow but *comprehensive* approach to analyzing heterogeneous **multiword expressions**, including those containing gaps—manually through linguistic annotation (ch. 3) and automatically through a discriminative sequence model with a modified chunking scheme (ch. 6);

- an approach to labeling semantic classes of noun and verb expressions using the **WordNet supersenses**, in a way that builds upon the MWE analysis and still lends itself to automatic sequence tagging (ch. 4, ch. 7); and

- an approach to describing the semantic functions of *prepositions* via a well-documented hierarchical taxonomy of **preposition supersenses** (ch. 5).

The operational details of this framework having been described in the aforementioned chapters, the following sections will elaborate on some of the broader issues raised by the thesis.

8.1.2 Limitations and Difficult Cases

The framework put forward here can be thought of as a compromise between the desire for explicit representations of meaning in context and the desire for practical and rapid corpus annotation with broad coverage. That the approach advocates a shallow treatment of multiword expressions and a coarse treatment of sense disambiguation

should not be interpreted as an argument that finer distinctions and details are irrelevant.[1]

As a reminder, here are some of the paricular difficulties encountered during annotation that may reflect limits on our representation's expressive power:

- There is no way to mark MWEs with overlapping words (§3.6).
- There is no way to mark that an MWE requires a possessive or reflexive constituent that might be a pronoun (§3.6).
- It can be extremely difficult to decide whether a preposition is selected by its governor, forming (e.g.) a prepositional verb (§3.7).[2]
- There is no way to mark constructions with only one lexicalized element (§3.6: footnote 27).
- For nouns referring to complex concepts such as businesses with a physical premises and staff, it is often difficult (and possibly misleading) to choose between ARTIFACT^, GROUP^, and LOCATION^ supersenses (§4.4.1).
- Prepositions can be viewed as having several "facets" of meaning, some of which are orthogonal. This complicates any ap-

[1] As an extreme example of a subtle nuance of (prepositional) meaning, Fillmore (1985) considers the circumstances under which someone can be described as being *in* a bus vs. *on* a bus: excluding the reading of *on* as 'atop', the same spatial configuration is implicated—but *on* is felicitous only if the bus is in service, i.e., the individual is inside the bus on the occasion of a scheduled trip for transporting passengers. Children playing in an abandoned bus are better described as *in* it (Fillmore, 1985, p. 235).

[2] We had hoped that semantic categorization of preposition functions would suggest a solution, namely that prepositions with anomalous functions would count as selected by their head. But this hope was not entirely borne out; it seems that many preposition functions are associated with clusters of semantically similar verbs (e.g., *look at*, *gaze at*, *glance at*, *take a gander at*, etc.). To decide which prepositions are selected may require attention to the equally hairy issue of arguments vs. adjuncts (Hwang, 2011).

proach based on assigning a single category label, even with a hierarchy over those labels.[3]

8.1.3 Regarding MWEs and Syntax

Syntax plays only a minor role in our MWE annotation scheme (ch. 3) and identification system (ch. 6). We use part-of-speech tags to detect candidate annotation targets and in features for the identification tool. But no step of the process relies on syntactic parses (even though gold phrase structure trees are available for the **REVIEWS** sentences).

There were several reasons behind this design decision. First, we see MWEs as primarily a phenomenon of lexical semantics, so we did not want our judgments of MWE-hood to be constrained or influenced by syntactic treebanking conventions. Second, we wanted to highlight that our framework is feasible for domains (and in principle, languages) without syntactic treebanks or parsers. Third, sequence models are computationally more efficient than parsing models. (Though our system requires POS tagging as preprocessing, that is also accomplished with an efficient sequence model.) And finally, including a parser in the pipeline would open up a large number of methodological possibilities to explore: What kind of syntactic formalism (e.g., phrase structure or dependency)? What kind of parsing algorithm (e.g., graph-based or transition-based dependency parsing)? Should the parser be trained in the domain of interest, and if not, how much would performance suffer? How should the parser output be exploited in features for MWE identification (see, e.g., Constant et al., 2012)? Is it better to identify MWEs *first*, as a preprocessing step for improved syntactic parsing (Nivre and Nilsson, 2004; Korkontzelos and Manandhar, 2010; Constant et al., 2012; Candito and Constant, 2014; de Lhoneux, 2014)? Or is MWE information best integrated into the syntax (à la Green et al., 2011, 2012; Vincze et al., 2013b; Candito and Constant, 2014), or can the MWE analyzer and the parser work simultaneously for mutual benefit (Le Roux et al., 2014)? These questions have been investigated to an extent in existing resources—primarily, for compounds in the French Treebank (see §3.3.4). We believe they are worth exploring thoroughly with heterogeneous MWEs, and so we leave this to future work, save for a brief comment in §8.3.1 below on the potential role of MWE identification in enhancing parsers.

NLP strategies aside, readers interested in linguistic theory are likely wondering what the present approach to lexical semantics means for a broader theory of grammar and compositionality (the "syntax-lexis nexus", if you will). Though ch. 3 describes MWEs using some of the formal tools that have been applied to syntax, I have made no claims about the status of these lexical semantic units in a syntactic analysis.

For theorists, MWEs challenge traditional assumptions about the separation between lexicon and grammar. The arguments have been made elsewhere (Fillmore et al., 1988; Nunberg et al., 1994; Sag et al., 2002), but to briefly list some of the possibilities: there are MWEs with

[3]Possible facets of preposition meaning include: (a) thematic roles with respect to an event/causality—our preposition supersenses largely express these; (b) spatial relations/image schemas (both static—being beside, above, on top, nearby, attached, etc., and dynamic—moving towards something, coming apart, etc.); (c) domain of experience—this is reflected in the PROFESSIONALASPECT' category for senses that didn't seem to fit well elsewhere; and (d) polarity, as evident in contrasts like *with*/*without* and *for*/*against*.

For example, dynamic spatial *off* involves moving away from a starting point (INITIALLOCATION') in a particular direction (DIRECTION') so that there is no longer contact with something (SEPARATION in the scheme of Srikumar and Roth, 2013a; we removed it in favor of thematic roles). See also footnote 1, which discusses a highly nuanced dimension of meaning. Allowing corpus annotators to apply multiple labels to the same token might be a solution to this problem.

- special component vocabulary, completely opaque form and meaning: *ceteris paribus*
- familiar component vocabulary, idiosyncratic syntax, and opaque meaning: *by and large*
- frozen and partially idiosyncratic syntax, and partially opaque meaning: *All your base are belong to us*
- familiar component vocabulary, familiar but frozen syntax, the head word inflecting regularly, and opaque meaning: *kick the bucket* (*kicked the bucket*, but **the bucket was kicked*)
- familiar and partially flexible syntax, requiring arguments and agreement, and mostly transparent meaning: *<one$_i$> give <something> <one$_i$>'s best shot* (*John gave the project his/*Mary's/*her best shot*)
- familiar and flexible syntax and figurative but decomposable meaning: *spill the beans* (*the beans were spilled*, *spill all the beans*, etc.)
- familiar but somewhat noncompositional parts arranged according to special syntactic rules: *Rev. Dr. Martin Luther King, Jr.*
- recognizable but frozen syntax and transparent semantics, but institutionalized with a special rhetorical function: *all things being equal*

In short, a full account of these expressions would need to mark what aspects of form and meaning are fixed vs. flexible and regular vs. idiosyncratic.

The philosophy of **Construction Grammar**—namely, that lexicon and grammar are endpoints on a spectrum of learned pattern/meaning associations, rather than separate mechanisms (Hoffmann and Trousdale, 2013)—seems a necessary background to such a theory. A **construction** (conventionalized form-meaning unit) can in principle map complex lexical and/or syntactic configurations

to a single meaning. Representing a language's lexicon-grammar as a network of constructions, with inheritance links between constructions that overlap in form and/or meaning, is a way to account for partially predictable but partially idiosyncratic patterns (Lakoff, 1987; Goldberg, 1995).

Computational Construction Grammar formalisms such as Embodied Construction Grammar (Bergen and Chang, 2005; Feldman et al., 2009) and Fluid Construction Grammar (Steels et al., 2011; Steels, 2012) have been implemented on a small scale, but lack a corpus for data-driven learning of broad-coverage parsers. On the other hand, approaches to parsing MWE structures with Tree Substitution Grammars (Green et al., 2011, 2012) have not incorporated any meaning representation, while for Combinatory Categorial Grammar (Steedman, 2000), semantics-enabled broad-coverage parsers (e.g., Bos et al., 2004) are not (yet) equipped to treat most kinds of multiword expressions (de Lhoneux, 2014). Thus, computationally efficient and data-driven parsing of complex, meaning-bearing constructions—MWEs as well as nonlexicalized constructions (Hwang et al., 2010b)—still presents a considerable challenge for future research (Schneider and Tsarfaty, 2013).

8.1.4 Regarding Empirical Support for Claims about Linguistic Theory and Annotation

When making hypotheses about natural language, applying relevant annotations to corpora, and building computational models to test those hypotheses, it is possible to fall into a trap of circular logic. Riezler (2013) raises several concerns about empirical validity in computational linguistics, some of which are on point here.

One concern is about the reproduceability of annotations. When, as in this thesis, a group of annotators are trained over a period of time—and especially when they are involved in shaping the guidelines themselves—it is likely that some of the consensus that emerges from that experience will be in the form of an unwritten understanding, rather than due to "pure" intuitions or articulated principles and conventions of the annotation scheme. Thus, high inter-annotator agreement may mask reliability that is due to factors outside of the annotation guidelines. Riezler suggests that ideally, naïve (even non-linguist) annotators be trained directly from the annotation guidelines to test the robustness of the scheme. Because of resource limitations, this was not possible for most aspects of the scheme proposed in this thesis, though we did find qualitatively that the noun supersense guidelines developed first for Arabic ported well to English with a different set of annotators.

Riezler also argues that extrinsic evaluations with a theory-neutral measure of "usefulness" are a valuable way to test an NLP system that produces theory-specific output. Of course, applications such as machine translation are also a major motivation for building linguistic analyzers in the first place. §8.3 considers the expected relevance of lexical semantic analysis to several extrinsic tasks.

8.1.5 Future Directions in Broad-Coverage Lexical Semantic Analysis

This thesis has not, of course, solved the lexical semantic analysis problem once and for all; much of the journey awaits. As the next step, we intend to deploy the preposition supersense scheme (ch. 5) to fully annotate a corpus and integrate preposition supersenses into the joint lexical analyzer (ch. 7). This should not require any major deviations from the approach taken for noun and verb supersenses.

Several technical directions hold promise for making models more robust. It should be possible to leverage additional unlabeled,

labeled, and type-level data sources, including data from other domains (much like Johannsen et al., 2014 have recently done for Twitter supersense tagging). We have not thoroughly inspected our annotations for consistency across sentences, so making existing data more consistent is a possible direction whose value should be weighed against the value of annotating new data. Of course, we hope that the annotation scheme will be applied to new corpora and languages, and that the guidelines can be improved where necessary to work across languages. The value of incorporating syntactic information into models deserves further investigation, as discussed in §8.1.3.

Finally, this thesis has only proposed supersense inventories for nouns, verbs, and prepositions, but the framework could be extended to additional parts of speech—ideally to the point that it is capable of covering most of the lexical semantic units in any sentence. Preliminary steps have already been taken to develop a supersense scheme for adjectives (Tsvetkov et al., 2014).

8.2 Linguistic Applications

We briefly point out that gold annotations (by humans) and silver annotations (by systems trained on the gold annotations) made possible by this thesis have the potential to enable new forms of corpus-based linguistic inquiry in lexical semantics. In particular, supersenses provide a level of abstraction that is often more conducive than words for positing and testing generalizations about language. In fact, similar schemes have been used by corpus linguists in the past (Zaenen et al., 2004). Furthermore, linguists studying idiomaticity in English will be in a much better position to use corpora (Moon, 1998, p. 51: "Ideally, the FEIs [fixed expressions and idioms] in a corpus would be identified automatically by machine, thus removing human error or partiality from the equation"). Of course, neither our annotated data nor our system is perfect. Still, we hope that our contributions will reduce the amount of manual coding required in new corpora and make possible linguistic analyses with much broader lexical coverage.

8.3 NLP Applications

With regard to other NLP tasks, prior work and future opportunities for applying broad-coverage lexical semantic analysis are worthy of comment.

8.3.1 Syntactic Parsing

MWEs. There has been some work connecting MWEs to parsing, either using parsing as a tool for identifying MWEs or using knowledge of MWEs to influence overall parsing accuracy (e.g., Nivre and Nilsson, 2004; Korkontzelos and Manandhar, 2010; Green et al., 2011, 2012; Constant et al., 2012; Candito and Constant, 2014; Le Roux et al., 2014; de Lhoneux, 2014; cf. §8.1.3 above). These attempts have met with mixed success. A concern is that it is not always clear which MWEs should be considered as syntactically idiosyncratic and which of them are merely semantically idiosyncratic, and how they should therefore be represented in a syntactic parse. One way to sidestep this issue would be to use the syntactic parse as a means to an end (such as semantic parsing or machine translation), and measure whether improved identification of MWEs in the parser correlates with downstream improvements.

Supersenses. Semantic senses and semantic classes such as the WordNet supersenses have been explored as additional information

for improving syntactic parsers (Agirre et al., 2008, 2011; Fujita et al., 2010; Bengoetxea et al., 2014). This line of work has been somewhat inconclusive, but may benefit from more accurate supervised statistical (rather than unsupervised or heuristic) supersense tagging, especially with semantic tags for prepositions as proposed in ch. 5. There has also been work specifically on the task of PP attachment (Hindle and Rooth, 1993; Niemann, 1998; Coppola et al., 2011; Greenberg, 2014, *inter alia*), which would obviously stand to benefit from a system that could semantically classify the preposition, its object, and its potential governors with high accuracy.

8.3.2 Semantic Parsing

As mentioned in §2.1, one goal in computational semantics is to analyze relationships among words or lexically-denoted concepts in a sentence via some meaning representation that provides abstraction and supports some sort of inference. Broadly speaking, this challenge is known as **semantic parsing**.

This section considers how lexical semantic analysis might aid semantic parsers. For concreteness, we focus on two of the computational representations for relational semantics in English: frame semantics and AMR.

FrameNet (Baker et al., 1998) is a linguistically rich semantic lexicon and corpus for predicate-argument structures that instantiates the theory of **frame semantics** (Fillmore, 1982) for English. The FrameNet lexicon is organized in terms of conceptual scenes, or **frames**. Associated with each frame definition is a list of **lexical units** (predicates) known to evoke the frame, as well as **frame elements**—roles that reflect conceptual attributes of the frame that may be elaborated when the frame is used. Each annotated sentence in FrameNet records one or more evoked frames; each frame

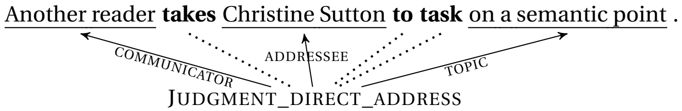


Figure 8.1: Example from the FrameNet lexicographic annotations. The gappy expression *takes… to task* is the frame-evoking target: it maps to the lexical unit *take to task.v* of the JUDGMENT_DIRECT_ADDRESS frame. The frame elements (roles) of this frame include COMMUNICATOR, ADDRESSEE, TOPIC, MEDIUM, and REASON, a subset of which are expressed overtly in the sentence. Other lexical units for this frame include *chide.v*, *compliment.{n,v}*, *harangue.v*, *tell off.v*, *telling off.n*, *tongue-lashing.n*, and *upbraid.v*.

annotation includes the evoking expression and any token spans associated with frame elements. Figure 8.1 gives an example sentence with a single frame annotation.

AMR (Banarescu et al., 2013) is a graph-based representation that canonicalizes certain aspects of logical meaning so as to abstract away from surface words and syntax, in a way that is human-readable and conducive to rapid annotation with broad coverage. Figure 8.2 displays an example. Designed primarily for English, AMR describes each sentence with a graph that encodes entity and predicate **concepts** as nodes, and semantic roles/relations as typed edges between concepts. Nodes can be shared to indicate within-sentence coreference (including implicit coreference implied by the syntax, such as with control structures and relative clauses). Event predicates and associated semantic roles are drawn from PropBank (Kingsbury and Palmer, 2002); the predicate-specific core roles from PropBank are supplemented with an inventory of non-core roles such as LOCATION, TOPIC, and POSS(ESSOR). There are also special conventions for named entities, time and value expressions, derivational morphology, modality and negation, and a host of special


```
(s / scold-01
 :ARG0 (p / person
       :ARG0-of (r / read-01)
       :mod (a / another))
 :ARG1 (p2 / person
       :name (n / name
             :op1 "Christine"
             :op2 "Sutton"))
 :ARG2 (p3 / point
       :mod (s2 / semantics)))
```

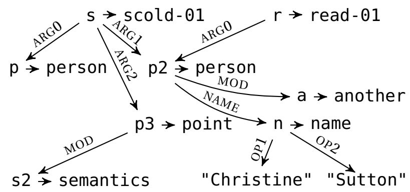


Figure 8.2: Possible AMR representation of the sentence from figure 8.1. The textual form on the left is equivalent to the graph structure on the right. An AMR is a graph whose nodes represent *concepts* and labeled edges represent *relations*. The concepts are not strictly restricted to words from the sentence: by convention, proper names are given an entity class (`person` for *Christine Sutton*); derivational morphology is "unpacked" (*reader* becomes 'person who reads'; the pertainym *semantic* is replaced with the noun *semantics*); and events are disambiguated to PropBank rolesets (`scold-01`, `read-01`). Core arguments of events are numbered, also following PropBank (`:ARG0`, etc.); non-core relations such as `:mod` and `:name` are AMR-specific.

phenomena. In contrast to FrameNet, an AMR graph is not aligned to the source sentence, and is structurally richer (hierarchical, covering more phenomena), but uses shallower lexical and relational labels.

Corpora annotated with both of these representations exist[4] and have been used to train statistical semantic parsers that take English sentences as input and predict meaning structures. The state-of-the-art system for frame-semantic parsing is SEMAFOR (Das et al., 2010, 2014).[5] To date, the only published system for AMR parsing

[4] See http://amr.isi.edu/ and https://framenet.icsi.berkeley.edu/fndrupal/index.php?q=fulltextIndex.

[5] http://www.ark.cs.cmu.edu/SEMAFOR/; https://github.com/sammthomson/semafor/

is JAMR (Flanigan et al., 2014).[6] Both of these systems attempt to build fairly rich structures but have limited data for supervision. Evaluations show there is a great deal of room for improvement in accuracy. A lexical semantic analyzer trained on other data could be incorporated as a preprocessing step to obtain additional features for the parser, alongside features derived from the output of other preprocessing steps already required by the tools.[7]

SEMAFOR and JAMR both consist of two main stages, executed in sequence. *First*, words/word sequences in the sentence are each mapped to a canonical conceptual representation—essentially, this entails word sense disambiguation of predicates. It should not be hard to see how supersenses could help such a system to disambiguate starkly polysemous lexical predicates, and (in the case of AMR) to add the appropriate semantic class for each named entity. Likewise, it is necessary to identify various kinds of multiword predicates, some of which are canonicalized to a multiword concept (e.g., *take to task* in FrameNet) while others are simplified to a single-word concept (*take to task* → `scold-01` in AMR; light verb constructions in both AMR and FrameNet). *Second*, typed links are added to indicate semantic relations—essentially, semantic role labeling. Lexical semantic analysis could help the parser avoid structural errors that would split MWEs across arguments, and the supersenses would inform relation labeling of non-core arguments (which often correspond closely to preposition supersenses[8]) as well as core arguments (whose selectional preferences could be modeled in terms of supersenses).

[6] `https://github.com/jflanigan/jamr/`

[7] Both SEMAFOR and JAMR require dependency parsing as a preprocessing step. JAMR additionally uses the output of a named entity recognizer.

[8] See table 5.1 (p. 102) for correspondences between preposition supersenses and AMR's non-core role labels. Similar correspondences should apply for FrameNet labels as well.

8.3.3 Machine Translation

Knowledge of lexical expressions and their meanings is surely integral to humans' ability to translate between two languages. But of course, machines and people work very differently. In practice, the modern **statistical machine translation (SMT)** systems with enormous amounts of data at their disposal may be coping indirectly with most of these phenomena. Would a monolingual computational model of lexical semantics be relevant to machine translation?

An example from an SMT system will be instructive. In Google Translate—for which English-French is the best language pair—both inputs in (28) are mapped to the nonsensical French output (29a) instead of to (29b), suggesting that *mind* is being translated separately from *make up*:

(28) a. She was unable to <u>make up</u> the Count's <u>mind</u>.
b. She was unable to <u>make up</u> the <u>mind</u> of the Count.

(29) a. Elle était incapable de compenser l'esprit du comte.
roughly: 'She was incapable of compensating for the spirit of the Count.'
b. Elle était incapable de convaincre le comte.
'She was incapable of convincing the Count.'

Failures such as these provide evidence that better treatment of lexical items is at least plausible as a path to better translation quality.

At the lexical level, current systems face the twin challenges of **sense ambiguity** and **multiword expressions**. The English WordNet senses of *make up* were enumerated on page 35 above. Among its major French translations are *constituer* (sense #1), *composer* (#1, #2), *fabriquer*, *faire*, and *préparer* (#2), *compenser* (#3, #7), *rattraper* (#4), *inventer* (#5), *ranger* (#6), *pallier* (#7), *se réconcilier* (#8), and *maquiller* (#9). Further, the idiom *make up… mind* translates to *se décider*. If the local context is insufficiently informative for the language model, an MT system might easily translate the wrong sense of *make up*. And if *make up* is not translated as part of the same unit (especially likely if it contains a gap), the overall bias for *make* translating as *faire* would probably prevail, and the *up* ignored entirely—or worse, mistranslated as a spatial term. Verb-noun constructions such as *make up... mind* are even more prone to disaster because they are more likely to be realized with a gap, as shown above.

Analysis and experimentation is therefore needed to establish the extent to which the explicit information in an English lexical semantic representation is orthogonal to, or redundant with, translation units learned and selected by a full-scale MT system.

Supersenses vs. WSD. Several attempts have been made to integrate word sense disambiguation into SMT systems. The disambiguation problem has been formulated with an explicit sense inventory (Carpuat and Wu, 2005), with lexical-level translations (Cabezas and Resnik, 2005; Chan et al., 2007; Carpuat and Wu, 2007), and with unsupervised topics (Xiong and Zhang, 2014; Hasler et al., 2014). In all of these methods, WSD is performed on the source side in order to capture wider context than is allowed in translation rules (cf. Gimpel and Smith, 2008). We are unaware of any WSD-for-SMT studies that have used prespecified *coarse-grained* senses such as supersenses, which would perhaps lead to better generalizations.

Name translation is a major obstacle in SMT due to unknown words (see Hermjakob et al., 2008 for a review), a problem which we do not expect supersenses to solve.

Prepositions. Prepositions are known to be especially challenging for machine translation (Gustavii, 2005), and are a high-value

target due to their frequency. Yet surprisingly, adpositions have received little attention in the SMT paradigm (Baldwin et al., 2009). Exceptions are the work of Toutanova and Suzuki (2007), who use a target side reranker for Japanese case-marking postpositions in an English-to-Japanese system, and the work of Shilon et al. (2012), who incorporate information about prepositions into translation rules for an Arabic-to-Hebrew system. Preposition supersenses, one hopes, would go a long way toward disambiguating the translation. For example, two of the French equivalents of *for* are the prepositions *pour* (GOAL, DESTINATION) and *pendant* (DURATION).

MWEs. Recent quantitative evaluations of MWEs in machine translation systems (especially for verb-particle constructions, prepositional verbs, and support verb constructions) underscore the challenges noted above (Barreiro et al., 2013, 2014; Ramisch et al., 2013). For instance: Barreiro et al. (2014), analyzing the performance of two MT systems across five language pairs (English into Portuguese, Spanish, French, Italian, and German), find that anywhere from 27% to 70% of support verb constructions are erroneously translated.

Techniques for adapting SMT systems to capture MWEs have included altering the tokenization of the text so MWEs constitute a single token; expanding the training data with monolingual paraphrases of MWEs; expanding the phrase table with a bilingual MWE lexicon; marking phrase table entries that capture MWEs with a feature that rewards their use in decoding; and constraining reorderings of words belonging to MWEs (Nakov, 2008; Ren et al., 2009; Carpuat and Diab, 2010; Ramisch, 2012; Ghoneim and Diab, 2013; Simova and Kordoni, 2013). Some of these strategies have been more successful than others, and different strategies work well for different kinds of MWEs. Because most of these methods rely on token-level identification of MWEs, it is hoped that upstream improvements to lexical semantic analysis will drive further gains.

8.3.4 Other Applications

Multiword chunks are an important phenomenon of study in both **first language acquisition** (Bannard and Lieven, 2012) and **second language acquisition** and education (Wray, 2000; Ellis et al., 2008).[9] Prepositions are notoriously difficult for second language learners, especially given their prevalence in multiword expressions, so they occupy a central place in the literature on automatic grammatical error correction (Chodorow et al., 2007; Hermet and Alain, 2009; Leacock et al., 2014). It would be interesting to see how well a lexical semantic analyzer trained on native English text would perform on nonnative writing, and whether supersense tagging could draw attention to anomalous usages.

Spatial and temporal analysis tasks such as SpaceEval[10] and TempEval (e.g., UzZaman et al., 2013), and related applications in robotics and computer vision (e.g., Dobnik and Kelleher, 2013, 2014), may benefit from the supersense analysis of prepositions, particularly the temporal (§5.4.3) and path (§5.4.5) portions of the hierarchy.

For information retrieval, segmenting or extracting multiword units in a text has been explored under various guises, including **keyphrase extraction** and **query segmentation** (e.g., Tomokiyo and Hurst, 2003; Tan and Peng, 2008; Acosta et al., 2011; Newman et al., 2012). Keyphrase extraction also has application to opinion mining (Berend, 2011). Segmentation of text on a semantic basis (though with looser criteria than proposed here) has been explored for distributional semantic models (Srivastava and Hovy, 2014). Tasks involving text chunking followed by semantic class assignment can be found in the biomedical information extraction literature (e.g., Stenetorp et al., 2014).

[9] A colleague in Pittsburgh reports that his young daughter says *upside up* by analogy to *upside down*, rather than the usual *right-side up*. Children are sometimes more logical than adults.

[10] http://alt.qcri.org/semeval2015/task8/

APPENDIX A

MWE statistics in PARSEDSEMCOR

A subset of Brown Corpus documents are both fully sense-tagged in SemCor (Miller et al., 1993; see §2.3.2) and parsed in version 3 of the Penn Treebank (Marcus et al., 1999). We will refer to this collection as **PARSEDSEMCOR**. A profile of the dataset appears in figure A.1.

Looking at the SemCor annotations of the 93 documents in the **PARSEDSEMCOR** collection, we find 220,933 words in 11,780 sentences. There are 5590 named entity mentions; of these, 1861 (1240 types) are multiword NEs, spanning 4323 word tokens (2% of the data).[1] An additional 6368 multiword expression mentions (3047 types) are annotated, encompassing 13,785 words (6% of the data). About 87% of these mentions (and 87% of types) are tagged

[1] For the type counts in this paragraph, mentions were grouped by their lowercased surface string.

# docs		genre
16	F	POPULAR LORE
15	G	BELLES-LETTRES (biographies, memoirs)
28	K	FICTION (General)
11	L	FICTION (Mystery/Detective)
2	M	FICTION (Science)
10	N	FICTION (Adventure/Western)
5	P	FICTION (Romance/Love Story)
6	R	HUMOR

Figure A.1: Composition of the **PARSEDSEMCOR** dataset, which is the parsed and fully sense-tagged subset of the Brown corpus. Parses and sense tags are gold standard. The 93 documents in this sample consist of about 2200–2500 words each, a total of 220,933 words in the SemCor tokenization.

with a WordNet sense.[2] All told, 8% of tokens in **PARSEDSEMCOR** belong to a SemCor-annotated MWE, with a 3-to-1 ratio of multiword idioms to multiword NEs.

A.1 Gappy MWEs

To identify gappy MWEs in the **PARSEDSEMCOR** collection, including those in figure 3.4, we extracted the sense-tagged items for which the number of words in the lemma differed from the number of words in the tagged surface span—this usually indicates a gap.[3]

[2]The 30 most frequent MWEs to be annotated without a sense tag are: *going to* (62), *had to* (34), *have to* (32), *most of* (28), *of it* (23), *no one* (19), *as well as* (15), *as long as* (13), *of this* (13), *in order* (13), *in this* (13), *in front of* (12), *in that* (10), *got to* (9), *as soon as* (9), *even though* (9), *many of* (9), *used to* (8), *as though* (8), *rather than* (8), *of what* (7), *up to* (7), *a lot* (6), *such as* (6), *as much as* (6), *want to* (6), *of that* (6), *out of* (6), *in spite of* (5), *according to* (5). These include complex prepositions, comparative expressions, and discourse connectives not in WordNet. The expression *a lot* is in WordNet, but is missing a sense tag in some of the documents.

[3]E.g., the lemma `make_up.05` would be marked for the verb and particle as a unit in *make up the story*, but for only the head verb in *make* ⟨ *the story* ⟩ *up*. Cases differing only in punctuation (e.g. hyphenation) were excluded.

There are 336 occurrences of mismatches, with 258 distinct lemma types. Of these types, a majority—about 160—are particle verbs or prepositional verbs. About 20 types are verb-noun constructions; 7 are verb-PP idioms. Roughly 30 are complex nominals, some of which are legitimately gappy and some of which have a lemma slightly more specific than the surface word (e.g. *the Church* mapped to `Roman_Catholic_Church.01`). Finally, 11 types are non-standard spellings (*suns of biches* is mapped to `son_of_a_bitch.01`), and 2 types were variant forms of the lemma: *physiotherapist* as `physical_therapist.01`, *co* as `commanding_officer.01`.

From these results we estimate that fewer than 2 gappy MWEs are annotated for every 1000 words of SemCor. However, we suspect SemCor annotators were conservative about proposing canonically gappy expressions like verb-noun constructions.

APPENDIX B

MWE Annotation Guidelines

These guidelines, published at `https://github.com/nschneid/nanni/wiki/MWE-Annotation-Guidelines`, document the conventions of grouping tokens of text into multiword expressions. See §3.6 for discussion.

MWE Annotation Guidelines

nschneid edited this page on Mar 26 · 18 commits

This document gives a detailed description of a linguistic annotation scheme for multiword expressions (MWEs) in English sentences. A conference paper summarizing the scheme and a dataset created with it are available at http://www.ark.cs.cmu.edu/LexSem/.

Markup

The input to annotators is a tokenized sentence. The goal is to join tokens where appropriate; this is done with the special characters underscore (_) for strong multiword links, a tilde (~) for weak links:

- This is a highly`~`recommended fast`_`food restaurant .

Weak links can join strongly-linked expressions, such as fast_food + chain:

- This is a highly`~`recommended fast`_`food`~`chain .

Where an expression is interrupted by other tokens, use trailing and leading joiners:

- Do n't give`_` Jon such `_`a`_`hard`_`time !

This even works when contiguous expression can fall within the gap:

- Do n't give`_` Jonathan`_`Q.`_`Arbuckle such `_`a`_`hard`_`time !

On rare occasion it may be necessary to use multiple gappy expressions, in which case indexing notation is available: `a|1 b|2 c|1$3 d|2 e$3` implies two strong expressions—a_c and b_d—and one weak expression, a_c~e. An example: `put_ whole$1 _heart_in$1` (amounts to: put_heart_in~whole). Also: `make_a_ big$1 _to_do$1`

Tokenization, hyphenation, spelling, and morphology

Don't correct spelling mistakes in general, but interpret misspelled words according to their intended meaning. (E.g., if *put it* is clearly a typo of the particle verb *put in*, mark put_in.)

Do add `_` if two separate tokens should really be one word. Include intermediate hyphens as part of an MWE:

anti - oil wells

or

anti - oil`_`wells

or

anti`_`-`_`oil`_`wells

NOT

anti`_`- oil`_`wells

or

anti -`_`oil`_`wells

If there's a nonstandard variant of a term (more than just a misspelling), don't join: craft_beer BUT handcraft beer

In general, don't worry about the **inflection** of the word: come_in, came_in, have been coming_in all map to the same MWE lemma; likewise for grocery_store, grocery_stores.

If different pronouns can be used in a slot, do not join the pronoun. But if the slot requires a possessive, do join the possessive clitic if present: take`_` her `_`time, take`_` one `_`'s`_`time

Figurative language

While many idioms/MWEs are figurative, not all figurative language is used in a lexically specific way. For example, "my Sicilian family" referring to a pizza community may be nonliteral, but is not an MWE.

Foreign languages

We do not annotate foreign sentences, but foreign names within an English sentence are MWEs.

Collocations vs. "strong" MWEs

A **collocation** is a pairing between two or more words that is unexpectedly frequent, but not syntactically or semantically unusual.

- eternally~grateful, can~not~wait, place~is~packed
- what s.o. has~to~say [is willing to say, brings to the conversation; not obligation (compare: *I have to say good things to my boss to get promoted*)]

A collocation may include components that are themselves MWEs:

- after~all~was_said_and_done

Drawing the line between free combination, collocation, and multiword is often difficult; annotators' opinions will vary.

Borderline/ambiguous cases

TODO (join with ~)

- Cj~and_company (ambiguous whether it is actually the name of a company or a guy and his crew)

Constructions with only 1 lexicalized word

Some semi-productive idioms are not well captured as lexicalized multiwords. These should not be joined:

- have + GOODNESS.ADJ + TIME.PERIOD: had a bad day, have a great year, etc.
- EVALUATIVE.ATTRIBUTE of s.o.: (real) Christian of you
- NUMERIC.QUANTITY PLURAL.TIME.NOUN running: two years running
- come + MENTAL.CHANGE.INFINITIVE: come to realize, believe, learn, adapt, have faith, …

Overlapping expressions

Rarely, a token will seemingly participate in multiple MWEs, which cannot be represented in our annotation scheme. Use your best judgment in such cases.

- I recently threw a surprise birthday party for my wife at Fraiser 's .

Possible pairs:

surprise_party

birthday_party

threw_party

Decision:

threw~birthday_party

- triple_chocolate_chunk brownie: multiplier+chocolate, chocolate_chunk

Syntactically perverted expressions

Don't worry if the parts of an expression are noncanonically ordered: gave`_`estimates, give`_` an `_`estimate, the estimate`_` that was `_`given

If one of the lexicalized parts is repeated due to coordination, attach the instance closest to the other lexicalized part: talked_to Bob and to Jill; North and South_America

Special kinds of expressions

- DO join Dr., Mr., etc. and other titles to a **personal name**: `Dr._Lori_Levin`, `Henry_,_Prince_of_Wales`, `Captain_Jack_Sparrow`
- DO join Ave., Rd., St., etc. in **street names**: Forbes_Ave.
- DO join **city-state-region** expressions: Bellevue~,~WA or Bellevue~WA (include the comma if there is one). Likewise: Ohiopyle_State_Park~,~Pennsylvania; Miami_University~,~Miami~,~Ohio; Amsterdam~,~The_Netherlands
- DON'T join normal **dates/times** together (but: Fourth_of_July for the holiday)
- Symbols
  - DON'T join normal **%** sign
  - DO join **letter grade followed by plus or minus**: A_+
  - DON'T join **mathematical operators**: 3 x the speed, 3 x 4 = 12 [x meaning "times"]
  - DO join # sign when it can be read as "number": #_1
- DO join a number and "star(s)" in the sense of a rating: 5_-_star
- When in doubt, join **cardinal directions**: north_east, north_west, south_east, south_west, north_-_northeast, …
- DO attach 's if part of the name of a **retail establishment**: Modell_'s
- DO join **product expressions** such as car Year/Make/Model or software Name/Version
  - excludes appositions that are not in a standard format (McDonald's Dollar_Menu Chicken Sandwich)
- DO join names of **foods/dishes** if (a) the expression is noncompositional in some way, or (b) there is well-established cultural knowledge about the dish. Use ~ if unsure. For example:
  - General_Tso_'s_chicken, macaroni_and_cheese, green_tea, red_velvet cake, ice_cream_sandwich, chicken_salad salad
  - triple_chocolate_chunk brownie [multiplier+chocolate, chocolate_chunk]
  - pizza~roll, ham~and~cheese, cheese~and~crackers, spaghetti~with~meatballs
  - grilled BBQ chicken, pumpkin spice latte, green pepper, turkey sandwich, eggplant parmesan, strawberry banana milkshake
- DO join established varieties of **animals/natural kinds**: yellow_lab, desert_chameleon, Indian_elephant, furcifer_pardalias; BUT: brown dog
- DO join **slogans**: Trust_The_Midas_Touch, Just_Do_It, etc.

By construction

A+[P+Pobj]

- pleased/happy/angry_with, mad_at
- good_for s.o. [healthy, desirable]

affective *on*

This is a special use of the preposition ‘on’, but it is does not generally join to form an MWE:

- drop_the_ball on s.o. [not literally!], die on s.o.
- hang_up~on s.o. [collocation]
- (‘step on s.o.’ is different: here it is the semantics of ‘step on’ that could convey negativity in certain contexts, not ‘on’ by itself)

age

- (appropriate) for_ {one's, a certain, ...} _age (of child)

age construction: TEMPORAL.QUANTITY *old*

3 years_old, month_old project. (Note that *ago* should NOT be joined because it is always postpositional.)

all + A

Join unless ‘all’ is paraphrasable as ‘completely’ or ‘entirely’:

- *participle:* all gone, all done, all_told [overall, in total]
- *other adj:* all ready, all_right well, OK

as X *as* Y

Do not join, even though the *as* PPs are correlated. Exceptions:

- as_long_as [while]

X *by* Y

- one_by_one [one at a time]
- Don’t join if *by* indicates a product, as in a multidimensional measurement: three by five

paper = 3 x 5 paper

classifiers: MEASURE.WORD *of* N

A few English nouns take idiosyncratic measure words: 3 sheets~of~paper, 2 pairs~of~pants, a piece~of~information

clear/straight/right + P

Do not attach the modifier if it has an ordinary meaning, e.g. *go clear through the wall*

complex adjective phrases

- highly~recommended, highly~trained
- family~owned company

complex nominals: A+N, N+N

- capital_punishment
- big_rig [slang for truck]
- road_construction [the road isn't actually being constructed, but reconstructed!]
- silver_ Mariott _member [rewards program]
- electric_blanket
- last_minute
- price_range
- second_chance
- grocery_stores
- pizza_parlor, pizza place, burger joint (diagnostic: does “favorite X” occur on the web? [to filter out proper names])
- little~danger/risk
- public~welfare
- this place is a hidden~gem
- strike_one/two/three (unusual syntax!)

complex prepositions

Cf. Quirk pp. 669–670

- out_of, in_between, in_front_of, next_to
- along_with
- as_well, as_well_as
- in_addition_to, in_light_of, on_account_of
- due_to, owing_to, because_of

complex subordinators

From Quirk et al. 1972:

- but_that, in_that, in_order_that, insofar_that, in_the_event_that, save_that, so_that, such_that, except_that, for_all_that, now_that
- `as_{far,long,soon}_as`, inasmuch_as, insofar_as, as_if, as_though, in_case
- Do NOT mark the participial ones: assuming, considering, excepting, … that

discourse connectives

- to_start_off_with
- that_said
- of_course

else

Though as a postmodifier it is a bit odd syntactically (*anything else*, *who else*, etc.), it does not generally participate in lexicalized idioms.

- “What does ‘else’ even *mean*?!” - Henrietta

exhortative, emotive, expletive, and proverb idioms

- `do_ X _a_favor_and Y`
  `do_ X _a_favor_, Y`
  vs. plain `do_favor`
- you_get_what_you_pay_for (NOT: ‘you get what you purchase’)
- get_ the_hell _out
- why in_the_hell [can be any WH word]
- do_n’t_forget, forget_it !, never_mind
- i have~to~say, gotta~say, etc.: semantics = obligation on self to say something, pragmatics = can’t restrain self from expressing an opinion
- Who_knows [rhetorical question]
- no_way
- **Phatic expressions**: I_’m_sorry, Thank_you

existential *there*

We do not mark ordinary *there be* existentials as multiwords.

get + "accomplishment" V

`get_upgraded`, `get_ cat _neutered`, `get_ a bill _passed`

get + destination

In the sense of ‘arrive’, not really a multiword:

- get back home, got to the school

get + result A

- get_ready, get_done, get_busy, get_older
- get_a_flat
- get_correct

infinitival *to*

If a verb, adjective, or noun typically takes an infinitival complement, and the infinitive verb is not fixed, don't join *to*:

- little to say
- important to determine his fate
- able/ability to find information
- chose/choice to do nothing
- willing(ness) to sail

But if it is a special construction, the *to* should be joined:

- in_order_to VP
- at_liberty_to VP
- ready_to_rumble
- special modal/tense constructions: ought_to, have_to (obligation or necessity), going_to, about_to (but want to, need to, try to)

long + TIME.PERIOD

- a long~{day, week, year} (long in the sense of bad/difficult; cannot be referring to an actual duration because the noun is a time unit with precise duration)

negative polarity items

Join these (including *n’t* and *a*, but not *do*) if sufficiently conventional: *did n't_sleep_a_wink*, *did n't_charge_a_cent/penny/dime*, *did n't_eat_a_morsel/scrap/bite/crumb* (of food)

on: see affective *on*

prepositional phrase idioms

These include the so-called **determiner-less PPs** (in_town vs. in the town).

- in_a_ nice/good/... _way
- out_of_site
- on_staff
- at_all
- at_liberty_to
- for_sure
- mediocre at_best
- to_boot
- in_town
- on_earth, on~the~planet, in~the~world, in the country/universe

in + method of payment

- in_cash/quarters/Euros

to + mental state

- to_ her *amusement, to* our *chagrin, to* the _surprise of all present
- to_ my _satisfaction

prepositional nouns: N+[P+Pobj]

- capacity_for love
- his problem_with the decision
- extensive damage_to the furniture

Sometimes these participate in verb-noun constructions:

- have_a_problem_with [be annoyed with], have_ a _problem_with [have trouble with something not working]
- do_damage_to the furniture

prepositional verbs: V+[P+Pobj]

- TODO: explain principles
- talk_with/to, speak_with/to, filled_with
- NOT: learn about, booked at (hotel)
- wait_for

- look_for
- test_for
- (a)rising_from
- disposed_of
- take_care_of
- trust_with
- listen_to, pay_attention_to
- `compare_ X _to Y`, `X compared_to Y`
- been_to LOCATION/doctor etc.
- trip_over, trip_on
- (do_)damage_to the furniture
- take_in [bring to an establishment for some purpose, e.g. a car for service]
- focus_on
- nibble/nosh/snack/munch_on
- kept_up_with [keep pace, manage not to fall behind]
- looking_for my friend [seeking out my friend] vs. looking for my friend [on behalf of]
- not multiword:
  - stay at hotel
  - supply with, fit_out with ['with' marks the thing transferred/given]

proximity expressions: A+P, A+P+P

Join: *close_to*, *close_by(_to) far_from*, *far_away(_from)*

same

- Join article if not followed by a noun (paraphrasable with 'identical'): his objective was the_same each time / each time he had the same objective
- exact_same

there: see existential *there*

verbs with intransitive prepositions

V+P

- walks_around (path covering an area)
- stay~away = keep_away
- run_out [get used up] vs. run out_of the filter [leak]
- *back*
  - Generally do not include literal uses: go back [motion], came/headed back [returned to a location]

    - money_back, cash_back [change of medium: overpaying with credit card so as to receive the difference in cash]
    - s.o.'s money_back, CONDITION or_your_money_back [refund]
    - pay_ s.o. _back, get money back [returning a loan; *get money back* is possible but not really idiomatic with this meaning]
    - brought back [taking a car to the shop again for further repairs] vs. brought_back [returning a purchase for a refund]
    - turned_back [turned around to travel back]
    - get_back_to s.o. [return to communicate information to s.o.]

particle verbs: V+P+Vobj, V+Vobj+P

- TODO: explain principles
- rent_out
    - with 'out', disambiguates the landlord/permanent owner vs. tenant/temporary user
    - (BUT: rent out an entire place?)
- turn_on, turn_off
- pick_up [retrieve from store]

quantifiers/quantity modifiers

- If prenominal, don't join *of*: a_lot/little/bit/couple/few (of), some/plenty of
    - EXCEPTION: a_number_of (TODO: why?) check what H did in xxxx5x
- Join 'square' or 'cubic' within a unit of measurement: square_miles/yards/..., cubic_centimeter/...
- Join half_a when modifying a quantity: half_a day's work, half_a million
- cf. **classifiers**

quantity comparisons

- *less than, more than*: Join if functioning as a relational "operator." Heuristic: can '<' or '>' be substituted?
    - less_than a week later ('< a week')
    - more happy than sad (NOT: '> happy than sad')
    - I agree with him more than with you (NOT: '> with you')

VP idioms

- trying_to_say (with implication of dishonesty or manipulativeness)
- went_out_of_ their _way (went to extra effort)
- go_on_and_on [= talk endlessly] (cf. go_on by itself meaning 'continue')
- went_so_far_as_to_say: include 'say' because it has a connotation of negativity (beyond

'went_so_far_as_to (do something)')
- have_a_gift_for [= possess talent]
- `love_ Boston _to_death`
- go_to_the_bathroom [use the potty]
- `give_ X _a_try` [test it out]
- dropped_the_issue, drop_the_subject, drop_it !
- `say/tell/lie_ (s.t.) to_ s.o. _'s_face`
- made_to_order
- took~forever

support verb constructions

A **support verb** is a semantically "light" (mostly contentless) verb whose object is a noun that by itself denotes a state, event, or relation; the noun can be thought of as providing the bulk of the meaning/determining the meaning of the verb [FN Book 2010, p. 31]. We join the verb to the noun in question:

- make_ a _decision/statement/request/speech/lecture
- take_ a(n) _test/exam
- take_ a _picture/photo
- give_speeches/lectures/interviews
- undergo/have/receive/get_ an _operation
- do/perform_surgery

Some useful properties:

1. Most commonly, support verbs are extremely frequent verbs that exhibit a certain degree of grammaticalization: *have*, *get*, *give*, *make*, *take*, etc.

2. One indication of lightness is when the noun cannot felicitously be omitted in a question (She made a decision. / #What did she make?; She had an operation. / ?#What did she have?; They perform surgery on newborns. / #What do they perform on newborns?)

3. Support verb constructions can often be paraphrased with a semantically "heavy" verb, which may be derivationally related to the noun: `make_ a _decision` = decide, `give_ an _interview` = be interviewed, `undergo_ an _operation` = be operated_on. (The noun *surgery* has no verb in English, but we could imagine "surgure" as a word! In other cases it would be not unreasonable to incorporate the noun into a verb: `take_ a _test` = test-take.)

4. Caution is required: some expressions are not support verbs, though they appear to be at first blush:

    - get a donation: *donation* refers to the money donated, not the act of donating. (What did she get in the mail? A donation.)
    - have a barbecue: here *have* has the sense of *hold* (an organized event). (What did she have at her house? A barbecue.)

- have a disease/an illness
- witnessed an operation: the verb and the noun refer to distinct events.

5. NOTE: We exclude the copula from our definition of support, though on rare occasions an idiom lexicalizes a copula: be_the_case.

Following [Calzolari et al. 2002], we distinguish “Type II” support verbs which do contribute some meaning, though it is understood in the context of the event/scenario established by the noun:

- start~ a ~race
  - most aspectual verbs—begin/end/start/stop/continue/finish/repeat/interrupt etc.—would qualify when used with an eventive noun
- pass~ an ~exam
- keep~ a ~promise
- answer~ a ~question
- execute~ a ~program, evaluate~ a ~function

Type II support verbs are lower priority for us than “core” support verbs.

verb-noun idioms

Some verb-object combinations are idiomatic, though they do not qualify as support verb constructions. We count these as multiwords as well.

- pay_attention
- *take...time*: There are several related idioms involving use of one’s time for some purpose. Include ‘the’ for the “extra effort” sense: take_the_time to help. Include a preposition for took_time_out_of (sacrifice), took_time_out/off (scheduled vacation).
- waste/spend/save/have~time/money
  - CHANGE: was waste_time
- `give_ an _estimate`, `give_ a _quote on something` [typically includes the process of estimation as well as offering that estimate to the customer]

well V-ed

Typically, don’t join:

- well done/made/oiled

Exceptions:

- my hamburger is well_done
- that was a_job_well_done
- a well_-_oiled_machine
- he is well_-_read
- well_fed

Changes requiring revisions of old annotations

- a_lot
- N_star
- in_cash
- possibly: get/have_done (hair, etc.)
- highly~recommended, highly~trained
- city, state, etc. locations: _ => ~
- waste_time, spend_time => ~

TODO

- good job, great job, look good
- good job, good work, hard work [I’d be OK with ~ for these but we decided previously that *good/great work* should be left alone]
- ‘look at it’: include ‘it’? could be specific or not
- Short of that, One more thing -- ?
- fix problem -- I'd say this is a collocation, so fix~problem
- best restaurant out_there
- fast and friendly [sounds slightly better than “friendly and fast”, but that probably reflects a preference for the word with fewer syllables to come first]
- walk_in_the_door: entering a room or establishment
- have/get_ nails/hair/tattoo/etc. _done (grooming)
- ?? have/get done [work/repairs]
- ?? do~work/job (cf. surgery)
- ?? do~dishes

APPENDIX C

MWE Patterns by POS

The following table shows all POS sequences occurring in at least 10 MWEs in version 1.0 of the CMWE corpus (49 patterns). Contiguous and gappy MWE instances are counted separately. POS groupings are abbreviated with a single character (`N` for common nouns, ˆ for proper nouns, `T` for particles, etc.). Strong MWEs are joined with _ and weak MWEs with ~; weak MWE examples are italicized. MWE types occurring at least 10 times are bolded.

POS pattern	MWEs *contig.*	*gappy*	most frequent types (lowercased lemmas) and their counts
N_N	331	1	**customer service: 31** oil change: 9 wait staff: 5 garage door: 4
ˆ_ˆ	325	1	santa fe: 4 dr. shady: 4
V_P	217	44	**work with: 27** **deal with: 16** **look for: 12** **have to: 12** ask for: 8
V_T	149	42	**pick up: 15** **check out: 10** show up: 9 end up: 6 give up: 5
V_N	31	107	take time: 7 give chance: 5 waste time: 5 have experience: 5
A_N	133	3	front desk: 6 top notch: 6 last minute: 5
V_R	103	30	**come in: 12** come out: 8 take in: 7 stop in: 6 call back: 5
D_N	83	1	**a lot: 30** **a bit: 13** a couple: 9
P_N	67	8	**on time: 10** in town: 9 in fact: 7
R_R	72	1	**at least: 10** at best: 7 as well: 6 of course: 5 at all: 5
V_D_N	46	21	**take the time: 11** do a job: 8
V~N	7	56	*do job: 9* *waste time: 4*

POS pattern	MWEs *contig.*	*gappy*	**most frequent types (lowercased lemmas) and their counts**
^_^_^	63		home delivery service: 3 lake forest tots: 3
R~V	49		***highly recommend: 43*** *well spend: 1* *pleasantly surprise: 1*
P_D_N	33	6	over the phone: 4 on the side: 3 at this point: 2 on a budget: 2
A_P	39		pleased with: 7 happy with: 6 interested in: 5
P_P	39		**out of: 10** due to: 9 because of: 7
V_O	38		**thank you: 26** get it: 2 trust me: 2
V_V	8	30	get do: 8 let know: 5 have do: 4
N~N	34	1	*channel guide: 2* *drug seeker: 2* *room key: 1* *bus route: 1*
A~N	31		*hidden gem: 3* *great job: 2* *physical address: 2* *many thanks: 2* *great guy: 1*
V_N_P	16	15	**take care of: 14** have problem with: 5
N_V	18	10	mind blow: 2 test drive: 2 home make: 2
^_$	28		bj s: 2 fraiser 's: 2 ham s: 2 alan 's: 2 max 's: 2
D_A	28		**a few: 13** **a little: 11**
R_P	25	1	all over: 3 even though: 3 instead of: 2 even if: 2
V_A	19	6	make sure: 7 get busy: 3 get healthy: 2 play dumb: 1
V_P_N	14	6	go to school: 2 put at ease: 2 be in hands: 2 keep in mind: 1
#_N	20		5 star: 9 2 star: 2 800 number: 1 one bit: 1 ten star: 1 360 restraunt: 1
N_A	18		year old: 9 month old: 3 years old: 2 cost effective: 1 lightning fast: 1
V~R	11	6	*stay away: 3* *go in: 2* *bring back: 2* *recommend highly: 2* *work hard: 1*
N_P_N	14	2	chest of drawers: 2 man of word: 1 bang for buck: 1 sister in law: 1
N~V	6	10	*job do: 2* *work do: 2* *picture take: 1* *care receive: 1* *operation run: 1*
R_V	15	1	well do: 4 never mind: 2 better believe: 1 well know: 1
N_R	15		night out: 3 hands down: 3 thanks again: 3
N_-_N	14		a / c: 2 jiu - jitsu: 2
P~D~N	14		*in the world: 3* *around the corner: 2* *for some reason: 2*
V_R_P	12	1	look forward to: 3 talk down to: 2 have yet to: 1 be there for: 1
A_A	13		west indian: 3 old fashioned: 1 up front: 1 spot on: 1 tip top: 1 dead on: 1
V_T_P	11	2	watch out for: 2 make up for: 2 put up with: 2 turn over to: 1
P_P_N	10	2	out of business: 3 out of town: 2 out of date: 1
N_P	12		nothing but: 2 increase in: 1 damage to: 1
P_N_P	11		in front of: 3 on top of: 2 in need of: 1 in spite of: 1 in search of: 1
A_N_N	11		criminal defense lawyer: 2 purple hull pea: 1 social security numbers: 1
N_N_N	11		search engine optimization: 2 kung pao chicken: 1
N_&_N	10		spay and neuter: 2 give and take: 1 bar and grill: 1 hit or miss: 1
G_A	10		over priced: 4 over cooked: 1 miss informed: 1 out standing: 1
^_^_^_^	10		bexar county tax office: 1 anna maria jose mudo: 1
P_R	10		by far: 8 if ever: 1 of late: 1

APPENDIX D

Noun Supersense Tagset

Here is the complete supersense tagset for nouns. Each tag is briefly described by its `symbol`, NAME, short description, and **examples**.

`O` **NATURAL OBJECT** natural feature or nonliving object in nature **barrier_reef nest neutron_star planet sky fishpond metamorphic_rock Mediterranean cave stepping_stone boulder Orion ember universe**

`A` **ARTIFACT** man-made structures and objects **bridge restaurant bedroom stage cabinet toaster antidote aspirin**

`L` **LOCATION** any name of a geopolitical entity, as well as other nouns functioning as locations or regions **Cote_d'Ivoire New_York_City downtown stage_left India Newark interior airspace**

`P` **PERSON** humans or personified beings; names of social groups (ethnic, political, etc.) that can refer to an individual in the singular **Persian_deity glasscutter mother kibbutznik firstborn worshiper Roosevelt Arab consumer appellant guardsman Muslim American communist**

`G` **GROUP** groupings of people or objects, including: organizations/institutions; followers of social movements **collection flock** army meeting clergy Mennonite_Church trumpet_section health_profession peasantry People's_Party U.S._State_Department University_of_California population consulting_firm communism Islam (= set of Muslims)

$ **SUBSTANCE** a material or substance krypton mocha atom hydrochloric_acid aluminum sand cardboard DNA

H **POSSESSION** term for an entity involved in ownership or payment birthday_present tax_shelter money loan

T **TIME** a temporal point, period, amount, or measurement 10_seconds day Eastern_Time leap_year 2nd_millenium_BC 2011 (= year) velocity frequency runtime latency/delay middle_age half_life basketball_season words_per_minute curfew August industrial_revolution instant/moment

= **RELATION** relations between entities or quantities, including ordinal numbers not used as fractions ratio scale reverse personal_relation exponential_function angular_position unconnectedness transitivity

Q **QUANTITY** quantities and units of measure, including cardinal numbers and fractional amounts 7_cm 1.8_million 12_percent/12% volume (= spatial extent) volt real_number square_root digit 90_degrees handful ounce half

F **FEELING** subjective emotions indifference wonder murderousness grudge desperation astonishment suffering

M **MOTIVE** an abstract external force that causes someone to intend to do something reason incentive

C **COMMUNICATION** information encoding and transmission, except in the sense of a physical object grave_accent Book_of_Common_Prayer alphabet Cree_language onomatopoeia reference concert hotel_bill broadcast television_program discussion contract proposal equation denial sarcasm concerto software

ˆ **COGNITION** aspects of mind/thought/knowledge/belief/ perception; techniques and abilities; fields of academic study; social or philosophical movements referring to the system of beliefs Platonism hypothesis logic biomedical_science necromancy hierarchical_structure democracy innovativeness vocational_program woodcraft reference visual_image Islam (= Islamic belief system) dream scientific_method consciousness puzzlement skepticism reasoning design intuition inspiration muscle_memory skill aptitude/talent method sense_of_touch awareness

S **STATE** stable states of affairs; diseases and their symptoms symptom reprieve potency poverty altitude_sickness tumor fever measles bankruptcy infamy opulence hunger opportunity darkness (= lack of light)

@ **ATTRIBUTE** characteristics of people/objects that can be judged resilience buxomness virtue immateriality admissibility coincidence valence sophistication simplicity temperature (= degree of hotness) darkness (= dark coloring)

! **ACT** things people do or cause to happen; learned professions meddling malpractice faith_healing dismount carnival football_game acquisition engineering (= profession)

E **EVENT** things that happens at a given place and time bomb_blast ordeal miracle upheaval accident tide

R **PROCESS** a sustained phenomenon or one marked by gradual changes through a series of states oscillation distillation overheating aging accretion/growth extinction evaporation

X **PHENOMENON** a physical force or something that happens/occurs electricity suction tailwind tornado effect

\+ **SHAPE** two and three dimensional shapes hexahedron dip convex_shape sine_curve groove lower_bound perimeter

D **FOOD** things used as food or drink Swiss_cheese rutabaga eggnog cranberry_sauce Guinness shrimp_cocktail

B **BODY** human body parts, excluding diseases and their symptoms femur prostate_gland ligament insulin gene hairstyle

Y **PLANT** a plant or fungus acorn_squash Honduras_mahogany genus_Lepidobotrys Canada_violet

N **ANIMAL** non-human, non-plant life cuckoo tapeworm carrier_pigeon Mycrosporidia virus tentacle egg

A few domain- and language-specific elaborations of the general guidelines are as follows:

Science chemicals, molecules, atoms, and subatomic particles are tagged as SUBSTANCE

Sports championships/tournaments are EVENTs

(Information) Technology Software names, kinds, and components are tagged as COMMUNICATION (e.g. **kernel**, **version**, **distribution**, **environment**). A **connection** is a RELATION; **project**, **support**, and a **configuration** are tagged as COGNITION; **development** and **collaboration** are ACTs.

Arabic conventions *Masdar* constructions (verbal nouns) are treated as nouns. Anaphora are not tagged.

APPENDIX E

Noun Supersense Annotation Guidelines

Supersense Tagging Guidelines

What should be tagged?

What counts as a noun?

For the current phase of annotation, we should be strict about only tagging things that (as a whole) serve as **nouns**. Though semantic categories like ATTRIBUTE (*modifiable*), LOCATION (*southwestern*, *underneath*), RELATION (*eleventh*), and TIME (*earlier*) may seem relevant to adjectives, adverbs, prepositions, or other parts of speech, worrying about those would make our lives too complicated.

Special cases:

- **Anaphora** (pronouns, etc.): if the supersense is clear in context—e.g. it has a clear nominal referent or obviously refers to a specific category (e.g. *someone* referring to a PERSON)—that supersense may be applied; leave blank otherwise (e.g. dummy *it*; *others* if too vague).
    - Never tag WH- or relative pronouns like *who* or *which*.
    - Never tag quantifiers in the gray area between determiners, adjectives, and pronouns: *some*, *all*, *much*, *several*, *many*, *most*, *few*, *none*, *each*, *every*, *enough*, *both*, *(n)either*, and generic senses of *one*. (These quantifiers often show up in partitives: *all/some/none of the X*, etc.)
    - For Arabic annotation we are not supersense-tagging ANY anaphora.
- **Verbal nouns/gerunds**
    - In Arabic, we have decided to tag *masdar* instances as nouns.
- **Mentions** of words (e.g., *The word "physics" means...*) should be tagged as COMMUNICATION because they are about the linguistic item.

Determining item boundaries

It is often difficult to determine which words should belong together as a unit (receiving a single supersense tag) vs. tagged separately. Some guidelines:

- Try to treat **proper names** as a unit. (Lack of capitalization makes this especially difficult for Arabic.)
    - Names of titles SHOULD be included if they appear as they might be used in addressing that person:
      President Obama
      United States *President Obama*
      Barack Obama, *president* of the *United States*
    - Honorific prefixes and suffixes should be included: *Dr. Fred Jelinek, Ph.D.*, *King Richard III*
- Other **multiword phrases** can be treated as a unit if they "go together strongly".
    - For example, *lexical semantics* is a standard term in linguistics and should therefore be considered a single unit. Note that *lexical* is not a noun, but it may be included as part of a term that overall functions as a noun.
    - Indications of whether an expression should be treated as a unit might include: conventionality (is it a particularly common way to refer to something?), predictability (if you had to guess how to express something, would you be likely to guess that phrase?), transparency (if you hadn't heard the whole expression before, would its meaning be clear from the individual words?), substitutability (could you replace a word with a similar word to get an equally normal expression

meaning the same thing?).
    - Consider: would you want to include the expression as a unit in a dictionary?

Vagueness and figurativity

Context and world knowledge should be used only to *disambiguate* the meaning of a word where it actually has multiple senses, not to refine it where it could refer to different things in context. For example, consider the sentences

(1) She felt a sense of shock at the outcome.
(2) She expressed her shock at the outcome.

The word 'shock' is ambiguous: as a technical term it could refer to a mechanical device, or to a medical state, but in the context of (1) and (2) it clearly has a sense corresponding to the FEELING tag.

You might notice that in (2) 'shock' is part of the content of a communication event. However, we do not want to say that 'shock' is ambiguous between an emotional state and something that is communicated; in (2) it is merely a feeling that happens to be communicated, while in (1) it is not communicated. Thus, we do *not* mark it as COMMUNICATION, because this meaning is not inherent to the word itself.

A similar problem arises with metaphor, metonymy, iconicity, and other figurative language. If a building is shaped like a pumpkin, given

(3) She lives in a pumpkin.

you might be tempted to mark 'pumpkin' as an ARTIFACT (because it is a building). But here 'pumpkin' is still referring to the normal sense of pumpkin—i.e. the PLANT—and from context you know that the typical appearance of a pumpkin plant is being used *in a novel (non-standard) way* to describe something that functions as a building. In other words, that buildings can be shaped like pumpkins is not something you would typically associate with the word 'pumpkin' (or, for that matter, any fruit). Similarly, in the sentence

(4) I gave her a toy lion.

'toy' should be tagged as ARTIFACT and 'lion' as ANIMAL (though it happens to be a nonliving depiction of an animal).

On the other hand, if it is highly conventional to use an expression figuratively, as in (5), we can decide that this figurative meaning has been lexicalized (given its own sense) and tag it as such:

(5) The White House said it would issue its decision on Monday.

According to WordNet, this use of 'White House' should be tagged as GROUP (not ARTIFACT) because it is a standard way to refer to the administration.

Highly idiomatic language should be tagged as if it were literal. For example, *road* in the phrase *road to success* should be tagged as ARTIFACT, even if it is being used metaphorically. Similarly, in an expression like

(6) behind the cloak of the Christian religion

(i.e., where someone is concealing their religious beliefs and masquerading as Christian), *cloak* should be tagged as an ARTIFACT despite being used nonliterally.

Supersense classification

Below are some examples of important words in specific domains, followed by a set of general-purpose rules.

Software domain

- pieces of software: COMMUNICATION
  - *version*, *distribution*
  - (software) *system*, *environment*
  - (operating system) *kernel*
- *connection*: RELATION
- *project*: COGNITION
- *support*: COGNITION
- *a configuration*: COGNITION
- *development*: ACT
- *collaboration*: ACT

Sports domain

- *championship*, *tournament*, etc.: EVENT

Science domain

- chemicals, molecules, atoms, and subatomic particles (*nucleus*, *electron*, *particle*, etc.): SUBSTANCE

Other special cases

- *world* should be decided based on context:
  - OBJECT if used like *Earth*/*planet*/*universe*
  - LOCATION if used as a place that something is located
  - GROUP if referring to humanity
  - (possibly other senses as well)
- someone's *life*:
  - TIME if referring to the time period (e.g. *during his life*)
  - STATE if referring to the person's (physical, cognitive, social, ...) existence
  - STATE if referring to the person's physical vitality/condition of being alive
  - (possibly others)
- *reason*: WordNet is kind of confusing here; I think we should say:
  - MOTIVE if referring to a (putative) cause of behavior (e.g. *reason for moving to Europe*)
  - COGNITION if referring to an understanding of what caused some phenomenon (e.g. *reason the sky is blue*)
  - COGNITION if referring to the abstract capacity for thought, or the philosophical notion of rationality
  - STATE if used to contrast reasonableness vs. unreasonableness (e.g. *within reason*)
  - [WordNet also includes COMMUNICATION senses for stated reasons, but I think this is splitting hairs. It makes more sense to contrast MOTIVE/COGNITION vs. COMMUNICATION for *explanation*, where communication seems more central to the lexical meaning. FrameNet seems

to agree with this: the Statement frame lists *explanation* but not *reason*.]

Decision list

This list attempts to make more explicit the semantic distinctions between the supersense classes for nouns. Follow the directions in order until an appropriate label is found.

1. If it is a **natural feature** (such as a mountain, valley, river, ocean, cave, continent, planet, the universe, the sky, etc.), label as OBJECT
2. If it is a **man-made structure** (such as a building, room, road, bridge, mine, stage, tent, etc.), label as ARTIFACT
   - includes venues for particular types of activities: *restaurant*, *concert hall*
   - *tomb* and *crypt* (structures) are ARTIFACTS, *cemetery* is a LOCATION
3. For **geopolitical entities** like cities and countries:
   - If it is a **proper name** that can be used to refer to a location, label as LOCATION
   - Otherwise, choose LOCATION or GROUP depending on which is the more salient meaning in context
4. If it describes a **shape** (in the abstract or of an object), label as SHAPE: *hexahedron*, *dip*, *convex shape*, *sine curve groove*, *lower bound*, *perimeter*
5. If it otherwise refers to an **space, area, or region** (not specifically requiring a man-made structure or describing a specific natural feature), label as LOCATION: *region*, *outside*, *interior*, *cemetery*, *airspace*
6. If it is a name of a **social group** (national/ethnic/religious/political) that can be made singular and used to refer to an individual, label as PERSON (*Arab*, *Muslim*, *American*, *communist*)
7. If it is a **social movement** (such as a religion, philosophy, or ideology, like *Islam* or *communism*), label as COGNITION if the belief system as a "set of ideas" sense is more salient in context (esp. for academic disciplines like *political science*), or as GROUP if the "set of adherents" is more salient
8. If it refers to an **organization or institution** (including companies, associations, teams, political parties, governmental divisions, etc.), label as GROUP: *U.S. State Department*, *University of California*, *New York Mets*
9. If it is a **common noun** referring to a **type or event of grouping** (e.g., *group*, *nation*, *people*, *meeting*, *flock*, *army*, *a collection*, *series*), label as GROUP
10. If it refers to something being used as **food or drink**, label as FOOD
11. If it refers to a **disease/disorder or physical symptom thereof**, label as STATE: *measles*, *rash*, *fever*, *tumor*, *cardiac arrest*, *plague* (= epidemic disease)
12. If it refers to **the human body or a natural part of the healthy body**, label as BODY: *ligament*, *fingernail*, *nervous system*, *insulin*, *gene*, *hairstyle*
13. If it refers to a **plant or fungus**, label as PLANT: *acorn squash*, *Honduras mahogany*, *genus Lepidobotrys*, *Canada violet*
14. If it refers to a **human or personified being**, label as PERSON: *Persian deity*, *mother*, *kibbutznik*, *firstborn*, *worshiper*, *Roosevelt*, *consumer*, *guardsman*, *glasscutter*, *appellant*
15. If it refers to **non-plant life**, label as ANIMAL: *lizard*, *bacteria*, *virus*, *tentacle*, *egg*
16. If it refers to a category of entity that pertains generically to **all life** (including both plants and animals), label as OTHER: *organism*, *cell*
17. If it refers to a prepared **drug** or health aid, label as ARTIFACT: *painkiller*, *antidepressant*, *ibuprofen*, *vaccine*, *cocaine*
18. If it refers to a **material or substance**, label as SUBSTANCE: *aluminum*, *steel* (= metal alloy), *sand*, *injection* (= solution that is injected), *cardboard*, *DNA*, *atom*, *hydrochloric acid*

19. If it is a term for an **entity that is involved in ownership or payment**, label as POSSESSION: *money*, *coin*, *a payment*, *a loan*, *a purchase* (= thing purchased), *debt* (= amount owed), one's *wealth/property* (= things one owns)
    - Does NOT include *acts* like *transfer*, *acquisition*, *sale*, *purchase*, etc.
20. If it refers to a **physical thing that is necessarily man-made**, label as ARTIFACT: *weapon*, *hat*, *cloth*, *cosmetics*, *perfume* (= scented cosmetic)
21. If it refers to a **nonliving object occurring in nature**, label as OBJECT: *barrier reef*, *nest*, *stepping stone*, *ember*
22. If it refers to a **temporal point, period, amount, or measurement**, label as TIME: *instant/moment*, *10 seconds*, *2011* (year), *2nd millenium BC*, *day*, *season*, *velocity*, *frequency*, *runtime*, *latency/delay*
    - Includes names of holidays: *Christmas*
    - *age* = 'period in history' is a TIME, but *age* = 'number of years something has existed' is an ATTRIBUTE
23. If it is a (non-temporal) **measurement or unit/type of measurement involving a relationship between two or more quantities**, including ordinal numbers not used as fractions, label as RELATION: *ratio*, *quotient*, *exponential function*, *transitivity*, *fortieth/40th*
24. If it is a (non-temporal) **measurement or unit/type of measurement**, including ordinal numbers and fractional amounts, label as QUANTITY: *7 centimeters*, *half*, *1.8 million*, *volume* (= spatial extent), *volt*, *real number*, *square root*, *decimal*, *digit*, *180 degrees*, *12 percent/12%*
25. If it refers to an **emotion**, label as FEELING: *indignation*, *joy*, *eagerness*
26. If it refers to an **abstract external force that causes someone to intend to do something**, label as MOTIVE: *reason*, *incentive*, *urge*, *conscience*
    - NOT *purpose*, *goal*, *intention*, *desire*, or *plan*
27. If it refers to a person's **belief/idea or mental state/process**, label as COGNITION: *knowledge*, *a dream*, *consciousness*, *puzzlement*, *skepticism*, *reasoning*, *logic*, *intuition*, *inspiration*, *muscle memory*, *theory*
28. If it refers to a **technique or ability**, including forms of perception, label as COGNITION: *a skill*, *aptitude/talent*, *a method*, *perception*, *visual perception/sight*, *sense of touch*, *awareness*
29. If it refers to an act of **information encoding/transmission** or the abstract information/work that is encoded/transmitted—including the use of language, writing, music, performance, print/visual/electronic media, or other form of signaling—label as COMMUNICATION: *a lie*, *a broadcast*, *a contract*, *a concert*, *a code*, *an alphabet*, *an equation*, *a denial*, *discussion*, *sarcasm*, *concerto*, *television program*, *software*, *input* (= signal)
    - Products or tools facilitating communication, such as books, paintings, photographs, or televisions, are themselves ARTIFACTS when used in the physical sense.
30. If it refers to a **learned profession** (in the context of practicing that profession), label as ACT: *engineering*, *law*, *medicine*, etc.
31. If it refers to a **field or branch of study** (in the sciences, humanities, etc.), label as COGNITION: *science*, *art history*, *nuclear engineering*, *medicine* (= medical science)
32. If it refers in the abstract to a **philosophical viewpoint**, label as COGNITION: *socialism*, *Marxism*, *democracy*
33. If it refers to a **physical force**, label as PHENOMENON: *gravity*, *electricity*, *pressure*, *suction*, *radiation*
34. If it refers to a **state of affairs**, i.e. a condition existing at a given point in time (with respect to some person/thing/situation), label as STATE: *poverty*, *infamy*, *opulence*, *hunger*, *opportunity*, *disease*, *darkness* (= lack of light)
    - heuristic: in English, can you say someone/something is "in (a state of) X" or "is full of X"?
    - let's exclude anything that can be an emotion [though WordNet also lists a STATE sense of *happiness* and *depression*]
    - easily confused with ATTRIBUTE and FEELING
35. If it refers to an **aspect/characteristic of something that can be judged** (especially nouns derived from adjectives), label as ATTRIBUTE: *faithfulness*, *clarity*, *temperature* (= degree of hotness), *valence*, *virtue*, *simplicity*, *darkness* (= dark coloring)
    - easily confused with STATE, FEELING, COGNITION
36. If it refers to the **relationship between two entities**, label as RELATION: *connection*, *marital relationship*, (non-person) *member*, (non-statistical) *correlation*, *antithesis*, *inverse*, *doctor-patient relationship*, *politics* (= social means of acquiring power), *causality*
37. If it refers to **"something that people do or cause to happen"**, label as ACT: *football game*, *acquisition* (= act of acquiring), *hiring*, *scoring*
    - Includes wars.
38. If it refers to **"something that happens at a given place and time"** label as EVENT: *tide*, *eclipse*, *accident*
    - Includes recurring events like sports tournaments.
39. If it refers to **"a sustained phenomenon or one marked by gradual changes through a series of states"** (esp. where the changes occur in a single direction), label as PROCESS: *evaporation*, *aging*, *extinction*, (economic) *inflation*, *accretion/growth*
40. If it refers to **something that happens/occurs**, label as PHENOMENON: *hurricane*, *tornado*, *cold front*, *effect*
41. If it is a synonym of ***kind/variety/type* (of something)**, label as COGNITION
42. If it is part of a **stock phrase used in discourse**, label as COMMUNICATION: for *example*, on the one *hand*, in the *case* of
43. If it is some other **abstract concept that can be known**, it should probably be labeled as COGNITION.

If you cannot decide based on these guidelines, use the "UNSURE" tag.

APPENDIX F

Verb Supersense Annotation Guidelines

Verb Supersense Tagging

Using WordNet as a guide, we should develop a tagging scheme for verbs along the lines of the one for nouns. (Verb tag names are lowercased to distinguish them from noun tags.)

`` `a `` **(auxiliary)**
: might/aux have/aux been/aux Ving

`` `j `` **(adjectival)**
: the written/adj message, a sinking/adj feeling

`body` **(grooming, dressing, bodily functions and care)**
: exercise = work out, cry (shed tears), wear (clothes), sweat, shiver, faint, burp, ache, tire, sleep, recuperate = convalesce, reproduce (biologically), die = cease to live [though WN puzzlingly has change], injure (physically)

`change` **(size, temperature change, intensifying, etc.)**
: grow (increase in size, age, or value), remove (physically), modify, revert, adjust, pop = burst

- includes verbs derived with -ify, -ize, -en, etc.: humidify, magnetize, strengthen

`cognition` **(thinking, judging, analyzing, doubting)**
: decide, think, rate (assign rating), respect = have respect for, memorize, learn, see = understand

contrast with `perception`, `communication`

`communication` **(verbal/linguistic or nonverbal gesturing: telling, asking, ordering)**
: speak, talk, write = communicate by writing, announce, type (on a keyboard), cry out, describe, argue, contest, petition, stammer, beg, mandate, veto, libel, preach, teach (education), fax, moo (animal noise)

- WN lists music production (a person singing/playing an instrument) as `creation`
- noises from inanimate objects (‘creak’, etc.) are `perception`
- contrast with `perception`, `cognition`

`competition` **(fighting, athletic activities)**
: compete, fight (with someone), play (sports), referee, duel [supersedes social?; superseded by communication for rhetorical senses of ‘attack’, ‘contend’, etc.; superseded by contact for moments of physical contact: ‘wrestle’, ‘box’, ‘punch’, ‘beat up’]

`consumption` **(ingesting, using, exploiting)**
: eat, picnic, thirst for (drink), digest (food), smoke (cigarette), use, waste [supersedes change and body?]

- *we* ***tasted*** *the food* (BUT: *the food* ***tastes*** *yummy* is `perception`)

`contact` **(touching, hitting, tying, digging)**
: fasten, overlay, slice, rub, pinch, box, punch, shoulder, yank, bump, release, lug, airlift, use/operate (an instrument or machine) [supersedes stative and motion, e.g. move something by vehicle or by carrying it]

`creation` **(sewing, baking, painting, performing)**
: create, bake (a cake), grow (agriculture), invent, write = produce a book, perform (give a performance)

`emotion` **(feeling)**
: fear, anger, hope = wish, trust [difficult to separate from social and communication, e.g. ‘amuse’, ‘encourage’]

- some lists of emotions: https://en.wikipedia.org/wiki/Contrasting_and_categorization_of_emotions

`motion` **(walking, flying, swimming)**
: travel, leap (physically), fly, vibrate, rotate (physically) [some synsets, e.g. for ‘reach’, conflate literal and metaphorical senses under motion (metaphorical should be change)?]

`perception` **(seeing, hearing, feeling)**
: see (visually), witness, feel (by touch), seem, thirst = feel thirsty, ache = feel physical pain, hallucinate, clang, twinkle (candle or bulb), creak (inanimate noise source)

- *the food* ***tastes*** *yummy* (BUT: *we* ***tasted*** *the food* is `consumption`)
- contrast with `communication`, `cognition`

`possession` **(buying, selling, owning)**
: receive = acquire, lend, sell, purchase, rob, want/ask=charge (an amount for), have = own, possess (a piece of property)

`social` **(social activities and events: law, politics, economy, education, family, religion, etc.)**
: meet (socially), celebrate, divorce, succeed (achieve success, or be successor to), respect = show respect towards, gerrymander, cheat (except in the competitive sense), spoil = mollycoddle

`stative` **(being, having, spatial relations)**
: be, have = feature, stagnate, equal, necessitate, lack, span, contain (physically), underlie [superseded by contact? encircle/surround, cover = serve as cover for]

- *what* ***separates*** *these things is...*

- *X **compared** to Y*
- active subject usually a theme??

`weather` (raining, snowing, thawing, thundering)
: rain, thunder, twinkle (star), warm up (climate)

Differentiation

`perception`, `communication`, `cognition`

- `communication` if there is necessarily a transfer of information from one party to another
  - ***showing** me various options*
- `perception` if the focus is on taking in information by one party (can often substitute "seem" or "feel")
  - *they **seemed** more interested*
  - *it **sounded** like it's done every day*
  - *it will **look** beautiful*
- `cognition` if about mentally processing information
  - ***rate** as five stars* (no communication necessarily implied)
  - *difficult to **see** what you're paying for* (= understand)

Precedence relations

- { `perception`, `consumption` } > `body` > `change`
- `motion` > `social` > `change`
- `emotion` > `change`
- `motion` > { `body`, `possession` } (e.g., stand_up, bring)
- `contact` > { `stative`, `motion` }
- { `contact`, `communication` } > `competition` > `social`
- `emotion` > `cognition`

Groupings

- **GENERAL**
  - motion, change, creation
  - contact

  - stative
  - weather
- **PERSONS**
  - body, consumption
  - perception
  - cognition
    - emotion
  - social
    - possession
    - communication
    - competition

Specific decisions

- fall_ill, fall_asleep, die, injure, give_birth: `body`
- fall_in_love: `emotion`

12/4

- have_been_to - as in visiting a location: `motion`
- cleft and existential 'be': `stative`
- give/receive a tattoo: `possession`; give/take a lesson: `social`

12/10

- wait: `cognition` if emphasis on expectation, `stative` if emphasis on not acting

12/12

- be_the_epitome_of or the_epitome_of?
- feel: `cognition` if having an opinion, `emotion` if experiencing a feeling, `perception` if experiencing a physical stimulus

12/19

- deal_with = come to grips with: `emotion`; deal_with = interact_with somebody: `social`
- found something to be adj: `cognition`

- require/take/let/permit with an inanimate subject (i.e., it took 1 hour) `stative`, but when the subject and object are both people it is `social`
- pass time: ??
- save/keep/prevent from: `stative`
- like/love/hate etc. when conveying an opinion: `emotion`
- get/receive/give~service: `social`
- bother_with meaning go to the trouble to do something: `social`
- We need to revisit *die* and *fall ill* (`body` vs. `change`)
- have_pedicure, have_ hair _done: `body`
  - We need to revisit *give/receive tattoo*, which is currently `possession`
- any commercial act (i.e., using their services): `possession`
- We need to revisit MWEs with *be*
- have_experience: `stative`

1/2

- revision of `body` to take precedence over `change` (for changes in health, etc.)
- use/operate an instrument/machine: `contact`

1/14

Several changes above, new section contrasting `perception` / `communication` / `cognition`, as well as:

- recommend, itemized (receipt), get_in_touch_with (a manager): `communication`
- beware of: `cognition`
- avoid:
  - (purposefully) **avoid** restaurant at all costs: `cognition`
  - something was **avoided** = it fortunately failed to happen: `change`
- `social`:
  - sabotaging
  - went_above_and_beyond
  - bait_and_switching
  - **helped_** me **_through** difficult times
  - she **makes_** you **_feel**
  - worst service I ever **experienced**
- arrive/got=made_it/leave/brought_in (= carried): `motion`
- got_upgraded_to (a corner suite): `change`
- sports **related**, **including** (parts): `stative`

1/15

- grab a cab: `possession`
- hard to find (restaurant): `cognition`
- do_research about a restaurant: `social`
- have_issues/problems: `cognition`
- use (someone's services): `possession`
- pay_attention: `perception`

1/16

- order a sandwich: `communication` (though it is part of a larger scenario of dining that might be `possession` or `social`)

Remarks

When tagging in context, note that (unless it is an MWE) only the meaning of the verb should be characterized. So the verb 'rising' in 'temperatures are rising' should be tagged as a change verb, NOT as a weather verb.

Inherently negative verbs like 'avoid', 'remove', and 'refuse' receive a sense depending on the sort of activity that is not performed.

Issues:

- modals, quasi-modals, count as auxes: be_supposed_to, be_going_to, do not V
- all uses of 'do', other weakly collocated light verbs not contributing their own meaning??
- have warning about s.t.: perception?

Bibliography

Anne Abeillé and Lionel Clément. Annotation morpho-syntaxique, January 2003. URL `http://www.llf.cnrs.fr/Gens/Abeille/guide-morpho-synt.02.pdf`.

Anne Abeillé, Lionel Clément, and François Toussenel. Building a treebank for French. In Anne Abeillé and Nancy Ide, editors, *Treebanks*, volume 20 of *Text, Speech and Language Technology*, pages 165–187. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2003.

Anne Abeillé, François Toussenel, and Martine Chéradame. Corpus le Monde. Annotations en constituants: Guide pour les correcteurs, March 2004. URL `http://www.llf.cnrs.fr/Gens/Abeille/guide-annot.pdf`.

Omri Abend and Ari Rappoport. Fully unsupervised core-adjunct argument classification. In *Proc. of ACL*, pages 226–236, Uppsala, Sweden, July 2010.

Otavio Acosta, Aline Villavicencio, and Viviane Moreira. Identification and treatment of multiword expressions applied to information retrieval. In *Proc. of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 101–109, Portland, Oregon, USA, June 2011.

Eneko Agirre and Philip Edmonds, editors. *Word Sense Disambiguation: Algorithms and Applications*, volume 33 of *Text, Speech and Language Technology*. Springer, Dordrecht, The Netherlands, 2006.

Eneko Agirre, Timothy Baldwin, and David Martinez. Improving parsing and PP attachment performance with sense information. In *Proc. of ACL-HLT*, pages 317–325, Columbus, Ohio, USA, June 2008.

Eneko Agirre, Kepa Bengoetxea, Koldo Gojenola, and Joakim Nivre. Improving dependency parsing with semantic classes. In *Proc. of ACL-HLT*, pages 699–703, Portland, Oregon, USA, June 2011.

Eneko Agirre, Nora Aranberri, Ander Barrena, António Branco, Martin Popel, Aljoscha Burchardt, Gorka Labaka, Petya Osenova, Kepa Sarasola, João Silva, and Gertjan van Noord. Report on the state of the art of named entity and word sense disambiguation. Technical Report QTLEAP-2013-D5.1, December 2013. URL http://qtleap.eu/wp-content/uploads/2014/04/QTLEAP-2013-D5.1.pdf. Version 3.0.

Hassan Al-Haj and Shuly Wintner. Identifying multi-word expressions by leveraging morphological and syntactic idiosyncrasy. In *Proc. of Coling*, pages 10–18, Beijing, China, August 2010.

Ash Asudeh and Richard Crouch. Glue semantics for HPSG. In Frank van Eynde, Lars Hellan, and Dorothee Beermann, editors, *Proc. of the 8th International HPSG Conference*, volume 3, page 5, Trondheim, Norway, August 2001.

Giuseppe Attardi, Stefano Dei Rossi, Giulia Di Pietro, Alessandro Lenci, Simonetta Montemagni, and Maria Simi. A resource and tool for supersense tagging of Italian texts. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proc. of LREC*, pages 2242–2248, Valletta, Malta, May 2010.

Giuseppe Attardi, Luca Baronti, Stefano Dei Rossi, and Maria Simi. SuperSense Tagging with a Maximum Entropy Markov Model. In Bernardo Magnini, Francesco Cutugno, Mauro Falcone, and Emanuele Pianta, editors, *Evaluation of Natural Language and Speech Tools for Italian*, number 7689 in Lecture Notes in Computer Science, pages 186–194. Springer, Berlin, January 2013.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The Berkeley FrameNet project. In *Proc. of COLING-ACL*, pages 86–90, Montreal, Quebec, Canada, August 1998.

Timothy Baldwin. Compositionality and multiword expressions: six of one, half a dozen of the other? Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties, July 2006. URL http://ww2.cs.mu.oz.au/~tim/pubs/colacl2006-mwe-invited.pdf.

Timothy Baldwin. A resource for evaluating the deep lexical acquisition of English verb-particle constructions. In *Proc. of MWE*, pages 1–2, Marrakech, Morocco, June 2008.

Timothy Baldwin and Su Nam Kim. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*, pages 267–292. CRC Press, Taylor and Francis Group, Boca Raton, Florida, USA, 2010.

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. An empirical model of multiword expression decomposability. In *Proc. of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 89–96, Sapporo, Japan, July 2003.

Timothy Baldwin, John Beavers, Leonoor van der Beek, Francis Bond, Dan Flickinger, and Ivan A. Sag. In search of a systematic treatment of determinerless PPs. In Patrick Saint-Dizier and Nancy Ide, editors, *Syntax and Semantics of Prepositions*, volume 29 of *Text, Speech and Language Technology*, pages 163–179. Springer, Dordrecht, The Netherlands, 2006.

Timothy Baldwin, Valia Kordoni, and Aline Villavicencio. Prepositions in applications: a survey and introduction to the special issue. *Computational Linguistics*, 35(2):119–149, 2009.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. Abstract Meaning Representation for sembanking. In

Proc. of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, pages 178–186, Sofia, Bulgaria, August 2013.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. Abstract Meaning Representation (AMR) 1.2 Specification, May 2014. URL https://github.com/amrisi/amr-guidelines/blob/b0fd2d6321ed4c9e9fa202b307cceeae36b8c25b/amr.md.

Colin Bannard and Elena Lieven. Formulaic language in L1 acquisition. *Annual Review of Applied Linguistics*, 32:3–16, 2012.

Colin Bannard, Timothy Baldwin, and Alex Lascarides. A statistical approach to the semantics of verb-particles. In *Proc. of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 65–72, Sapporo, Japan, July 2003.

Anabela Barreiro, Johanna Monti, Brigitte Orliac, and Fernando Batista. When multiwords go bad in machine translation. In Johanna Monti, Ruslan Mitkov, Gloria Corpas Pastor, and Violeta Seretan, editors, *Proc. of the MT Summit 2013 Workshop on Multi-word Units in Machine Translation and Translation Technologies*, pages 26–33, Nice, France, September 2013.

Anabela Barreiro, Johanna Monti, Brigitte Orliac, Susanne Preuß, Kutz Arrieta, Wang Ling, Fernando Batista, and Isabel Trancoso. Linguistic evaluation of support verb constructions by OpenLogos and Google Translate. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proc. of LREC*, pages 35–40, Reykjavík, Iceland, May 2014.

Eduard Bejček and Pavel Straňák. Annotation of multiword expressions in the Prague dependency treebank. *Language Resources and Evaluation*, 44(1):7–21, 2010.

Eduard Bejček, Pavel Straňák, and Pavel Pecina. Syntactic identification of occurrences of multiword expressions in text using a lexicon with dependency structures. In *Proc. of the 9th Workshop on Multiword Expressions*, pages 106–115, Atlanta, Georgia, USA, June 2013.

Kepa Bengoetxea, Eneko Agirre, Joakim Nivre, Yue Zhang, and Koldo Gojenola. On WordNet semantic classes and dependency parsing. In *Proc. of ACL*, pages 649–655, Baltimore, Maryland, USA, June 2014.

Gábor Berend. Opinion expression mining by exploiting keyphrase extraction. In *Proc. of 5th International Joint Conference on Natural Language Processing*, pages 1162–1170, Chiang Mai, Thailand, November 2011.

Benjamin K. Bergen and Nancy Chang. Embodied Construction Grammar in simulation-based language understanding. In Jan-Ola Östman and Mirjam Fried, editors, *Construction grammars: cognitive grounding and theoretical extensions*, pages 147–190. John Benjamins, Amsterdam, 2005.

Archna Bhatia, Chu-Cheng Lin, Nathan Schneider, Yulia Tsvetkov, Fatima Talib Al-Raisi, Laleh Roostapour, Jordan Bender, Abhimanu Kumar, Lori Levin, Mandy Simons, and Chris Dyer. Automatic classification of communicative functions of definiteness. In *Proc. of COLING*, pages 1059–1070, Dublin, Ireland, August 2014.

Ann Bies, Justin Mott, Colin Warner, and Seth Kulick. English Web Treebank. Technical Report LDC2012T13, Linguistic Data Consortium, Philadelphia, PA, 2012. URL http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2012T13.

Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit.* O'Reilly Media, Inc., Sebastopol, California, USA, June 2009.

Phil Blunsom and Timothy Baldwin. Multilingual deep lexical acquisition for HPSGs via supertagging. In *Proc. of EMNLP*, pages 164–171, Sydney, Australia, July 2006.

Dwight Le Merton Bolinger. *The phrasal verb in English.* Harvard University Press, Cambridge, MA, 1971.

Claire Bonial, William Corvey, Martha Palmer, Volha V. Petukhova, and Harry Bunt. A hierarchical unification of LIRICS and VerbNet semantic roles. In *Fifth IEEE International Conference on Semantic Computing*, pages 483–489, Palo Alto, CA, USA, September 2011.

Claire Bonial, Julia Bonn, Kathryn Conger, Jena D. Hwang, and Martha Palmer. PropBank: semantics of new predicate types. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proc. of LREC*, pages 3013–3019, Reykjavík, Iceland, May 2014a.

Claire Bonial, Meredith Green, Jenette Preciado, and Martha Palmer. An approach to *take* multi-word expressions. In *Proc. of the 10th Workshop on Multiword Expressions*, pages 94–98, Gothenburg, Sweden, April 2014b.

Johan Bos, Stephen Clark, Mark Steedman, James R. Curran, and Julia Hockenmaier. Wide-coverage semantic representations from a CCG parser. In *Proc. of Coling*, pages 1240–1246, Geneva, Switzerland, August 2004.

Ram Boukobza and Ari Rappoport. Multi-word expression identification using sentence surface features. In *Proc. of EMNLP*, pages 468–477, Singapore, August 2009.

Melissa Bowerman and Soonja Choi. Shaping meanings for language: universal and language-specific in the acquisition of spatial semantic categories. In Melissa Bowerman and Stephen Levinson, editors, *Language Acquisition and Conceptual Development*, number 3 in Language, Culture & Cognition, pages 475–511. Cambridge University Press, Cambridge, UK, January 2001.

S.R.K. Branavan, Luke Zettlemoyer, and Regina Barzilay. Reading between the lines: learning to map high-level instructions to commands. In *Proc. of ACL*, pages 1268–1277, Uppsala, Sweden, July 2010.

Julian Brooke, Vivian Tsang, Graeme Hirst, and Fraser Shein. Unsupervised multiword segmentation of large corpora using prediction-driven decomposition of n-grams. In *Proc. of COLING*, pages 753–761, Dublin, Ireland, August 2014.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, December 1992.

Claudia Brugman. *The story of 'over': polysemy, semantics and the structure of the lexicon.* MA thesis, University of California, Berkeley, Berkeley, CA, 1981. Published New York: Garland, 1981.

Clara Cabezas and Philip Resnik. Using WSD techniques for lexical selection in statistical machine translation. Technical Report CS-TR-4736, University of Maryland, College Park, Maryland, USA, July 2005. URL http://www.dtic.mil/cgi-bin/GetTRDoc?Location=U2&doc=GetTRDoc.pdf&AD=ADA453538.

Nicoletta Calzolari, Charles J. Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod, and Antonio Zampolli. Towards best practice for multiword expressions in computational lexicons. In *Proc. of LREC*, pages 1934–1940, Las Palmas, Canary Islands, May 2002.

Marie Candito and Matthieu Constant. Strategies for contiguous multiword expression analysis and dependency parsing. In *Proc. of ACL*, pages 743–753, Baltimore, Maryland, USA, June 2014.

Jaime G. Carbonell. POLITICS: automated ideological reasoning. *Cognitive Science*, 2(1):27–51, January 1978.

Marine Carpuat and Mona Diab. Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In *Proc. of NAACL-HLT*, pages 242–245, Los Angeles, California, USA, June 2010.

Marine Carpuat and Dekai Wu. Word sense disambiguation vs. statistical machine translation. In *Proc. of ACL*, pages 387–394, Ann Arbor, Michigan, June 2005.

Marine Carpuat and Dekai Wu. Improving statistical machine translation using word sense disambiguation. In *Proc. of EMNLP-CoNLL*, pages 61–72, Prague, Czech Republic, June 2007.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. Word sense disambiguation improves statistical machine translation. In *Proc. of ACL*, pages 33–40, Prague, Czech Republic, June 2007.

Martin Chodorow, Joel Tetreault, and Na-Rae Han. Detection of grammatical errors involving prepositions. In *Proc. of the Fourth ACL-SIGSEM Workshop on Prepositions*, pages 25–30, Prague, Czech Republic, June 2007.

Kenneth Church, William Gale, Patrick Hanks, and Donald Hindle. Using statistics in lexical analysis. In Uri Zernik, editor, *Lexical acquisition: exploiting on-line resources to build a lexicon*, pages 115–164. Lawrence Erlbaum Associates, Hillsdale, New Jersey, USA, 1991.

Massimiliano Ciaramita and Yasemin Altun. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proc. of EMNLP*, pages 594–602, Sydney, Australia, July 2006.

Massimiliano Ciaramita and Mark Johnson. Supersense tagging of unknown nouns in WordNet. In *Proc. of EMNLP*, pages 168–175, Sapporo, Japan, July 2003.

Silvie Cinková, Jan Hajič, Marie Mikulová, Lucie Mladová, Anja Nedolužko, Petr Pajas, Jarmila Panevová, Jiří Semecký, Jana Šindlerová, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. Annotation of English on the tectogrammatical level: reference book. Technical report, Charles University, Prague, 2006. URL http://ufal.mff.cuni.cz/pcedt2.0/publications/TR_En.pdf.

Martin Čmejrek, Jan Cuřín, Jan Hajič, and Jiří Havelka. Prague Czech-English Dependency Treebank: resource for structure-based MT. In *Proc. of EAMT*, pages 73–78, Budapest, Hungary, May 2005.

Michael Collins. Discriminative training methods for Hidden Markov Models: theory and experiments with perceptron algorithms. In *Proc. of EMNLP*, pages 1–8, Philadelphia, Pennsylvania, USA, July 2002.

Matthieu Constant and Anthony Sigogne. MWU-aware part-of-speech tagging with a CRF model and lexical resources. In *Proc. of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 49–56, Portland, Oregon, USA, June 2011.

Matthieu Constant, Anthony Sigogne, and Patrick Watrin. Discriminative strategies to integrate multiword expression recognition and parsing. In *Proc. of ACL*, pages 204–212, Jeju Island, Korea, July 2012.

Paul Cook and Suzanne Stevenson. Classifying particle semantics in English verb-particle constructions. In *Proc. of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 45–53, Sydney, Australia, July 2006.

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A. Sag. Minimal Recursion Semantics: an introduction. *Research on Language and Computation*, 3(2-3):281–332, July 2005.

Gregory F. Coppola, Alexandra Birch, Tejaswini Deoskar, and Mark Steedman. Simple semi-supervised learning for prepositional phrase attachment. In *Proc. of IWPT*, pages 129–139, Dublin, Ireland, October 2011.

William Croft. *Radical Construction Grammar: syntactic theory in typological perspective.* Oxford University Press, Oxford, UK, 2001.

William Croft. *Verbs: Aspect and Causal Structure.* Oxford University Press, Oxford, UK, March 2012.

Hubert Cuyckens and Günter Radden, editors. *Perspectives on prepositions.* Max Niemeyer Verlag, Tübingen, Germany, 2002.

Daniel Dahlmeier, Hwee Tou Ng, and Tanja Schultz. Joint learning of preposition senses and semantic roles of prepositional phrases. In *Proc. of EMNLP*, pages 450–458, Singapore, August 2009.

Dipanjan Das, Nathan Schneider, Desai Chen, and Noah A. Smith. Probabilistic frame-semantic parsing. In *Proc. of NAACL-HLT*, pages 948–956, Los Angeles, California, USA, June 2010.

Dipanjan Das, Desai Chen, André F. T. Martins, Nathan Schneider, and Noah A. Smith. Frame-semantic parsing. *Computational Linguistics*, 40 (1):9–56, March 2014.

Hal Daumé, III. *Practical structured learning techniques for natural language processing.* Ph.D. dissertation, University of Southern California, Los Angeles, California, USA, 2006. URL `http://hal3.name/docs/daume06thesis.pdf`.

Miryam de Lhoneux. *CCG Parsing and Multiword Expressions.* MS thesis, University of Edinburgh, Edinburgh, Scotland, UK, August 2014.

Paul D. Deane. Multimodal spatial representation: On the semantic unity of *over*. In *From Perception to Meaning: Image Schemas in Cognitive Linguistics*, number 29 in Cognitive Linguistics Research, pages 235–282. Mouton de Gruyter, Berlin, 2005.

Nicolas Denand and Monique Rolbert. Contextual processing of locative prepositional phrases. In *Proc. of Coling*, pages 1353–1359, Geneva, Switzerland, August 2004.

Robert B. Dewell. *Over* again: Image-schema transformations in semantic analysis. *Cognitive Linguistics*, 5(4):351–380, January 1994.

Mona Diab and Pravin Bhutada. Verb noun construction MWE token classification. In *Proc. of MWE*, pages 17–22, Suntec, Singapore, August 2009.

Mona Diab and Madhav Krishna. Unsupervised classification of verb noun multi-word expression tokens. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 5449 of *Lecture Notes in Computer Science*, pages 98–110. Springer, Berlin, 2009.

Arantza Díaz de Ilarraza, Koldo Gojenola, and Maite Oronoz. Detecting erroneous uses of complex postpositions in an agglutinative language. In *Coling 2008: Posters and Demonstrations*, pages 31–34, Manchester, UK, August 2008.

Simon Dobnik and John D. Kelleher. Towards an automatic identification of functional and geometric spatial prepositions. In *Proc. of PRE-CogSci*, Berlin, July 2013.

Simon Dobnik and John D. Kelleher. Exploration of functional semantics of prepositions from corpora of descriptions of visual scenes. In *Proc. of the Workshop on Vision and Language*, pages 33–37, Dublin, Ireland, August 2014.

Florian Dömges, Tibor Kiss, Antje Müller, and Claudia Roch. Measuring the productivity of determinerless PPs. In *Proc. of the Fourth ACL-SIGSEM Workshop on Prepositions*, pages 31–37, Prague, Czech Republic, June 2007.

Sabri Elkateb, William Black, Horacio Rodríguez, Musa Alkhalifa, Piek Vossen, Adam Pease, and Christiane Fellbaum. Building a WordNet for Arabic. In *Proc. of LREC*, pages 29–34, Genoa, Italy, May 2006.

Nick C. Ellis, Rita Simpson-Vlach, and Carson Maynard. Formulaic language in native and second language speakers: psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly*, 42(3):375–396, 2008.

Rod Ellis. *The study of second language acquisition.* Oxford University Press, Oxford, 2nd edition, 2008.

Vyvyan Evans and Melanie Green. *Cognitive linguistics: an introduction.* Edinburgh University Press, Edinburgh, Scotland, UK, March 2006.

Afsaneh Fazly, Suzanne Stevenson, and Ryan North. Automatically learning semantic knowledge about multiword predicates. *Language Resources and Evaluation*, 41(1):61–89, 2007.

Susanne Feigenbaum and Dennis Kurzon, editors. *Prepositions in their syntactic, semantic, and pragmatic context*. Number 50 in Typological Studies in Language. John Benjamins, Amsterdam, 2002.

Jerome Feldman, Ellen Dodge, and John Bryant. Embodied Construction Grammar. In Bernd Heine and Heiko Narrog, editors, *The Oxford Handbook of Linguistic Analysis*, pages 111–138. Oxford University Press, Oxford, December 2009.

Christiane Fellbaum. English verbs as a semantic net. *International Journal of Lexicography*, 3(4):278–301, December 1990.

Christiane Fellbaum, editor. *WordNet: an electronic lexical database*. MIT Press, Cambridge, Massachusetts, USA, 1998.

David A. Ferrucci. Introduction to “This is Watson”. *IBM Journal of Research and Development*, 56(3.4):1:1–1:15, June 2012.

Charles J. Fillmore. Frame Semantics. In *Linguistics in the Morning Calm*, pages 111–137. Hanshin Publishing Co., Seoul, South Korea, 1982.

Charles J. Fillmore. Frames and the semantics of understanding. *Quaderni di Semantica*, 6(2):222–254, 1985.

Charles J. Fillmore, Paul Kay, and Mary Catherine O’Connor. Regularity and idiomaticity in grammatical constructions: the case of ‘let alone’. *Language*, 64(3):501–538, September 1988.

Jeffrey Flanigan, Sam Thomson, Jaime Carbonell, Chris Dyer, and Noah A. Smith. A discriminative graph-based parser for the Abstract Meaning Representation. In *Proc. of ACL*, pages 1426–1436, Baltimore, Maryland, USA, June 2014.

Richard Fothergill and Timothy Baldwin. Combining resources for MWE-token classification. In *Proc. of *SEM*, pages 100–104, Montréal, Canada, June 2012.

Yoav Freund and Robert E. Schapire. Large margin classification using the perceptron algorithm. *Machine Learning*, 37(3):277–296, December 1999.

Annemarie Friedrich and Alexis Palmer. Automatic prediction of aspectual class of verbs in context. In *Proc. of ACL*, pages 517–523, Baltimore, Maryland, USA, June 2014.

Sanae Fujita, Francis Bond, Stephan Oepen, and Takaaki Tanaka. Exploiting semantic information for HPSG parse selection. *Research on Language and Computation*, 8(1):1–22, March 2010.

Mahmoud Ghoneim and Mona Diab. Multiword expressions in the context of statistical machine translation. In *Proc. of IJCNLP*, pages 1181–1187, Nagoya, Japan, October 2013.

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

Kevin Gimpel and Noah A. Smith. Rich source-side context for statistical machine translation. In *Proc. of WMT*, pages 9–17, Columbus, Ohio, USA, June 2008.

Kevin Gimpel and Noah A. Smith. Generative models of monolingual and bilingual gappy patterns. In *Proc. of WMT*, pages 512–522, Edinburgh, Scotland, UK, July 2011.

Adele E. Goldberg. *Constructions at work: the nature of generalization in language*. Oxford University Press, Oxford, UK, 2006.

Adele E. Goldberg. *Constructions: a construction grammar approach to argument structure*. University of Chicago Press, Chicago, Illinois, USA, 1995.

Laura Gonnerman and Mary-Jane Blais. L2 processing of English phrasal verbs. The 31st Second Language Research Forum, October 2012.

Edouard Grave, Guillaume Obozinski, and Francis Bach. Hidden Markov tree models for semantic class induction. In *Proc. of CoNLL*, pages 94–103, Sofia, Bulgaria, August 2013.

Spence Green, Marie-Catherine de Marneffe, John Bauer, and Christopher D. Manning. Multiword expression identification with tree substitution grammars: a parsing *tour de force* with French. In *Proc. of EMNLP*, pages 725–735, Edinburgh, Scotland, UK, July 2011.

Spence Green, Marie-Catherine de Marneffe, and Christopher D. Manning. Parsing models for identifying multiword expressions. *Computational Linguistics*, 39(1):195–227, November 2012.

Sidney Greenbaum. *The Oxford English Grammar*. Oxford University Press, Oxford, UK, 1996.

Clayton Greenberg. Disambiguating prepositional phrase attachment sites with sense information captured in contextualized distributional data. In *Proc. of the ACL 2014 Student Research Workshop*, pages 71–77, Baltimore, Maryland, USA, June 2014.

Stefan Th. Gries. Phraseology and linguistic theory: a brief survey. In Sylviane Granger and Fanny Meunier, editors, *Phraseology: an interdisciplinary perspective*, pages 3–25. John Benjamins, Amsterdam, 2008.

Cyril Grouin, Sophie Rosset, Pierre Zweigenbaum, Karën Fort, Olivier Galibert, and Ludovic Quintard. Proposal for an extension of traditional named entities: from guidelines to evaluation, an overview. In *Proc. of the 5th Linguistic Annotation Workshop*, pages 92–100, Portland, Oregon, USA, June 2011.

Ebba Gustavii. Target language preposition selection – an experiment with transformation-based learning and aligned bilingual data. In *Proc. of EAMT*, pages 112–118, Budapest, Hungary, May 2005.

Claude Hagège. *Adpositions*. Oxford University Press, Oxford, UK, December 2009.

Jan Hajič. Building a syntactically annotated corpus: the Prague Dependency Treebank. In Eva Hajičová, editor, *Issues of Valency and Meaning. Studies in Honor of Jarmila Panevová*, pages 12–19. Prague Karolinum, Charles University Press, Prague, 1998.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. Prague Czech-English Dependency Treebank 2.0. Technical Report LDC2012T08, Linguistic Data Consortium, Philadelphia, PA, June 2012. URL `http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2012T08`.

Silvana Hartmann, György Szarvas, and Iryna Gurevych. Mining multiword terms from Wikipedia. In Maria Teresa Pazienza and Armando Stellato, editors, *Semi-Automatic Ontology Development*, pages 226–258. IGI Global, Hershey, PA, February 2012.

Chikara Hashimoto and Daisuke Kawahara. Construction of an idiom corpus and its application to idiom identification based on WSD incorporating idiom-specific features. In *Proc. of EMNLP*, pages 992–1001, Honolulu, Hawaii, USA, October 2008.

Eva Hasler, Barry Haddow, and Philipp Koehn. Dynamic topic adaptation for SMT using distributional profiles. In *Proc. of the Ninth Workshop on Statistical Machine Translation*, pages 445–456, Baltimore, Maryland, USA, June 2014.

Karl Moritz Hermann, Phil Blunsom, and Stephen Pulman. An unsupervised ranking model for noun-noun compositionality. In *Proc. of *SEM*, pages 132–141, Montréal, Canada, June 2012a.

Karl Moritz Hermann, Chris Dyer, Phil Blunsom, and Stephen Pulman. Learning semantics and selectional preference of adjective-noun pairs. In *Proc. of *SEM*, pages 70–74, Montréal, Canada, June 2012b.

Matthieu Hermet and Désilets Alain. Using first and second language models to correct preposition errors in second language authoring. In *Proc. of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 64–72, Boulder, Colorado, June 2009.

Ulf Hermjakob, Kevin Knight, and Hal Daumé III. Name translation in statistical machine translation - learning when to transliterate. In *Proc. of ACL-HLT*, pages 389–397, Columbus, Ohio, USA, June 2008.

Donald Hindle and Mats Rooth. Structural ambiguity and lexical relations. *Computational Linguistics*, 19(1):103–120, 1993.

Thomas Hoffmann and Graeme Trousdale, editors. *The Oxford Handbook of Construction Grammar*. Oxford University Press, Oxford, UK, 2013.

Munpyo Hong, Chang-Hyun Kim, and Sang-Kyu Park. Treating unknown light verb construction in Korean-to-English patent MT. In Tapio Salakoski, Filip Ginter, Sampo Pyysalo, and Tapio Pahikkala, editors, *Advances in Natural Language Processing*, volume 4139 of *Lecture Notes in Computer Science*, pages 726–737. Springer, Berlin, 2006.

Dirk Hovy, Stephen Tratz, and Eduard Hovy. What's in a preposition? Dimensions of sense disambiguation for an interesting word class. In *Coling 2010: Posters*, pages 454–462, Beijing, China, August 2010.

Dirk Hovy, Ashish Vaswani, Stephen Tratz, David Chiang, and Eduard Hovy. Models and training for unsupervised preposition sense disambiguation. In *Proc. of ACL-HLT*, pages 323–328, Portland, Oregon, USA, June 2011.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. OntoNotes: the 90% solution. In *Proc. of HLT-NAACL*, pages 57–60, New York City, USA, June 2006.

Ruihong Huang and Ellen Riloff. Inducing domain-specific semantic class taggers from (almost) nothing. In *Proc. of ACL*, pages 275–285, Uppsala, Sweden, July 2010.

Rodney Huddleston. The clause: complements. In Rodney Huddleston and Geoffrey K. Pullum, editors, *The Cambridge Grammar of the English Language*, pages 213–321. Cambridge University Press, Cambridge, UK, 2002.

Jena D. Hwang. Making verb argument adjunct distinctions in English. Synthesis paper, University of Colorado, Boulder, Colorado, November 2011. URL `http://verbs.colorado.edu/~hwangd/docs/synthesis-jena_d_hwang.pdf`.

Jena D. Hwang. *Identification and representation of caused motion constructions*. Ph.D. dissertation, University of Colorado, Boulder, Colorado, 2014.

Jena D. Hwang, Archna Bhatia, Claire Bonial, Aous Mansouri, Ashwini Vaidya, Nianwen Xue, and Martha Palmer. PropBank annotation of multilingual light verb constructions. In *Proc. of the Fourth Linguistic Annotation Workshop*, pages 82–90, Uppsala, Sweden, July 2010a.

Jena D. Hwang, Rodney D. Nielsen, and Martha Palmer. Towards a domain independent semantics: enhancing semantic representation with construction grammar. In *Proc. of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics*, pages 1–8, Los Angeles, California, June 2010b.

Jena D. Hwang, Annie Zaenen, and Martha Palmer. Criteria for identifying and annotating caused motion constructions in corpus data. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proc. of LREC*, pages 1297–1304, Reykjavík, Iceland, May 2014.

Christian Hying. A corpus-based analysis of geometric constraints on projective prepositions. In *Proc. of the Fourth ACL-SIGSEM Workshop on Prepositions*, pages 1–8, Prague, Czech Republic, June 2007.

Rubén Izquierdo, Sonia Vázquez, and Andrés Montoyo. Semantic classes and relevant domains on WSD. In Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala, editors, *Text, Speech and Dialogue*, number 8655 in Lecture Notes in Computer Science, pages 166–172. Springer International Publishing, Cham, Switzerland, January 2014.

Anders Johannsen, Dirk Hovy, Héctor Martínez Alonso, Barbara Plank, and Anders Søgaard. More or less supervised supersense tagging of Twitter. In *Proc. of *SEM*, pages 1–11, Dublin, Ireland, August 2014.

Mark Johnson. *The body in the mind: the bodily basis of meaning, imagination, and reason.* University of Chicago Press, Chicago, 1987.

Daniel Jurafsky and James H. Martin. *Speech and Language Processing.* Pearson Education, Inc., Upper Saddle River, New Jersey, 2nd edition, 2009. An introduction to natural language processing, computational linguistics, and speech recognition.

Su Nam Kim and Timothy Baldwin. How to pick out token instances of English verb-particle constructions. *Language Resources and Evaluation*, 44(1):97–113, 2010.

Paul Kingsbury and Martha Palmer. From TreeBank to PropBank. In *Proc. of LREC*, pages 1989–1993, Las Palmas, Canary Islands, May 2002.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. A large-scale classification of English verbs. *Language Resources and Evaluation*, 42(1):21–40, March 2008.

Tibor Kiss, Katja Keßelmeier, Antje Müller, Claudia Roch, Tobias Stadtfeld, and Jan Strunk. A logistic regression model of determiner omission in PPs. In *Coling 2010: Posters*, pages 561–569, Beijing, China, August 2010.

Terry Koo, Xavier Carreras, and Michael Collins. Simple semi-supervised dependency parsing. In *Proc. of ACL-08: HLT*, pages 595–603, Columbus, Ohio, USA, June 2008.

Ioannis Korkontzelos and Suresh Manandhar. Can recognising multiword expressions improve shallow parsing? In *Proc. of NAACL-HLT*, pages 636–644, Los Angeles, California, June 2010.

Koenraad Kuiper, Heather McCann, Heidi Quinn, Therese Aitchison, and Kees van der Veer. SAID. Technical Report LDC2003T10, Linguistic Data Consortium, Philadelphia, Pennsylvania, USA, June 2003. URL http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2003T10.

Dennis Kurzon and Silvia Adler, editors. *Adpositions: Pragmatic, semantic and syntactic perspectives.* Number 74 in Typological Studies in Language. John Benjamins, Amsterdam, May 2008.

Henry Kučera and W. Nelson Francis. *Computational analysis of present-day American English.* Brown University Press, Providence, RI, 1967.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML*, pages 282–289, Williamstown, MA, USA, June 2001.

George Lakoff. *Women, fire, and dangerous things: what categories reveal about the mind.* University of Chicago Press, Chicago, 1987.

Mirella Lapata and Alex Lascarides. Detecting novel compounds: the role of distributional evidence. In *Proc. of EACL*, pages 235–242, Budapest, Hungary, April 2003.

Joseph Le Roux, Matthieu Constant, and Antoine Rozenknop. Syntactic parsing and compound recognition via dual decomposition: application to French. In *Proc. of COLING*, pages 1875–1885, Dublin, Ireland, August 2014.

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault. *Automated Grammatical Error Detection for Language Learners.* Number 25 in Synthesis Lectures on Human Language Technologies. Morgan & Claypool, San Rafael, CA, 2 edition, February 2014.

Xin Li and Dan Roth. Learning question classifiers. In *Proc. of COLING*, pages 1–7, Taipei, Taiwan, August 2002.

Percy Liang. *Semi-supervised learning for natural language*. Master's thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, May 2005. URL http://people.csail.mit.edu/pliang/papers/meng-thesis.pdf.

Seth Lindstromberg. *English Prepositions Explained*. John Benjamins Publishing, Amsterdam, revised edition, 2010.

Ken Litkowski. The Preposition Project corpora. Technical Report 13-01, CL Research, Damascus, MD, 2013. URL http://www.clres.com/online-papers/TPPCorpora.pdf.

Ken Litkowski. Pattern Dictionary of English Prepositions. In *Proc. of ACL*, pages 1274–1283, Baltimore, Maryland, USA, June 2014.

Ken Litkowski and Orin Hargraves. The Preposition Project. In *Proc. of the Second ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, pages 171–179, Colchester, Essex, UK, April 2005.

Ken Litkowski and Orin Hargraves. SemEval-2007 Task 06: Word-Sense Disambiguation of Prepositions. In *Proc. of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 24–29, Prague, Czech Republic, June 2007.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. The Penn Treebank: annotating predicate argument structure. In *Proc. of HLT*, pages 114–119, Plainsboro, NJ, USA, March 1994.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

Mitchell P. Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. *Treebank-3*. Linguistic Data Consortium, Philadelphia, PA, 1999. LDC99T42.

Diana McCarthy, Bill Keller, and John Carroll. Detecting a continuum of compositionality in phrasal verbs. In *Proc. of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 73–80, Sapporo, Japan, July 2003.

I. Dan Melamed. Automatic discovery of non-compositional compounds in parallel data. In *Proc. of EMNLP*, pages 97–108, Providence, Rhode Island, USA, August 1997.

Igor Aleksandrovič Mel'čuk. Collocations and lexical functions. In A. P. Cowie, Richard W. Bailey, Noel Osselton, and Gabriele Stein, editors, *Phraseology: Theory, Analysis and Applications*, Oxford Studies in Lexicography and Lexicology, pages 23–54. Clarendon Press, Oxford, 1998.

Lukas Michelbacher, Alok Kothari, Martin Forst, Christina Lioma, and Hinrich Schütze. A cascaded classification approach to semantic head recognition. In *Proc. of EMNLP*, pages 793–803, Edinburgh, Scotland, UK., July 2011.

George A. Miller. Nouns in WordNet: a lexical inheritance system. *International Journal of Lexicography*, 3(4):245–264, December 1990.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. A semantic concordance. In *Proc. of HLT*, pages 303–308, Plainsboro, New Jersey, USA, March 1993.

Scott Miller, Jethran Guinness, and Alex Zamanian. Name tagging with word clusters and discriminative training. In *Proc. of HLT-NAACL*, pages 337–342, Boston, Massachusetts, USA, May 2004.

Behrang Mohit, Nathan Schneider, Rishav Bhowmick, Kemal Oflazer, and Noah A. Smith. Recall-oriented learning of named entities in Arabic Wikipedia. In *Proc. of EACL*, pages 162–173, Avignon, France, April 2012.

Begoña Villada Moirón and Jörg Tiedemann. Identifying idiomatic expressions using automatic word-alignment. In *Proc. of the EACL 2006 Workshop on Multi-word Expressions in a Multilingual Context*, pages 33–40, Trento, Italy, April 2006.

Rosamund Moon. *Fixed expressions and idioms in English: a corpus-based approach*. Oxford Studies in Lexicography and Lexicology. Clarendon Press, Oxford, UK, 1998.

Antje Müller, Olaf Hülscher, Claudia Roch, Katja Keßelmeier, Tobias Stadtfeld, Jan Strunk, and Tibor Kiss. An annotation schema for preposition senses in German. In *Proc. of the Fourth Linguistic Annotation Workshop*, pages 177–181, Uppsala, Sweden, July 2010.

Antje Müller, Claudia Roch, Tobias Stadtfeld, and Tibor Kiss. Annotating spatial interpretations of German prepositions. In *Proc. of ICSC*, pages 459–466, Palo Alto, CA, September 2011.

David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticæ Investigationes*, 30(1):3–26, 2007.

István Nagy T. and Veronika Vincze. Identifying verbal collocations in Wikipedia articles. In Ivan Habernal and Václav Matoušek, editors, *Text, Speech and Dialogue*, volume 6836 of *Lecture Notes in Computer Science*, pages 179–186. Springer, Berlin, 2011.

Preslav Nakov. Improved statistical machine translation using monolingual paraphrases. In Malik Ghallab, Constantine D. Spyropoulos, Nikos Fakotakis, and Nikos Avouris, editors, *Proc. of ECAI*, volume 178 of *Frontiers in Artificial Intelligence and Applications*, pages 338–342, Patras, Greece, July 2008.

Srinivas Narayanan. Moving right along: a computational model of metaphoric reasoning about events. In *Proc. of AAAI*, pages 121–128, Orlando, Florida, July 1999.

Roberto Navigli. Word sense disambiguation: a survey. *ACM Computing Surveys*, 41(2):10:1–10:69, February 2009.

David Newman, Nagendra Koilada, Jey Han Lau, and Timothy Baldwin. Bayesian text segmentation for index term identification and keyphrase extraction. In *Proc. of COLING 2012*, pages 2077–2092, Mumbai, India, December 2012.

Michael Niemann. Determining PP attachment through semantic associations and preferences. In Dominique Estival, editor, *Abstracts for the ANLP Post Graduate Workshop*, pages 25–32, Melbourne, Australia, January 1998.

Joakim Nivre and Jens Nilsson. Multiword units in syntactic parsing. In *Workshop on Methodologies and Evaluation of Multiword Units in Real-World Applications*, pages 39–46, Lisbon, Portugal, May 2004.

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. Idioms. *Language*, 70 (3):491–538, September 1994.

Elizabeth M. O'Dowd. *Prepositions and particles in English: a discourse-functional account*. Oxford University Press, New York, June 1998.

Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Dan Flickinger, Jan Hajic, Angelina Ivanova, and Yi Zhang. SemEval 2014 Task 8: Broad-Coverage Semantic Dependency Parsing. In *Proc. of SemEval*, pages 63–72, Dublin, Ireland, August 2014.

Tom O'Hara and Janyce Wiebe. Preposition semantic classification via Treebank and FrameNet. In Walter Daelemans and Miles Osborne, editors, *Proc. of CoNLL*, pages 79–86, Edmonton, Canada, June 2003.

Tom O'Hara and Janyce Wiebe. Exploiting semantic role resources for preposition disambiguation. *Computational Linguistics*, 35(2):151–184, 2009.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *Proc. of NAACL-HLT*, pages 380–390, Atlanta, Georgia, USA, June 2013.

Gerhard Paaß and Frank Reichartz. Exploiting semantic constraints for estimating supersenses with CRFs. In *Proc. of the Ninth SIAM International Conference on Data Mining*, pages 485–496, Sparks, Nevada, USA, May 2009.

Martha Palmer, Daniel Gildea, and Nianwen Xue. *Semantic Role Labeling*. Number 6 in Synthesis Lectures on Human Language Technologies. Morgan & Claypool, San Rafael, CA, January 2010.

Rebecca J. Passonneau, Ansaf Salleb-Aoussi, Vikas Bhardwaj, and Nancy Ide. Word sense annotation of polysemous words by multiple annotators. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proc. of LREC*, pages 3244–3249, Valletta, Malta, May 2010.

Pavel Pecina. Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44(1):137–158, 2010.

Scott S. L. Piao, Paul Rayson, Dawn Archer, Andrew Wilson, and Tony McEnery. Extracting multiword expressions with a semantic tagger. In *Proc. of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 49–56, Sapporo, Japan, July 2003.

Scott Songlin Piao, Paul Rayson, Dawn Archer, and Tony McEnery. Comparing and combining a semantic tagger and a statistical tool for MWE extraction. *Computer Speech & Language*, 19(4):378–397, October 2005.

Davide Picca, Alfio Massimiliano Gliozzo, and Massimiliano Ciaramita. Supersense Tagger for Italian. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, and Daniel Tapias, editors, *Proc. of LREC*, pages 2386–2390, Marrakech, Morocco, May 2008.

Davide Picca, Alfio Massimiliano Gliozzo, and Simone Campora. Bridging languages by SuperSense entity tagging. In *Proc. of NEWS*, pages 136–142, Suntec, Singapore, August 2009.

Karl Pichotta and John DeNero. Identifying phrasal verbs using many bilingual corpora. In *Proc. of EMNLP*, pages 636–646, Seattle, Washington, USA, October 2013.

Vinodkumar Prabhakaran, Michael Bloodgood, Mona Diab, Bonnie Dorr, Lori Levin, Christine D. Piatko, Owen Rambow, and Benjamin Van Durme.

Statistical modality tagging from rule-based annotations and crowdsourcing. In *Proc. of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 57–64, Jeju, Republic of Korea, July 2012.

Geoffrey K. Pullum and Rodney Huddleston. Prepositions and preposition phrases. In Rodney Huddleston and Geoffrey K. Pullum, editors, *The Cambridge Grammar of the English Language*, pages 579–611. Cambridge University Press, Cambridge, UK, 2002.

James Pustejovsky. *The Generative Lexicon*. MIT Press, Cambridge, MA, January 1998.

Ashequl Qadir and Ellen Riloff. Ensemble-based semantic lexicon induction for semantic tagging. In *Proc. of *SEM*, pages 199–208, Montréal, Canada, June 2012.

Likun Qiu, Yunfang Wu, Yanqiu Shao, and Alexander Gelbukh. Combining contextual and structural information for supersense tagging of Chinese unknown words. In *Computational Linguistics and Intelligent Text Processing: Proceedings of the 12th International Conference (CICLing'11)*, volume 6608 of *Lecture Notes in Computer Science*, pages 15–28. Springer, Berlin, 2011.

Anita Rácz, István Nagy T., and Veronika Vincze. 4FX: Light verb constructions in a multilingual parallel corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proc. of LREC*, pages 710–715, Reykjavík, Iceland, May 2014.

Carlos Ramisch. *A generic and open framework for multiword expressions treatment: from acquisition to applications*. Ph.D. dissertation, University of Grenoble and Federal University of Rio Grande do Sul, Grenoble, France, 2012. URL http://www.inf.ufrgs.br/~ceramisch/download_files/thesis-getalp.pdf.

Carlos Ramisch, Aline Villavicencio, Leonardo Moura, and Marco Idiart. Picking them up and figuring them out: verb-particle constructions, noise and idiomaticity. In *Proc. of CoNLL*, pages 49–56, Manchester, England, August 2008.

Carlos Ramisch, Aline Villavicencio, and Christian Boitet. mwetoolkit: a framework for multiword expression identification. In *Proc. of LREC*, pages 662–669, Valletta, Malta, May 2010.

Carlos Ramisch, Vitor De Araujo, and Aline Villavicencio. A broad evaluation of techniques for automatic acquisition of multiword expressions. In *Proc. of ACL 2012 Student Research Workshop*, pages 1–6, Jeju Island, Korea, July 2012.

Carlos Ramisch, Laurent Besacier, and Alexander Kobzar. How hard is it to automatically translate phrasal verbs from English to French? In Johanna Monti, Ruslan Mitkov, Gloria Corpas Pastor, and Violeta Seretan, editors, *Proc. of the MT Summit 2013 Workshop on Multi-word Units in Machine Translation and Translation Technologies*, pages 53–61, Nice, France, September 2013.

Lance A. Ramshaw and Mitchell P. Marcus. Text chunking using transformation-based learning. In *Proc. of the Third ACL Workshop on Very Large Corpora*, pages 82–94, Cambridge, Massachusetts, USA, June 1995.

Mohammad Rasooli, Heshaam Faili, and Behrouz Minaei-Bidgoli. Unsupervised identification of Persian compound verbs. In Ildar Batyrshin and Grigori Sidorov, editors, *Advances in Artificial Intelligence*, volume 7094 of *Lecture Notes in Computer Science*, pages 394–406. Springer, Berlin, 2011.

Lev Ratinov and Dan Roth. Design challenges and misconceptions in named entity recognition. In *Proc. of CoNLL*, pages 147–155, Boulder, Colorado, USA, June 2009.

Gisa Rauh. *Approaches to prepositions*. Tubinger Beiträge zur Linguistik. G. Narr, Tübingen, Germany, 1991.

Gisa Rauh. On the grammar of lexical and non-lexical prepositions in English. In Cornelia Zelinsky-Wibbelt, editor, *The Semantics of Prepositions: From Mental Processing to Natural Language Processing*, pages 99–150. Mouton de Gruyter, New York, 1993.

Paul Rayson, Dawn Archer, Scott Piao, and Tony McEnery. The UCREL semantic analysis system. In *Proc. of the Workshop on Beyond Named Entity Recognition: Semantic Labelling for NLP Tasks*, pages 7–12, Lisbon, Portugal, May 2004.

Marta Recasens and Eduard Hovy. BLANC: Implementing the Rand index for coreference evaluation. *Natural Language Engineering*, 17(04):485–510, 2011.

Michael J. Reddy. The conduit metaphor: a case of frame conflict in our language about language. In Andrew Ortony, editor, *Metaphor and Thought*, pages 284–324. Cambridge University Press, Cambridge, UK, 1979.

Terry Regier. *The human semantic potential: spatial language and constrained connectionism*. MIT Press, Cambridge, MA, September 1996.

Roi Reichart and Ari Rappoport. Tense sense disambiguation: a new syntactic polysemy task. In *Proc. of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 325–334, Cambridge, MA, October 2010.

Zhixiang Ren, Yajuan Lü, Jie Cao, Qun Liu, and Yun Huang. Improving statistical machine translation using domain bilingual multiword expressions. In *Proc. of MWE*, pages 47–54, Suntec, Singapore, August 2009.

Stefan Riezler. On the problem of theoretical terms in empirical computational linguistics. *Computational Linguistics*, 40(1):235–245, November 2013.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. Named entity recognition in tweets: an experimental study. In *Proc. of EMNLP*, pages 1524–1534, Edinburgh, Scotland, UK, July 2011.

Stefano Dei Rossi, Giulia Di Pietro, and Maria Simi. Description and results of the SuperSense tagging task. In Bernardo Magnini, Francesco Cutugno, Mauro Falcone, and Emanuele Pianta, editors, *Evaluation of Natural Language and Speech Tools for Italian*, number 7689 in Lecture Notes in Computer Science, pages 166–175. Springer, Berlin, January 2013.

J. K. Rowling. *Harry Potter and the Half-Blood Prince*. Arthur A. Levine Books, New York, NY, 2005.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword expressions: a pain in the neck for NLP. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 2276 of *Lecture Notes in Computer Science*, pages 189–206. Springer, Berlin, Germany, 2002.

Patrick Saint-Dizier. PrepNet: a multilingual lexical description of prepositions. In *Proc. of LREC*, volume 6, pages 1021–1026, Genoa, Italy, May 2006a.

Patrick Saint-Dizier. Introduction to the Syntax and Semantics of Prepositions. In Patrick Saint-Dizier and Nancy Ide, editors, *Syntax and Semantics of Prepositions*, volume 29 of *Text, Speech and Language Technology*, pages 1–25. Springer, Dordrecht, The Netherlands, 2006b.

Patrick Saint-Dizier and Nancy Ide, editors. *Syntax and Semantics of Prepositions*, volume 29 of *Text, Speech and Language Technology*. Springer, Dordrecht, The Netherlands, 2006.

Bahar Salehi, Narjes Askarian, and Afsaneh Fazly. Automatic identification of Persian light verb constructions. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 7181 of *Lecture Notes in Computer Science*, pages 201–210. Springer, Berlin, 2012.

Bahar Salehi, Paul Cook, and Timothy Baldwin. Using distributional similarity of multi-way translations to predict multiword expression compositionality. In *Proc. of EACL*, pages 472–481, Gothenburg, Sweden, April 2014.

Nathan Schneider and Reut Tsarfaty. *Design Patterns in Fluid Construction Grammar*, Luc Steels (editor). *Computational Linguistics*, 39(2):447–453, June 2013.

Nathan Schneider, Behrang Mohit, Kemal Oflazer, and Noah A. Smith. Coarse lexical semantic annotation with supersenses: an Arabic case study. In *Proc. of ACL*, pages 253–258, Jeju Island, Korea, July 2012.

Nathan Schneider, Behrang Mohit, Chris Dyer, Kemal Oflazer, and Noah A. Smith. Supersense tagging for Arabic: the MT-in-the-middle attack. In *Proc. of NAACL-HLT*, pages 661–667, Atlanta, Georgia, USA, June 2013.

Nathan Schneider, Emily Danchik, Chris Dyer, and Noah A. Smith. Discriminative lexical semantic segmentation with gaps: running the MWE gamut. *Transactions of the Association for Computational Linguistics*, 2: 193–206, April 2014a.

Nathan Schneider, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T. Mordowanec, Henrietta Conrad, and Noah A. Smith. Comprehensive annotation of multiword expressions in a social web corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proc. of LREC*, pages 455–461, Reykjavík, Iceland, May 2014b.

Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. Extended named entity hierarchy. In *Proc. of LREC*, pages 1818–1824, Las Palmas, Canary Islands, May 2002.

Yutaro Shigeto, Ai Azuma, Sorami Hisamoto, Shuhei Kondo, Tomoya Kouse, Keisuke Sakaguchi, Akifumi Yoshimoto, Frances Yung, and Yuji Matsumoto. Construction of English MWE dictionary and its application to POS tagging. In *Proc. of the 9th Workshop on Multiword Expressions*, pages 139–144, Atlanta, Georgia, USA, June 2013.

Reshef Shilon, Hanna Fadida, and Shuly Wintner. Incorporating linguistic knowledge in statistical machine translation: translating prepositions. In *Proc. of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, pages 106–114, Avignon, France, April 2012.

Iliana Simova and Valia Kordoni. Improving English-Bulgarian statistical machine translation by phrasal verb treatment. In Johanna Monti, Ruslan Mitkov, Gloria Corpas Pastor, and Violeta Seretan, editors, *Proc. of the MT Summit 2013 Workshop on Multi-word Units in Machine Translation and Translation Technologies*, pages 62–71, Nice, France, September 2013.

Rita Simpson and Dushyanthi Mendis. A corpus-based study of idioms in academic speech. *TESOL Quarterly*, 37(3):419–441, 2003.

Noah A. Smith. *Linguistic Structure Prediction*. Number 13 in Synthesis Lectures on Human Language Technologies. Morgan & Claypool, San Rafael, CA, May 2011.

Vivek Srikumar and Dan Roth. An inventory of preposition relations. Technical Report arXiv:1305.5785, May 2013a. URL `http://arxiv.org/abs/1305.5785`.

Vivek Srikumar and Dan Roth. Modeling semantic relations expressed by prepositions. *Transactions of the Association for Computational Linguistics*, 1:231–242, May 2013b.

Shashank Srivastava and Eduard Hovy. Vector space semantics with frequency-driven motifs. In *Proc. of ACL*, pages 634–643, Baltimore, Maryland, June 2014.

Manfred Stede. *Discourse Processing*. Number 15 in Synthesis Lectures on Human Language Technologies. Morgan & Claypool, San Rafael, CA, November 2011.

Mark Steedman. *The Syntatic Process*. MIT Press, Cambridge, MA, 2000.

Luc Steels, editor. *Computational Issues in Fluid Construction Grammar*. Number 7249 in Lecture Notes in Computer Science. Springer, Berlin, 2012.

Luc Steels, Jan-Ola Östman, and Kyoko Ohara, editors. *Design patterns in Fluid Construction Grammar*. Number 11 in Constructional Approaches to Language. John Benjamins, Amsterdam, 2011.

Pontus Stenetorp, Sampo Pyysalo, Sophia Ananiadou, and Jun'ichi Tsujii. Generalising semantic category disambiguation with large lexical resources for fun and profit. *Journal of Biomedical Semantics*, 5(1):26, June 2014.

Leonard Talmy. Force dynamics in language and cognition. *Cognitive Science*, 12(1):49–100, January 1988.

Leonard Talmy. Fictive motion in language and "ception". In Paul Bloom, Mary A. Peterson, Nadel Lynn, and Merrill F. Garrett, editors, *Language and Space*, pages 211–276. MIT Press, Cambridge, MA, 1996.

Bin Tan and Fuchun Peng. Unsupervised query segmentation using generative language models and wikipedia. In *Proc. of WWW*, pages 347–356, Beijing, China, April 2008.

Yee Fan Tan, Min-Yen Kan, and Hang Cui. Extending corpus-based identification of light verb constructions using a supervised learning framework. In *Proc. of the EACL Workshop on Multi-Word Expressions in a Multilingual Context*, pages 49–56, Trento, Italy, April 2006.

Takaaki Tanaka and Timothy Baldwin. Noun-noun compound machine translation: a feasibility study on shallow processing. In *Proc. of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 17–24, Sapporo, Japan, July 2003.

Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R. Walter, Ashis Gopal Banerjee, Seth J. Teller, and Nicholas Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proc. of AAAI*, pages 1507–1514, San Francisco, California, USA, August 2011.

James W. Thatcher. Characterizing derivation trees of context-free grammars through a generalization of finite automata theory. *Journal of Computer and System Sciences*, 1(4):317–322, December 1967.

Erik F. Tjong Kim Sang and Sabine Buchholz. Introduction to the CoNLL-2000 Shared Task: Chunking. In *Proc. of the Fourth Conference on Computational Natural Language Learning and of the Second Learning Language in Logic Workshop*, pages 127–132, Lisbon, Portugal, September 2000.

Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In Walter Daelemans and Miles Osborne, editors, *Proc. of CoNLL*, pages 142–147, Edmonton, Canada, June 2003.

Takashi Tomokiyo and Matthew Hurst. A language model approach to keyphrase extraction. In *Proc. of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 33–40, Sapporo, Japan, July 2003.

Kristina Toutanova and Hisami Suzuki. Generating case markers in machine translation. In *Proc. of NAACL-HLT*, pages 49–56, Rochester, New York, April 2007.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proc. of HLT-NAACL*, pages 173–180, Edmonton, Alberta, Canada, June 2003.

Stephen Tratz. *Semantically-enriched parsing for natural language understanding*. Ph.D. dissertation, University of Southern California, Los Angeles, California, December 2011.

Stephen Tratz and Dirk Hovy. Disambiguation of preposition sense using linguistically motivated features. In *Proc. of NAACL-HLT Student Research Workshop and Doctoral Consortium*, pages 96–100, Boulder, Colorado, June 2009.

Stephen Tratz and Eduard Hovy. A fast, accurate, non-projective, semantically-enriched parser. In *Proc. of EMNLP*, pages 1257–1268, Edinburgh, Scotland, UK, July 2011.

Beata Trawinski. Licensing complex prepositions via lexical constraints. In *Proc. of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 97–104, Sapporo, Japan, July 2003.

Jesse L. Tseng. *The representation and selection of prepositions*. Ph.D. dissertation, University of Edinburgh, Edinburgh, Scotland, UK, 2000. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.70.4995&rep=rep1&type=pdf.

Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6:1453–1484, September 2005.

Yulia Tsvetkov and Shuly Wintner. Extraction of multi-word expressions from small parallel corpora. In *Coling 2010: Posters*, pages 1256–1264, Beijing, China, August 2010.

Yulia Tsvetkov and Shuly Wintner. Identification of multi-word expressions by combining multiple linguistic information sources. In *Proc. of EMNLP*, pages 836–845, Edinburgh, Scotland, UK, July 2011.

Yulia Tsvetkov, Nathan Schneider, Dirk Hovy, Archna Bhatia, Manaal Faruqui, and Chris Dyer. Augmenting English adjective senses with supersenses. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proc. of LREC*, pages 4359–4365, Reykjavík, Iceland, May 2014.

Yuancheng Tu and Dan Roth. Learning English light verb constructions: contextual or statistical. In *Proc. of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 31–39, Portland, Oregon, USA, June 2011.

Yuancheng Tu and Dan Roth. Sorting out the most confusing English phrasal verbs. In *Proc. of *SEM*, pages 65–69, Montréal, Quebec, Canada, June 2012.

Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning. In *Proc. of ACL*, pages 384–394, Uppsala, Sweden, July 2010.

Andrea Tyler and Vyvyan Evans. *The Semantics of English Prepositions: Spatial Scenes, Embodied Meaning and Cognition.* Cambridge University Press, Cambridge, UK, 2003.

Kiyoko Uchiyama and Shun Ishizaki. A disambiguation method for Japanese compound verbs. In *Proc. of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 81–88, Sapporo, Japan, July 2003.

Zdeňka Urešová, Jan Hajič, Eva Fučíková, and Jana Šindlerová. An analysis of annotation of verb-noun idiomatic combinations in a parallel dependency corpus. In *Proc. of the 9th Workshop on Multiword Expressions*, pages 58–63, Atlanta, Georgia, USA, June 2013.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. SemEval-2013 Task 1: TempEval-3: Evaluating Time Expressions, Events, and Temporal Relations. In *Proc. of SemEval*, pages 1–9, Atlanta, Georgia, USA, June 2013.

VerbNet Annotation Guidelines. VerbNet Annotation Guidelines. URL http://verbs.colorado.edu/verb-index/VerbNet_Guidelines.pdf.

Torben Vestergaard. *Prepositional phrases and prepositional verbs: a study in grammatical function.* Mouton, The Hague, 1977.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. A model-theoretic coreference scoring scheme. In *Proc. of MUC-6*, pages 45–52, Columbia, Maryland, USA, November 1995.

Aline Villavicencio. Verb-particle constructions and lexical resources. In *Proc. of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 57–64, Sapporo, Japan, July 2003.

Veronika Vincze. Light verb constructions in the SzegedParalellFX English-Hungarian parallel corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proc. of LREC*, pages 2381–2388, Istanbul, Turkey, May 2012.

Veronika Vincze, István Nagy T., and Gábor Berend. Multiword expressions and named entities in the Wiki50 corpus. In *Proc. of RANLP*, pages 289–295, Hissar, Bulgaria, September 2011.

Veronika Vincze, István Nagy T., and János Zsibrita. Learning to detect English and Hungarian light verb constructions. *ACM Transactions on Speech and Language Processing*, 10(2):6:1–6:25, June 2013a.

Veronika Vincze, János Zsibrita, and István Nagy T. Dependency parsing for identifying Hungarian light verb constructions. In *Proc. of the Sixth International Joint Conference on Natural Language Processing*, pages 207–215, Nagoya, Japan, October 2013b.

Ralph Weischedel and Ada Brunstein. BBN pronoun coreference and entity type corpus. Technical Report LDC2005T33, Linguistic Data Consortium, Philadelphia, PA, 2005. URL http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2005T33.

Alison Wray. Formulaic sequences in second language teaching: principle and practice. *Applied Linguistics*, 21(4):463–489, December 2000.

Stefanie Wulff. *Rethinking idiomaticity: a usage-based approach.* Research in Corpus and Discourse. Continuum International Publishing Group, London, November 2008.

Stefanie Wulff. Marrying cognitive-linguistic theory and corpus-based methods: on the compositionality of English V NP-idioms. In Dylan Glynn and Kerstin Fischer, editors, *Corpus-driven Cognitive Semantics*, pages 223–238. Mouton, Berlin, 2010.

Deyi Xiong and Min Zhang. A sense-based translation model for statistical machine translation. In *Proc. of ACL*, pages 1459–1469, Baltimore, Maryland, USA, June 2014.

Yang Xu and Charles Kemp. Constructing spatial concepts from universal primitives. In Stellan Ohlsson and Richard Catrambone, editors, *Proc. of CogSci*, pages 346–351, Portland, Oregon, August 2010.

Patrick Ye and Timothy Baldwin. MELB-YB: Preposition sense disambiguation using rich semantic features. In *Proc. of SemEval*, pages 241–244, Prague, Czech Republic, June 2007.

Liang-Chih Yu, Chung-Hsien Wu, Ru-Yng Chang, Chao-Hong Liu, and Eduard Hovy. Annotation and verification of sense pools in OntoNotes. *Information Processing & Management*, 46(4):436–447, July 2010.

Annie Zaenen, Jean Carletta, Gregory Garretson, Joan Bresnan, Andrew Koontz-Garboden, Tatiana Nikitina, M. Catherine O'Connor, and Tom Wasow. Animacy encoding in English: why and how. In Bonnie Webber and Donna K. Byron, editors, *ACL 2004 Workshop on Discourse Annotation*, pages 118–125, Barcelona, Spain, July 2004.

Cornelia Zelinsky-Wibbelt. Interpreting and translating prepositions: a cognitively based formulation. In Cornelia Zelinsky-Wibbelt, editor, *The Semantics of Prepositions: From Mental Processing to Natural Language Processing*, number 3 in Natural Language Processing, pages 351–390. Mouton de Gruyter, Berlin, 1993a.

Cornelia Zelinsky-Wibbelt, editor. *The semantics of prepositions: from mental processing to natural language processing*. Number 3 in Natural Language Processing. Mouton de Gruyter, Berlin, 1993b.

John M. Zelle and Raymond J. Mooney. Learning to parse database queries using inductive logic programming. In *Proc. of AAAI*, pages 1050–1055, Portland, Oregon, August 1996.

Luke S. Zettlemoyer and Michael Collins. Learning to map sentences to logical form: structured classification with probabilistic categorial grammars. In *Proc. of UAI*, pages 658–666, Edinburgh, Scotland, UK, July 2005.

Joost Zwarts and Yoad Winter. Vector space semantics: a model-theoretic analysis of locative prepositions. *Journal of Logic, Language and Information*, 9:169–211, 2000.

Index