Lexical Semantic Analysis in Natural Language Text

PH.D. THESIS PROPOSAL

Nathan Schneider
Language Technologies Institute ⋄ School of Computer Science
Carnegie Mellon University
November 17, 2012

0 Introduction

This thesis concerns the lexical semantics of natural language text, studying from a computational perspective how words in sentences ought to be analyzed, how this analysis can be automated, and to what extent such analysis matters to other natural language processing (NLP) problems.

It may not be obvious that words of text should be analyzed at all. After all, superficial uses of word tokens—most famously, **bag-of-words** representations and **n-grams**—are quite successful in settings ranging from information retrieval to language modeling.

On the other hand, it is clear that there is a fuzzy relationship between the use of a word and the intended meaning, even when orthographic and morphological normalization (such as lemmatization or stemming) are applied. The word lexicon and its derivatives offer a good case in point:

- A specific **lexicon** (or **dictionary**) is a list of natural language words or expressions. The list may be flat or structured (e.g., into a taxonomy). Entries (called **lexical items**, **lexical units**, or **lexemes**) may be associated with metadata such as definitions, etymologies, and corpus frequency counts. A **bilingual lexicon** includes translation mappings between the vocabularies of two languages.

- More abstractly, **lexicon** (or **lexis**) can refer to the vocabulary that a speaker of a language has at his disposal, or to the language's collective vocabulary aggregated over all speakers.
- Some linguistic theories posit a formal distinction between the **lexicon** and **grammar** (Bloomfield, 1933; Chomsky, 1965). In generative grammar these are taken to be separate modules in the mental architecture of language, with the former consisting of an exhaustive list of atomic entries, and the latter consisting of abstract **rules** (syntactic, morphological, etc.) for assembling well-formed utterances. By contrast, theories such as Construction Grammar (Fillmore et al., 1988; Goldberg, 2006) disavow the strong modularity assumption, instead viewing a speaker's linguistic knowledge as spanning a continuum from the most **lexical** expressions—where a single concept is expressed by a single, specific word like boy—to the most grammatical, e.g. the abstract syntactic pattern NP.subj BE V.pastpart (the English passive construction) to indicate an action on something while deemphasizing the party responsible for that action.
- In computational linguistics, type-level generalizations may—for theoretical or practical reasons—be made specific to individual vocabulary items, or may abstract away from the vocabulary. For instance, the rules in a syntactic grammar may include a category generalizing over all verbs, which is said to be **unlexicalized**, or may have a separate category for each verb lexeme so as to capture valency distinctions at a finer level of granularity, in which case that category (as well as the rules using it) are said to be **lexicalized**.

The most frequent words present an extreme case of semantic promiscuity: for instance, the verb make is ambiguous between highly contentful usages (*make a salad*), grammaticized/semantically "light" usages (*make a decision*, *make up a story*), and names (the software utility `make`). And just as a word can have many meanings (or shades of meaning), many different words may have synonymous or similar meanings. We would therefore expect information provided by models of lexical meaning in context to benefit problems of sentence-level analysis (e.g., syntactic and semantic parsing) and generation (e.g., machine translation).

The traditional approach to lexical semantics calls for a detailed characterization of meanings within a meticulously crafted lexical resource, the chief example being English WordNet (Fellbaum, 1998). But even listing a single word's possible meanings at a level of granularity that everyone can agree on is far from simple (Hovy et al., 2006). Further, if the sense tagging scheme is lexicalized—that is, ev-

In WordNet 3.1	OOV	In WordNet 3.1	OOV
take place (‘occur’)	hold hostage	DNA	IPA
carry out (‘execute’)	stress out	haute couture	crème brûlée
extreme unction	anatomical snuffbox	ricin	Thebacon
proper name	named entity	bring home the bacon	gum up the works
word salad	ice cream sandwich	from time to time	beyond repair
kind of, kinda, sort of	sorta	tank top	hoodie

Figure 1: Examples of the vocabulary coverage of the WordNet lexicon.

Carnegie Mellon University {N:ARTIFACT} an engineering university in Pittsburgh
Andrew Carnegie {N:PERSON} United States industrialist and philanthropist who endowed education and public libraries and research trusts (1835-1919)
Andrew Mellon {N:PERSON} United States financier and philanthropist (1855-1937)
Andrew McCallum OOV
UNIX {N:COMMUNICATION} trademark for a powerful operating system

Apple {N:FOOD} fruit with red or yellow or green skin and sweet to tart crisp whitish flesh; {N:PLANT} native Eurasian tree widely cultivated in many varieties for its firm rounded edible fruits
Microsoft OOV
Google {N:COMMUNICATION} a widely used search engine that uses text-matching techniques to find web pages that are important and relevant to a user's search; {V:COGNITION} search the internet (for information) using the Google search engine

Figure 2: Named entities and WordNet.

ery sense category is specific to a lexical item—then it has nothing to say about out-of-vocabulary words, or about hitherto unseen meanings of known words.

A further complication is that what we regard as a “basic” lexical meaning may be expressed with more than one orthographic word. For example, we analyze the sentence

(1) A minute later they turned the corner into the side street where the Hog 's Head 's sign creaked a little , though there was no breeze . [*HBP*, p. 554]

as having four multiword units: the nominal compound side street; the named entity Hog 's Head (the name of a pub); the measurement phrase a little; and the discontiguous expression turned. . . corner, which is a verb-noun construction. (§2 confronts the issue of multiword units in greater detail.) Lexical resources such as WordNet contain many multiword units, but the treatment of these units appears to be largely ad hoc (figures 1 and 2). Even if they are known at the type level, token-level ambiguity requires techniques for identifying multiwords in context. Indeed, entire

A minute later they turned the corner into the side street where the Hog 's Head 's
N:TIME V:MOTION N:LOC N:ARTIFACT N:ARTIFACT
sign creaked a little , though there was no breeze .
N:ARTIFACT V:PERCEPTION N:PHENOMENON

Figure 3: Sentence (1) annotated for supersenses. The label N:LOC indicates the nominal LOCATION category.

literatures on named entity recognition (Nadeau and Sekine, 2007) and multiword expressions (MWEs; see §2) have sought to tackle the many subtleties of multiword lexical phenomena. Yet from this literature it is not clear whether a coherent account of these phenomena can be formulated. Individual kinds of multiword constructions are typically addressed in isolation; no comprehensive multiword annotation scheme, let alone datasets or models, has been put forward.

We believe that the current state of affairs warrants a more “pragmatic” approach to computational lexical semantics as applied to tokens of text. Specifically, we have in mind the following desiderata for a token-level representation: it should answer the questions

1. **What are the lexical units in the text?**
2. **What semantics are associated with the lexical units?**

and in doing so should be

- *robust*, with the potential for full token coverage and strong performance regardless of topic, genre, or language;
- *explicit*, analyzing tokens in a well-defined and intuitive representation; and
- *efficient*, facilitating rapid human annotation as well as scalable machine learning algorithms.

We elaborate on these three criteria in turn.

0.1 Robust

A truly robust approach to semantic analysis would cover most tokens with few language- or domain-specific dependencies. A primary consideration is the availability of **lexical resources**. While English WordNet is quite extensive, having benefited from decades of lexicographic research, most languages are not so lucky. When faced with developing text processing tools for such a language, rather than accept

the necessarily limited coverage of a WordNet-style semantic lexicon or give up on semantics altogether, we advocate a middle ground in the form of an **unlexicalized semantic representation**. The so-called **supersense tags** are one such scheme: they represent coarse conceptual groupings such as PERSON and ARTIFACT (for nouns) and CREATION (for verbs). (Figure 3 gives an example sentence annotated with supersenses.) Though the supersense categories originated from the WordNet project, they are *general* enough to be assigned to nouns and verbs without being bound by the availability or coverage of a lexicon: for instance, the tagging guidelines presented in §1 specify COMMUNICATION as the appropriate tag for software, which governs annotation of the noun kernel in an operating systems context even though its only WordNet senses fall under PLANT and COGNITION. Supersenses are also more conducive than fine-grained sense lexicons to *rapid* and *reliable* full-text annotation. §1 discusses our approach to annotation and automatic tagging with supersenses. §2 proposes to generalize the English supersense tagging task to include additional kinds of multiword expressions.

A second aspect of robustness is the **oracle token coverage**, by which we mean the number of tokens that should be analyzed as part of the representation given perfect input. The tokenized sentence in Figure 3 comprises 28 tokens, but only 11 of them are part of a nominal or verbal expression. The supersense tagging task (as traditionally defined) therefore covers about 40% of the tokens in this sentence. In particular, a majority of the remaining tokens are **function words**. §3 expands the semantic tagging scheme to include **prepositions**, a closed class of highly frequent words that serve richly diverse functions. For example, the sentence

(2) It even got a little worse during a business trip to the city , so on the advice of a friend I set up an appointment with True Massage . [Yelp.com user review of a massage therapist]

has 6 prepositions: 2 mark spatiotemporal relations (during, to), 1 marks reciprocality (with), 1 marks an agent-source function (of), and the remaining 2 participate in a multiword expression with a content word (on. . . advice, set up). We therefore define, annotate for, and model preposition functions in tandem with our treatment of multiword expressions.

A third aspect of robustness concerns the expected **input**. We think it is reasonable to assume the text has been tokenized and morphologically preprocessed (e.g. with a part-of-speech tagger or lemmatizer), as these are fundamental NLP components that can be constructed with limited resources (Beesley and Karttunen, 2003). However, more sophisticated components like parsers are not generally

available for most languages, and the ones that are available may not generalize well to new domains. We will therefore aim to develop a semantic tagger that does not depend on syntax. Given that prepositions typically serve a linking function between syntactic phrases, and that many multiword expressions have internal syntactic structure (e.g. verb-object idioms), it is an open question to what extent a full syntactic parse is necessary for our task. Our evaluation will include English treebank data in order to test the effects of gold syntax vs. parser output vs. no syntax.

0.2 Explicit

Basing a representation on a fixed inventory of interpretable categories facilitates understanding of linguistic phenomena through annotation and error analysis.

In this thesis we adopt three semantic sense tagsets: one for nominal expressions (25 supersenses), a second for verbal expressions (15 supersenses), and a third for prepositions (≈20–40 senses, yet to be finalized). Defining and refining explicit categories through human annotation is a data-driven process that gives greater insight into the linguistic phenomena at work, and produces readily interpretable training data and (semi-)supervised system outputs. Some of the approaches below will exploit unsupervised methods, but only as a means to an explicitly defined end.

The long tail of language phenomena is assuredly a concern for lexical semantics. A compromise between **tagset complexity** and **tag specificity** is therefore needed in order to attain high coverage. In past work have found 10s of tags to be a manageable number for annotators (Gimpel et al., 2011; Schneider et al., 2012), so we stick with the original supersense categories; some of these are general enough to assure full coverage of nominal and verbal expressions. Through iterative annotation and discussion we develop lucid definitions of the supersense categories under which inter-annotator agreement is satisfactory. A small number of preposition sense tags are likewise expected to cover most cases, and a "miscellaneous" category will be reserved for the most idiosyncratic usages.

0.3 Efficient

A semantic representation would ideally lend itself to rapid, low-cost annotation of free text as well as computationally efficient modeling techniques.

Human time is precious. Under certain conditions the cost of annotation (greatest when annotators require extensive training or prior expertise) can be mitigated

by strategies such as active learning and crowdsourcing. But here, we argue for a cost-effective **annotation task design**—encompassing the annotation formalism (e.g., tagset), instructions/guidelines, interface, and training/review processes. Our lexical semantic representation promises to fit the bill because it presents a manageable number of options, which simplifies each decision, and because these options are consistent across tokens (unlexicalized), which makes it easier to remember the meaning of each option.

On the computational side, the representation allows for a **sequence tagging** formulation of the analysis task. Sequence tagging is a central problem in NLP; in particular, chunking problems are typically reduced to tagging problems (Ramshaw and Marcus, 1995), permitting inference algorithms that scale linearly with the length of the sequence. This will naturally facilitate efficient joint modeling of the grouping of tokens into lexical units and the assignment of semantic sense categories to the units. We show that even gappy chunks with arbirarily large gaps can be accommodated in this framework under some linguistically reasonable assumptions about the nesting/interleaving of chunks. Moreover, we note that without syntactic parsing (which will be avoided for robustness—see above), all of the necessary preprocessing should be achievable in linear time, since morphological preprocessing/part-of-speech tagging tools typically use token-level or sequence models.

We will use the term **lexical semantic analysis** (**LxSA**) for the problem of detecting lexical units in text and assigning semantic information (here, supersenses and preposition functions) to these units.

The central *linguistic* challenge of the thesis will be to precisely define this task in a way that meets the demands of robustness, explicitness, and efficiency. This process will produce annotation guidelines, annotated datasets, and quantitative measurements of inter-annotator agreement. The new datasets, spanning multiple domains, will be in Arabic (nominal supersenses only; see §1) and English (the full LxSA representation). Where possible, we will also adapt existing corpora to test individual components of the LxSA task, e.g. lexical unit detection for multiword expressions in the French Treebank (Abeillé et al., 2003).

The central *computational* challenge will be to show that, given some human-annotated data, the LxSA task can be automated. The product of this component will be open-source lexical semantic tagging software based on a discriminative statistical sequence model. A crucial innovation here will be the extension of the standard tagging-chunking paradigm to detect discontiguous ("gappy") multiword

units. Other techniques to be explored aim to exploit indirect evidence from unlabeled data or from other languages within semi-supervised learning. We will conduct empirical intrinsic and extrinsic evaluations of our approach in multiple languages and genres, measuring the quality of the system's predictions relative to human annotations as well as reporting efficiency measures.

Next we address the core components of our formulation of the LxSA problem: supersense tagging (§1), multiword expression identification (§2), and preposition function tagging (§3). §4 then describes how we plan to integrate these components in a single, unified model. We explore prospects for applying the output of LxSA to extrinsic tasks, namely frame-semantic parsing (§5) and machine translation (§6). Finally, we wrap up with concluding remarks in §7 and a proposed timeline in §8.

1 Supersense Tagging of Nouns and Verbs

In the face of limited lexical semantic resources, what is the most practical approach to semantic annotation that would lead to a useful dataset and NLP tool? This is the question we faced having created a named entity corpus and tagger for Arabic Wikipedia (Mohit et al., 2012). Aside from named entities, the standard kinds of general-purpose semantic annotation—e.g., WordNet-style word senses or predicate-argument structures—would not have been feasible (or would have been severely limited in coverage) for a small corpus creation effort in Arabic.

In completed work that forms the first part of this thesis, we proposed that the WordNet **supersenses** be used directly for *annotation*, and developed and released a small, multi-domain corpus of Arabic Wikipedia articles with nominal supersenses (Schneider et al., 2012). The highlights of that work are summarized here.

1.1 Semantic Categorization Schemes

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 4. 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 **splitting** the word into multiple **lexicalized** senses (figure 4a). **Named entity 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

(a) splitting 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) lumping {ARTIFACT: seal, stamp, ...}, {ANIMAL: seal, Tasmanian devil, Burmese python, ...}

Figure 4: Categorization schemes for two senses of the noun seal and related concepts.

way the task is **unlexicalized** and contributes information by **lumping** together multiple lexical items that belong to the same (coarse) semantic class.

1.1.1 WordNet

Figure 4b is a flattened, partial view of the taxonomy of the **WordNet** semantic lexicon (Fellbaum, 1998). This approach can be considered a hybrid—it both lumps and splits lexical items in mapping them to **synsets** (senses possibly shared by multiple lexemes) and defining groupings over synsets. But WordNet is fundamentally lexicalized: every semantic category is associated with at least one lexical item.

1.1.2 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. Named entities not in WordNet (most of them) were tagged with a coarse class.

Below, we will make use of the subset of Brown Corpus documents that are fully sense-tagged in SemCor 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 5.

# 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 5: 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.

1.1.3 Supersense Tags

In this work we will use the lumping scheme illustrated in figure 4c. Like NER, we seek to tag tokens with a coarse semantic class, regardless of whether those tokens are present in a lexicon. But instead of limiting ourselves to proper names, we use WordNet's **supersense** categories, 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]

The 25 noun supersense categories are:

(3) 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

§A gives several examples for each of the noun tags. There are 15 tags for verbs:

(4) 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

[1] WordNet synset entries were originally partitioned into **lexicographer files** for these coarse categories, which became known as "supersenses." The **lexname** function returns the supersense of a given synset.

principal, teacher, and student. A different sense of principal falls under the category POSSESSION.

1.2 Supersense Annotation

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 framework adapted from NER. (We return to the supersense tagging task in §1.3.) 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 decided to test whether the supersense categories offered a practical scheme for direct lexical semantic *annotation*, especially in a language and domain where no high-coverage WordNet is available.[2] Our annotation project for Arabic Wikipedia articles validated this approach.[3]

1.2.1 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 6.

1.2.2 Annotation Process

This project focused on annotating the free text Arabic Wikipedia data with the 25 noun supersenses of (3) and §A. The goal was to mark all common and proper

[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.

[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).

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

أقدم	المغرب	فاس	في	القيروان	جامعة	أن	القياسية	للأرقام	جينيس	كتاب	يعتبر
oldest		Morocco	in	Al-Karaouine	university	that	the-standard	for-records	Guinness	book	considers
		N:LOCATION		N:ARTIFACT			N:COMMUNICATION				

ميلادي	859	سنة	في	تأسيسها	تم	حيث	العالم	في	جامعة	.
	AD	year	in	established	was	where	the-world	in	university	
	N:TIME			N:ACT			N:LOCATION		N:GROUP	

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

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 7 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. Each supersense was assigned an 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 undergraduates 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 anno-

tation rounds were interspersed with discussions about the tagset. The annotation guidelines, §B, 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/

1.2.3 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. Additional details and analysis can be found in the paper (Schneider et al., 2012).

1.2.4 English Data

The methodology developed for Arabic supersense annotation was designed to be as general as possible. Only minor modifications should be necessary to adapt

the noun tagging conventions to a new language. We propose to conduct a small-scale supersense annotation effort on English text within domains not represented in SemCor (§1.1.2). This will be on top of the multiword expression annotations (§2.2.6). The primary methodological contribution of this will be an extension of the tagging guidelines from (Schneider et al., 2012) to include verb supersenses. The resource will be part of a multi-domain evaluation of automatic supersense tagging. We turn now to this NLP task.

1.3 Automatic Supersense Tagging

Here we discuss the current state of the art for automatic supersense tagging of English, which is based on a supervised statistical sequence model. Then we present techniques for addressing the more difficult problem of inducing supersense tags in Arabic text given only indirect evidence in the form of a small lexicon or automatically tagged machine translations.

1.3.1 Prior Work: English Supersense Tagging with a Discriminative Sequence Model

The model of Ciaramita and Altun (2006) represents the state of the art for English supersense tagging, achieving an F_1 score of 77% on the SemCor test set. It is a feature-based **tagging-chunking** sequence model learned in a supervised fashion. The goodness of the tagging $\mathbf{y}$ for the observed sequence $\mathbf{x}$ is modeled as a linear function (with real vector–valued feature function $\mathbf{g}$) parametrized by a real weight vector $\mathbf{w}$:

$$score(\mathbf{x}, \mathbf{y}; \mathbf{w}) = \mathbf{w}^\top \mathbf{g}(\mathbf{x}, \mathbf{y}) \tag{1}$$

The decoding problem given the weights $\mathbf{w}$ and input $\mathbf{x}$ is to construct the tag sequence $\mathbf{y}$ 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:[4]

$$\mathbf{g}(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{|\mathbf{x}|+1} \mathbf{f}(\mathbf{x}, y_j, y_{j-1}, j) \tag{2}$$

Many supervised learning algorithms are available for linear models (Smith, 2011). The input to such an algorithm is a training corpus that is a set of labeled

[4] 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 $\mathbf{x}$.

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

Figure 8: A supersense tagging shown with per-token BIO labels.

sequences, $\mathscr{D} = \{\langle \mathbf{x}^{(1)}, \mathbf{y}^{(1)} \rangle, \ldots, \langle \mathbf{x}^{(D)}, \mathbf{y}^{(D)} \rangle\}$; the output is the feature weight vector $\mathbf{w}$. Ciaramita and Altun (2006) use the **structured perceptron** (Collins, 2002), a generalization of the perceptron algorithm to sequences.

For Ciaramita and Altun (2006) and hereafter, sequences correspond to sentences, with each sentence segmented into words according to some tokenization. Any ordering or grouping of sentences (e.g., into documents) is disregarded by our models.

A **chunking** model is designed to group sequence elements (tokens) into units. The most popular flavor, **BIO chunking**, accomplishes this by assigning each token one of three labels: 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 of any chunk (Ramshaw and Marcus, 1995). Only contiguous chunks are allowed by this representation (we propose to relax this constraint in §2). A **tagging-chunking** model assigns a tag to each chunk as follows: each in-chunk label combines a chunk position (B or I) with a tag such as a supersense tag, and the decoding algorithm is constrained to only consider compatible label bigrams. For example, "non-initial word of a PERSON chunk" can be denoted as $_I$PERSON, and this is only allowed to follow $_B$PERSON or $_I$PERSON. With T tags, the number of labels is therefore $2T+1$, and the number of legal token label bigrams is $2T^2+5T+1$. At each time step the Viterbi algorithm considers all label bigrams, so decoding time is linear in this value and also linear in the length of the sentence.

The Ciaramita and Altun (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 the first WordNet sense of the current word. On SemCor data, the model achieves a 10% absolute improvement in F_1 over the first sense baseline.

1.3.2 English Supersense Tagging in New Domains

Ideally an English supersense tagger would perform well across a variety of topics and genres. The English supersense annotation effort proposed in §1.2.4 affords us the opportunity to assess the performance of supersense tagging in non-SemCor data. We will retrain Ciaramita and Altun's (2006) model on these data, and also experiment with adding **distributional cluster features** derived from large quantities of web data, as we have found such features to be worthwhile when tagging words in informal and noisy web text (Owoputi et al., 2012).

1.3.3 Arabic Supersense Tagging with Indirect Evidence

The supervised learning approach described in the previous section was made possible by SemCor, a 360,000 word sense-tagged corpus (§1.1.2). Unfortunately, for most languages—even languages with considerable corpus resources—no comparable semantically-annotated corpus is available. Such is the case of Arabic. In the absence of semantically-annotated corpus data, we turn to learning paradigms that exploit *indirect* evidence towards the semantics of words in context. Supersense tagging is an appropriate testing ground for this goal as it encodes major semantic category distinctions, covering most content words and generalizing across languages. In our experiments, *no* supersense-annotated training sentences in Arabic will be assumed; the small supersense-annotated dataset we have constructed for Arabic (§1.2) will be reserved for tuning and evaluation only.

Lexicon evidence. The first source of indirect evidence we use is **Arabic WordNet (AWN)** (Elkateb et al., 2006), a small lexical resource modeled after its English counterpart. Notably, many of the lexical entries in AWN are glossed with English WordNet sense mappings. From these mappings we can recover English supersense tags for Arabic lemmas, which we use to construct a supersense tagging lexicon.

We can then heuristically tag sentences using this lexicon alone; however, this faces two major limitations. First, without generalizing beyond the lexicon, noun and verb coverage will be poor. (On the development set this gives an F_1 score in the low 20% range.) Better generalization should be attainable with graph-based semi-supervised learning (Das and Petrov, 2011), which hypothesizes supersense tags for new words on the basis of labeled seeds (AWN words with supersenses) and a semantic similarity metric between word types. Second, unlike the fully supervised supersense tagging models above, it does not allow for neighboring context to inform the labeling of each token. A solution is to learn Das and Petrov's (2011) unsupervised sequence tagger—which uses the expanded (semi-supervised) lexicon for constraints—on Arabic Wikipedia data.

Cross-lingual evidence. If the Arabic Wikipedia sentences were parallel with English, we could supersense-tag the English side with Ciaramita and Altun's 2006 system and project its tagging via word alignments to Arabic, as has been done with named entities and word sense annotations in previous work (Yarowsky et al., 2001; Diab and Resnik, 2002). However, in this case we are faced with non-parallel data.

One strategy would be to project automatic predictions across an unrelated parallel corpus, and then train a monolingual Arabic supersense tagger on these projections. Downsides of such an approach are that (a) both the source-language tagging and the projection process are extremely noisy; and (b) the parallel data would be in a different domain. In preliminary experiments we found that this technique was actually worse than the AWN heuristics.

A second idea is to elicit (noisy) English translations from an MT system, automatically tag those with supersenses, and then project the tags onto the Arabic sentence. Preliminary tests of this technique with this technique were positive, indicating it is more effective than the AWN heuristics.

Combination. If both of the above ideas show promise, the lexicon-constrained unsupervised learning could be conducted on Arabic Wikipedia data and incorporate cross-lingual features acquired using machine translation and bilingual projection.

This work is ongoing; we expect experimental results within the next few weeks.

2 Open-ended Identification of Multiword Expressions

Thus far, we have considered the supersense tagging scheme for nouns and verbs. That scheme reflects the choices of WordNet lexicographers, capturing some kinds of multiword units (especially names and other nominal expressions, discussed below). In general, however, it is worth developing a resource-agnostic understanding of which multiword combinations cohere strongly enough to count as units. The many kinds of putative MWEs and gradient lexicality make this difficult to do even for specific constructions, let alone in a general-purpose manner. Rather than search for clear-cut tests of MWE-hood, we therefore endeavor to provide brief exemplar-based guidelines to annotators and then set them loose on free text. This section motivates and describes this approach and proposes techniques for modeling and evaluating multiword expressions at the token level.

2.1 What is a Multiword Expression?

Much ink has been spilt over the definition of multiword expressions/units, idioms, collocations, and the like.[5] 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. 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.*

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.[6] 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).

[5] 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).

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

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).

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[7]	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 *street-light*/ *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.

In this work we will use the term **multiword unit (MWU)** for any two or more words that function together as a **multiword expression (MWE)** or **named entity (NE)**.[8]

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

[8] We may also include **value expressions** like dates, times, and monetary quantities in our definition of multiword unit; in fact, many of these are tagged by existing NER systems (e.g. Bikel et al., 1999).

2.1.1 Polysemy

Figure 9 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).[9] 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:

1. {V:STATIVE} form or compose
2. {V:CREATION} devise or compose
3. {V:POSSESSION} do or give something to somebody in return
4. {V:SOCIAL} make up work that was missed due to absence at a later point
5. {V:CREATION} concoct something artificial or untrue
6. {V:CHANGE} put in order or neaten
7. {V:STATIVE} adjust for
8. {V:COMMUNICATION} come to terms
9. {V:BODY} apply make-up or cosmetics to one's face to appear prettier

Some of these senses are radically different: making up a story, a bed, a missed exam, one's face, and (with) a friend have very little in common![10] Reassuringly, the supersenses attest to major differences, which suggests that the MWU grouping and supersense tags offer complementary information (we propose in §4 to exploit this complementarity in a unified model).

2.1.2 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). 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).[11] An additional 6368 multiword expression mentions (3047 types) are annotated, encompassing 13,785 words (6% of the data).

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

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

[11] For the type counts in this paragraph, mentions were grouped by their lowercased surface string.

'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*

particle verbs (2): *from what Harry could make out* (make out = 'reckon'), *make up to well-connected people* (make up to = 'cozy/kiss/suck up to')

light verb with eventive noun (11): *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*

miscellaneous multiword expressions (11): *make mistakes* (×2), *make different arrangements, make sure* (×5), *make do, make sense, make any sign of recognition*

Figure 9: Occurrences of the bare verb make in a small text sample.

About 87% of these mentions (and 87% of types) are tagged with a WordNet sense.[12] All told, 8% of tokens in **ParsedSemCor** belong to a SemCor-annotated MWU, with a 3-to-1 ratio of MWEs to multiword NEs.

2.1.3 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)

[12] 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.

- 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); 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 10);[13]
- When an internal modifier such as an adjective, adverb, or determiner is present (the **modifier gap** column of figure 10);
- 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>).[14]

To identify gappy MWEs in the **ParsedSemCor** collection, including those in figure 10, we extracted the sense-tagged items for which the number of words in

[13] This is not to suggest that the syntactic arguments MWEs always fall between lexicalized words: with prepositional and particle verbs, 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.

[14] 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 collocations. 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,*	*runs ⟨ too great ⟩ a risk,*
	made up ⟨ her ⟩ mind	*paid ⟨ no ⟩ attention*
Verb-PP	*put ⟨ many persons ⟩ to death*	*falls ⟨ hopelessly ⟩ in love*
Verb-adverb		*stood ⟨ very ⟩ still*

Figure 10: Examples of gappy MWEs in the SemCor corpus.

the lemma differed from the number of words in the tagged surface span—this usually indicates a gap.[15] 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. One of our pilot annotation studies (below, §2.2.5) is designed in part to compare the MWE coverage of SemCor annotations versus our annotators' judgments.

One final point worth making is that multiword expressions create syntactic ambiguity. For example, somone might *make [up to 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 §2.3).

[15] 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.

2.1.4 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.

2.2 Multiword Annotation Paradigms

2.2.1 Prior Work

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.[16] On the other, studies in the MWE literature (and of lexical semantics more broadly) tend to (a) build lexicons capturing corpus-level generalizations, or (b) use a specific class of expressions in a known lexicon to reason about tokens in sentence context. In the case of (a), there is no need to commit to any token-level analysis; in the case of (b) there is not an expectation that the lexicon will provide good coverage of every sentence. Without getting into the details of automatic multiword analysis tasks here just yet (they will appear in §2.3), we take the position that a comprehensive treatment requires corpora annotated for a broad variety of multiword expressions.

To our knowledge, only a few corpora approach this goal:

SemCor. As discussed above, SemCor includes many multiword expressions, most of which are tagged with WordNet senses. Exactly how the lexicographic decisions were made is unclear, but SemCor seems to prioritize complex nominals and particle verbs over other kinds of multiword constructions.

[16] 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.

The Prague Dependency Treebanks. The **Prague Dependency Treebank (PDT)** (Hajič, 1998) and the **Prague Czech-English Dependency Treebank (PCEDT)** (Čmejrek et al., 2005) 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. 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 may be possible to automatically extract shallow multiword groupings given that we are not seeking to model their syntax.

The French Treebank. The French Treebank specially designates a subclass of MWEs, which it terms *compounds* (Abeillé et al., 2003, p. 172). This category appears to be rather narrow, excluding (for example) prepositional verbs (Abeillé and Clément, 2003, p. 53):

> On ne considère pas les combinaisons clitiques-verbes comme formant un V composé, même pour les clitiques intrinsèques (*s'apercevoir de*, *en avoir assez de*...)

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).
>
> *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*

The French Treebank has been used to train and evaluate multiword expression identification systems, but to our knowledge, none of this work has attempted to model the gaps due to internal modifiers. We address this issue in §2.3 below.

2.2.2 Towards an Open-ended Paradigm

We have begun studying how annotators respond when given a text (but no dictionary) and asked to find multiword expressions. The difficulty of achieving high

inter-annotator agreement for dictionary-free labeling of MWEs has been noted anecdotally (e.g., Piao et al., 2003) but (to our knowledge) never quantified at the token level.

2.2.3 Pilot Annotation Study 1

The purpose of this study was to test the viability of a simple token-grouping scheme for multiword expression annotation. We wanted to know:

- How do annotators vary when they have received minimal instruction in the task? Are there systematic kinds of disagreement that suggest revisions to the guidelines?
- How much time is required for the task?
- What kinds of gappy expressions are found in practice?

SETUP

Participants. There were four annotators (the author and three colleagues), all of them graduate students in LTI. All are native speakers of American English.

Task. Participants were directed to a website which provided sentences to annotate. The instructions on the website are reproduced in full below (§C); part of the explanation is as follows:

> You are given a (pre-tokenized) English sentence. **Your mission is to partition the sentence into lexical expressions, each consisting of one or more tokens.**
>
> Most tokens will remain as they are, but some belong to a multiword expression and should therefore be joined to other tokens. What is meant by *multiword expression*? Intuitively, **any expression that (a) is a proper name, or (b) ought to be listed in a dictionary because its structure and/or meaning and/or frequency are not predictable solely on the basis of its components.** (This definition includes, but is not limited to, idioms and noncompositional expressions.)

The instructions include a list of contiguous and gappy MWE examples, as well as the sample annotated sentence:

```
It even got_ a little _worse during a business_trip to the city , so
on|1 the advice|1 of a friend I set_up an appointment with True_Massage .
```

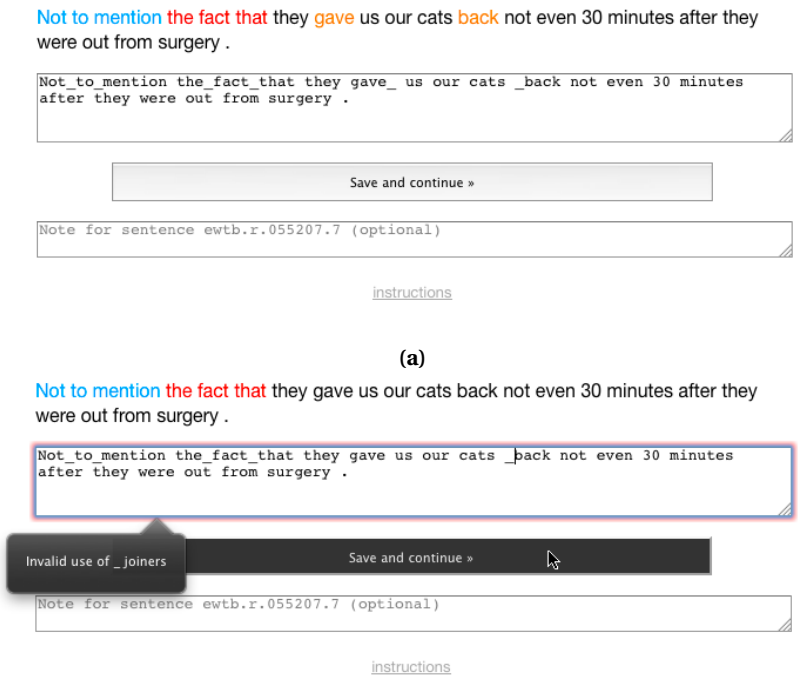

Figure 11: Multiword expression annotation interface.

The annotation scheme allows for arbitrary groupings of words into multiword expressions, so long as no word belongs to more than one expression. Aside from §C, annotators received no additional information about the annotation scheme, and were asked not to confer with one another about the task.

Interface. The annotation interface, figure 11a, consists of a webpage with a text input box for the marked-up sentence. Above this is a rendered version of the sentence illustrating the annotated groupings by color-coding the tokens. The rendering is updated as the contents of the text box are modified. Client-side input validation ensures that the words themselves do not change and that the multiword markup is valid (figure 11b depicts an error state resulting from an incomplete gappy expression).

N	IAP	①	②	③	④
128	①		.65	.74	.69
124	②	.62		.80	.80
102	③	.59	.60		.71
80	④	.43	.52	.56	

IAF

$F_1(①,②) = .63$ $F_1(①,③) = .65$
$F_1(①,④) = .53$ $F_1(②,③) = .66$
$F_1(②,④) = .63$ $F_1(③,④) = .63$

Figure 12: Total mentions annotated and inter-annotator precision and F_1 scores for Pilot Study 1. For instance, in the top row, $|① \cap ②|/|②| = .65$ ($|②| = 124$). $F_1(①,②) = 2 \cdot IAP(① \mid ②) \cdot IAP(② \mid ①)/[IAP(① \mid ②) + IAP(② \mid ①)] = 2(.65)(.62)/(.65 + .62) = .63$.

Source Data. The sentences for this study were drawn from documents in the reviews portion of the English Web Treebank (Bies et al., 2012). The online reviews genre was chosen for its informal style in which idiomatic expressions are frequent. We used the first 100 ASCII-ified and tokenized sentences of a document-level split[17] of the corpus, amounting to a total of 1321 words from 24 documents. Items (sentences) were presented to annotators one at a time in their original order. Every participant annotated all 100 items. Annotators did not have the opportunity to review or revise previous annotations.

Results

Below, we use ①, ②, ③, and ④ to denote the respective annotators. ① corresponds to the author, who designed the task and selected the source data.

Time. Page load and submit times were recorded for each annotated sentence. Median sentence-annotation times (in seconds) were as follows: $t(①) = 14$, $t(②) = 23$, $t(③) = 14$, $t(④) = 9$. Overall, the median time to annotate a sentence was 13 seconds.

Inter-Annotator Agreement. The 4 annotators found an average of 109 multiword mentions (tokens). Figure 12 gives a breakdown by annotator. While there is some variation (e.g., ④ spent less time and was more conservative), the average pairwise strict inter-annotator mention F_1 score was 62%—surprisingly high given the limited nature of the instructions.

[17] Specifically, this split consisted of all documents with IDs of the form `xxxx0x` (`x` being any digit).

190 mentions were given by one or more annotators. This breaks down into 41 for which all annotators were in agreement, 39 marked by 3/4 annotators, 43 marked by 2/4 annotators, and 67 marked by only one annotator. Thus by the strict (exact-match) criterion, there was a 67/190 = 35% "non-agreement" rate. Given that some annotator found a mention, the expected number of other annotations of that mention was 1.3.

Inter-Annotator Overlap and Disagreements. 47 of the mentions given by one or more annotators overlapped partially with some other mention from another annotator.[18] Merging overlapping mentions yields 23 groups. We categorized the disagreements within these groups: notably, 3 groups concerned article inclusion (*(the) hustle and bustle, make (a) order, (a) hour and a half*); 5 concerned complex nominals (e.g. *pumpkin spice (latte), (criminal) defense attorney, mental health (counselor), low oil pressure light* vs. *oil pressure*); 4 concerned verb inclusion (e.g. *(has) much to offer, (do…) good job*); and 3 concerned preposition inclusion in prepositional verbs (*when it came (to), make up (for), spreading the word (about)*). In a couple of cases there were multiple discrepancies: two annotators provided *had a problem* while the other two marked *had… problem with*; and for the phrase *the number 9 Bus route*, one annotator had only *Bus route*, two had *number 9* and *Bus route* as separate expressions, and one marked the full phrase as a single expression.

Annotator ② consistently attached tokenized clitics like *'s* and *n't*, whereas the others did not mark them as multiwords. Clarifying how to handle these in the guidelines should improve inter-annotator consistency.

Merging partially overlapping mentions and removing clitic attachments leaves 166 mention groups and only 32/166 = 19% single-annotation mentions. These are listed in figure 13. Interestingly, a plurality involve prepositions; we expect that improving inter-annotator agreement for such cases will go hand in hand with developing a systematic treatment of prepositions, as proposed in §3.

Gappy Expressions. Figure 14 lists the gappy expressions marked in the study, along with the annotator(s) responsible for each. The number of words in the gap ranged from 1 (in 10 of the 16 mentions) to 5 (2 mentions).

Note that in two cases the gap between two parts of an expression included a contiguous multiword expression (*Ford Fusion* and *rear window*). Annotators did

[18] The mentions from any single annotator were required to be disjoint.

Involving Prepositions (13): *work with, variety of, lots of, using on, for years, style of, sent… to, sensitivity for, capable of, Even if, damage to, something as simple as, years of*
Verbal Support (8): *getting it done, answered… phone, had… spayed, got infections, did… surgery, direction… take, realized… mistake, had… replaced*
Other Nominal (11): *Sheer contrast, Stationery store, whatever else, no doubt, the fact that, not event, great ear, level of skill, place of beauty, A couple, All of this*

Figure 13: Mentions from only one annotator.

they gave us our cats back ①②③④	*the vet they sent us to was* ②
to let her know that Yelp may ①②③	*the vet that did the surgery* ①
will work every possible legal " angle " ①②③	*think they would do a good job but* ①
had taken her '07 Ford Fusion in for ①②③	*direction they want their lessons to take* ①
to learn more about my practice ①③	*To make a order you may have* ①
never had a problem with their ①②	*Once they realized their mistake they* ①
the boy who answered the phone ①	*they will overcharge you for just* ①
I had my cat spayed ①	*had my bmw z3 rear window replaced* ①

Figure 14: Gappy mentions (with some context) and annotators. Annotator ① was the most liberal about marking gappy expressions (15); ④ was the most conservative (1).

not ever nest or interleave two *gappy* expressions, though the annotation scheme allowed them to do so.

Discussion

Overall, we were pleasantly surprised with the level of agreement given the underspecification of the task. To better understand the results we decided to undertake two additional pilot studies.

2.2.4 Pilot Annotation Study 2

In this study we investigate the intuitions of *nonnative speakers* at multiword expression annotation. We use the same setup as Pilot Study 1, but different participants (LTI graduate students for whom English is a second language). The goal is to determine if their responses differ noticeably from native speakers', and if so whether the differences reflect systematic biases. This study is currently underway.

2.2.5 Pilot Annotation Study 3

This study has two goals: (a) testing whether revisions to the annotation guidelines improve inter-annotator agreement relative to Pilot Study 1; and (b) assessing agreement between participants in the study vs. SemCor annotations (§1.1.2). Participants will be the Pilot Study 1 annotators. Part of the data sample (web reviews) will be repeated from Pilot Study 1, reflecting the first goal; and part of it will be new (from SemCor, per the second goal). This study is ongoing.

2.2.6 New Datasets

Once a stable annotation protocol has been developed, we will apply it to the reviews subcorpus of the English Web Treebank (Bies et al., 2012) (50,000 words). University of Pittsburgh undergraduates majoring in linguistics will be enlisted as annotators (they will be compensated financially and/or with internship credit). In addition, we hope to be able to use the MWE annotations in SemCor and the English side of the Prague Czech-English Dependency Treebank (§2.2.1). Some supplementary annotation in these datasets may be required to resolve inconsistencies between different conventions.

2.2.7 Leveraging Multiple Annotations

Traditionally in NLP it is assumed that the goal of human annotation is to create a single "gold-standard" dataset against which systems can be evaluated. Yet there are contexts in which raw labels from annotators cannot necessarily be trusted as gold-standard. A line of research stimulated especially with the advent of crowdsourcing has developed methods for analyzing annotation quality and individual annotator biases (Dawid and Skene, 1979; Wiebe et al., 1999; Snow et al., 2008; Carpenter, 2008; Munro et al., 2010). In the case of crowdsourcing, quality assurance is necessary because the annotators are *untrusted*—they may not understand the task or may not take it seriously.

Here we face a slightly different problem. First, we are dealing with structured annotations, not independent labels. Second, our annotators are trusted but the task is open-ended enough that they might reasonably be expected to come to different conclusions. While training, discussion, and guidelines refinement should minimize confusion over the annotation standard, we expect the inherently statistical and "fuzzy" nature of collocation and idiomaticity will leave cases of legitimate

disagreement (perhaps attributable to idiolect). Rather than ask the original annotators or a third party to adjudicate cases of disagreement, we hypothesize that models can take advantage of inter-annotator variability to learn more robust generalizations.

Instead of a producing a single adjudicated consensus, we will elicit multiple annotations (at least 3) for each item and experiment with the following supervised learning regimes:

- Train on all annotations. That is, the learning algorithm will see the same sentence multiple times, with potentially different labelings. The data points could be weighted to account for known annotator biases.
- Train with a loss function that imposes a cost on incorrect predictions where the cost function considers multiple annotations (cf. Mohit et al., 2012, which uses a cost function for sequence tagging). Intuitively, a predicted expression that is at least partially consistent with at least one annotation should cost less than a wholly unsupported prediction.
- Train on an automatically-inferred consensus annotation of the corpus. The consensus could be produced on a sentence-by-sentence basis by searching for the labeling maximally agreeing with the human annotations (under some agreement measure[19]).[20]

Each of these systems could then be evaluated on held-out test data prepared using the same criteria as the training data.

2.3 Automatic Identification of Multiword Expressions

2.3.1 Prior Work on Processing Multiword Expressions

There is a sizeable literature concerning multiword expressions in NLP: automatic techniques have been developed to create multiword lexicons from raw corpora (**extraction**), recognize MWEs in context (**identification**), infer their internal syntactic or semantic structure at the type level (**interpretation**), and classify an MWE's

[19] E.g., average of the pairwise F_1 scores with the individual human annotations. F_1, however, does not factor and is therefore not amenable to an exact dynamic programming solution.

[20] A baseline strategy would be to choose the single human labeling that is in highest agreement on average with the other human annotations—either on a sentence-by-sentence basis, or for the corpus as a whole.

meaning in context (**disambiguation**).[21] Some of these studies have targeted NLP applications such as machine translation. Baldwin and Kim (2010) and Ramisch (2012) offer extensive reviews.

Here we seek a general-purpose solution to the identification problem. Many identification approaches assume an MWE lexicon as input, and heuristically match n-grams against its entries. Sometimes this is followed by a classification step to determine if the candidate expression is being used literally or idiomatically (Birke and Sarkar, 2006; Hashimoto and Kawahara, 2008; Fazly et al., 2009; Boukobza and Rappoport, 2009; Li and Sporleder, 2010; Michelbacher et al., 2011; Fothergill and Baldwin, 2011, 2012). For morphologically and syntactically flexible expressions, however, this may not be sufficient. Other approaches use *syntax* or integrate MWE identification within syntactic parsing, in research reviewed by Seretan (2011) as well as research conducted more recently by Green et al. (2011, to appear) and Constant et al. (2012). But the resources and computing power required for full syntactic parsing are not always available in practice.

Most relevant here are approaches that cast MWE identification as a *sequence labeling* problem. Diab and Bhutada (2009) trained an SVM-based sequence tagger to detect literal and idiomatic verb-noun constructions represented with the BIO chunking scheme (cf. §1.3.1). In their formulation, the chunks are contiguous: determiners and other modifiers between the verb and the noun are included. Their model included features over context word n-grams, character n-grams, POS tags, and lemmas. They also used features based on the output of an NER system—an ablation study proved these features to be most useful. Subsequent studies have used CRFs for supervised BIO chunking of MWEs, namely noun compounds (Vincze et al., 2011) and verb-noun constructions (Vincze, 2011) in English, reduplicated MWEs in Manipuri (Nongmeikapam et al., 2011), and MWEs in French (Constant and Sigogne, 2011; Constant et al., 2012). Ciaramita and Altun (2006) (discussed in §1.3.1 above) similarly train a supervised sequence tagger for English lexical units in SemCor, some of which are MWEs.

To our knowledge, the only work on statistical identification of *gappy* multiword expressions was the generative model of Gimpel and Smith (2011). Their model assigns a "color" to each word of the sentence, such that all words labeled with the same color are interpreted as belonging to the same expression ("pattern"); it is learned in an *unsupervised* fashion, with priors on the inferred expression lexicon

[21] The extraction and identification tasks are sometimes grouped together under the label **acquisition** (Ramisch, 2012, p. 50).

encouraging a reasonable number of patterns. A bilingual variant assigns colors to both source and target tokens of word-aligned parallel sentences. Rather than seeking to match human annotations, Gimpel and Smith's quantitative evaluation embeds the gappy pattern model within a machine translation system, achieving modest BLEU score gains over a baseline.

2.3.2 Towards Discriminative Gappy Chunking

In this section we seek to incorporate gaps into the supervised chunking regime that has been used to identify the sorts of MWEs that human annotators provide. Importantly, we aim to identify all MWEs in a given sentence—not just a single variety in the manner of some previous work. Results from the pilot annotation study in §2.2.3 indicate it is necessary to support limited nesting of MWEs: specifically, contiguous MWEs may fall within a gap, as in the following sentence that was excerpted in figure 14:

(5) My wife had taken her '07 Ford Fusion in for a routine oil change .
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14

As the standard BIO chunking scheme can only encode contiguous chunks, we need to alter the representation to accommodate the gappy particle verb. There are many possible solutions. The most incremental change is to assume that there are no more than two levels of chunk structure—that is, gaps are allowed for top-level expressions, but not for expressions falling within the gap (so structures such as $a\langle\ b\langle\ c\ d\ \rangle e\ \rangle f$ would be disallowed). We further assume that gappy expressions never interleave (prohibiting e.g. $a\ B\ c\ d\ E$, where capitalization indicates chunk membership).[22] All of the annotations from the pilot annotation study are consistent with these constraints.

Based on these assumptions, we propose the following **BbIiOo** scheme:

(6) My wife had taken her '07 Ford Fusion in for a routine oil change .
O O O B o o b i I O O O B I O

Here we have introduced three new lowercase labels to encode the chunking of the words *inside the gap*. The (capital) letter following the gap is to be read as if the

[22] Throughout we assume that gaps are never empty and are always flanked by words in the expression. We will not impose any limits on the number of gaps in an expression, so long as two gaps are not adjacent.

gap were not present. As before, I is prohibited from occurring at the beginning of the sentence or immediately following O. Additionally, i must not follow o, and no lowercase label may begin or end the sentence or be adjacent to O. The regular expression $(O|BI*\ (o|bi*)*\ I*)*$ describes the language of valid sequences.[23]

When developing a sequence model for a task, it is not enough to consider the formal constraints on the output structure—we must take into account the *features* that can be represented without sacrificing algorithmic correctness or efficiency. In a first-order discriminative model like that of §1.3.1, a **local feature** may consider a label bigram as well as the observations $\mathbf{x}$ (all of the words of the sentence and any auxiliary information from preprocessing). So, for example, we can represent a feature that fires for the bigram *oil*/B *change*/I, and another feature that fires for the bigram *Ford*/b *Fusion*/i. We can also specify a set of features that ignore the case of the labels, e.g. *Ford*/$\{B,b\}$ *Fusion*/$\{I,i\}$—this is likely desirable as examples of MWEs within gaps are expected to be sparse. Features like *taken*/$\{B,I\}$ $*$/$\{b,i,o\}$ would fire whenever the word *taken* is followed by the start of a gap. Finally, *taken*/$*$ $\ldots$ $*$/$\{b,i,o\}$ *in*/I ("the word *in* occurs at least two words after *taken* and resumes a gappy MWE") would be a local feature.

What would not be a local feature is anything specifying two nonadjacent words as belonging to the same expression. Such a feature may be essential to model gappy MWEs. A variety of approximate decoding techniques have been used in NLP to make predictions with **nonlocal features** or constraints, including beam search, reranking, integer linear programming, cube pruning (Chiang, 2007), and stacking (Cohen and Carvalho, 2005). But perhaps the necessary features can be made local by enhancing the state space of the labels. Based on the observation that most gappy MWEs in practice contain a verb, we propose to extend the BbIiOo scheme by attaching indices pointing to the MWE's verb:

(7) My wife had taken her '07 Ford Fusion in for a routine oil change .
O O O B-3 o-3 o-3 b-3 i-3 I-3 O O O B I O

Crucially, the 3 indexing the verb *taken* is copied through the gap. As a result, the label following the gap (I-3) can be locally constrained to have the same index. Because the verb's index is now in the label for the *in* token, features local to the particle can consult the observations $\mathbf{x}$ to specify that *taken*$\ldots$*in* occur within the same expression. Requiring the specified index to belong to the verb, regardless

[23] Relaxing the above assumptions would lead us to a larger language class—such as the context-free family—requiring sacrifices in computational complexity.

of the verb's position in the expression, allows the phrase *direction they want their lessons to take* (figure 14) to be handled as well without exploding the search space: in decoding, only verbs in the sentence need to be considered as possible MWE "anchors".[24] Unadorned B and I will continue to be allowed for expressions not containing verbs. The number of possible label bigrams is thus quadratic in V, the number of verbs in the sentence. More precisely, if sentence boundary constraints are ignored, there are $(V+1)(2(V+1)+22)+1 = 2V^2+26V+1$ label bigram types. Decoding time will be linear in this value.

2.3.3 Proposed Features and Experiments

The model will be trained and tested on our benchmark dataset of English treebank data augmented with open-ended MWE annotations (§2.2.6) as well as the French Treebank, which includes some MWE annotations (§2.2.1).

Our model's features will take inspiration from previous MWE identification work (Diab and Bhutada, 2009; Constant et al., 2012). **Basic features** will consist of token and character n-grams—including features like the previous verb token—as well as automatic part-of-speech tags, lemmas, and named entity labels from existing tools. After Constant et al. (2012), we will incorporate two sets of **exogenous features**. Some of these will leverage existing lexicons, namely WordNet and existing MWE-specific datasets like (Cook et al., 2008; Simpson-Vlach and Ellis, 2010; Martinez and Schmitt, 2012). The others will encode association measures computed from raw corpora, which are often indicative of MWE-hood and widely used for MWE extraction (Pecina, 2010).

We are especially interested in the following questions:

1. To what extent will performance degrade without *syntax*?
2. Can *gappy* MWEs be handled in a sequence model without sacrificing exactness or efficiency?
3. How should *multiple annotations* be combined when training a chunking model?

The first question explains our choice of treebank data. To assess the impact of syntactic features in MWE identification, we will experiment with three feature sets: (a) the basic and exogenous features only (baseline); (b) baseline plus syntactic features derived from the gold parses; and (c) baseline plus syntactic features

[24] If verbs are used, no additional annotation is necessary to choose the anchors. (Part-of-speech tagging as preprocessing is assumed.)

the	11,953	**in**	3483	his	2007	**as**	1274	she	957	**by**	768	there	651
,	11,637	he	3340	had	1766	is	1247	him	925	**from**	763	or	649
.	10,686	was	2683	i	1592	you	1150	her	916	this	739	an	643
of	5469	“	2293	**for**	1434	**at**	1078	they	887	n’t	731	all	635
and	5433	”	2272	**with**	1400	but	1017	?	817	have	717	**out**	585
to	4797	that	2210	’s	1288	not	984	-	792	were	664	we	559
a	4531	it	2087	**on**	1285	be	969	would	784	one	655	said	548

Figure 15: Counts of the top 49 most frequent words in **ParsedSemCor**. Prepositions are bolded; others in the top 100 include up (#51), into (#65), about (#67), than (#77), down (#79), over (#80), back (#82), and before (#93). (The most frequent noun, time, ranks 73rd with a count of 332.)

derived from a parser. We hypothesize that the condition with gold parse features will result in the best performance, but it may be a modest gain over the syntax-free baseline.

A second dimension to explore is the model’s ability to predict gappy expressions, which are beyond the expressive power of the BIO labeling scheme. The previous section proposed two alternative representations: BbIiOo with and without verb anchors (the anchors allow for more expressive features at the cost of increasing the search space). These two conditions will be compared against a baseline in which gappy MWEs have been removed.

§2.2.7 presents an experimental setup for the multiannotator case.

Performance measures for the experiments will be: token-level accuracy, precision, recall, and F_1; strict mention-level precision, recall, and F_1; and training and test runtimes. Error analysis will examine the model’s behavior for the major classes of English multiword expressions (complex nominals, verb-particle constructions, verb-noun constructions, etc.).

3 Functional Tagging of Prepositions

Prepositions are perhaps the most beguiling yet pervasive lexicosyntactic class in English. They are everywhere (figure 15); their functional versatility is unrivaled and largely idiosyncratic (8). 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 (8):

(8) a. My cake is **to** die for. *(nonfinite verb idiom)*
b. If you want I can treat you **to** some. *(prepositional verb idiom[25])*
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)*
h. and put it in the oven for 30 **to** 40 minutes *(temporal range endpoint)*
i. so you will arrive **to** the sweet smell of chocolate. *(FrameNet* Coordinated_event *(Ruppenhofer et al., 2010)?)*
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 (8h). 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 (8g). As we have seen in §2 above, prepositions play a key role in multiword expressions, as in (8a), (8l), the prepositional verbs in (8b) and (8k), and arguably (8e). Other prepositions can be intransitive: *brought **down** the bed*/ *brought the bed **down*** (non-idiomatic verb particle; Huddleston, 2002, p. 280), *take **down** the message*/ *take the message **down*** (idiomatic verb particle), and *the car broke **down*** (verb–intransitive preposition idiom; Huddleston, 2002, p. 285).

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, 2006), 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). Here we opt to represent and model prepositions from the combined perspectives of WSD and multiword expressions, and explore the relevance of this approach to two applications (§5 and §6).

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

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

The Penn Treebank. As detailed in (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 builds on top of FrameNet annotations. The data for the SemEval-2007 shared task on preposition WSD were drawn from TPP, consisting of 34 prepositions with a total of 332 senses attested in over 25,000 sentences (Litkowski and Hargraves, 2007). TPP now incorporates additional prepositions and resources (Litkowski, 2012).

Studies in preposition sense disambiguation have evaluated systems against one or more of the above resources (O'Hara and Wiebe, 2003, 2009; Ye and Baldwin, 2007; Dahlmeier et al., 2009; Tratz and Hovy, 2009; Hovy et al., 2010, 2011). Unfortunately, all three 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 (see §5). 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 (cf. §1.1.2). A **token-driven** approach would be more in line with the philosophy advocated here for lexical semantic annotation and modeling.[26]

[26] 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.

	SPATIAL/MOTION	TEMPORAL	COMMUNICATN	TRANSFER	MEASURMNT	COMPARISON	CAUSATION
part1	figure/traj	event	topic/content	given	measured	compfigure	agent/witness
part2	colocatedthing	simulevt		exchngdfor	measured2	similarthing	patient/affectd
state	location/ground	time	corpus		scale		manner
path	trajectory	timespan	medium	medium	range		instrument
source	origin	starttime	speaker	giver	lowerbound	lesservalue	cause/conditn
goal	destination	endtime	audience	recipient	upperbound	greatervalue	purpose/result
extent	size/distance	duration		price	amount	difference	

Figure 16: Coarse semantic senses for prepositions (preliminary). For convenience they are organized into generic "scenes" and "roles". Additional senses like COMITATIVE may be necessary.

We therefore plan to develop a medium-grained inventory of preposition functions in the spirit of supersense tags (§1), and to deploy it for annotating the English datasets proposed in §2.2.6. The preposition sense inventory will resemble figure 16, though further analysis and refinement is needed. It takes inspiration from an ongoing corpus creation project for German preposition senses (Müller et al., 2010, 2011). Like their approach, our sense inventory will be cross-cutting (unlexicalized), owing to the fact that certain senses can be realized by multiple prepositions—for example, both to and for can be used to mark a PURPOSE:[27]

(9) a. We bought a new TV (in order) **to** watch the election coverage.
b. We bought a new TV **for** (the purpose of) watching the election coverage.

An important and novel aspect of our approach will be the use of multiword expression annotations to inform preposition annotations. In a nutshell: if a preposition lies within an MWE, the annotator can elect not to tag it with a semantic sense. Otherwise, an explicit annotation is required—either a semantic sense from a predefined list of about 20–40 (which is expected to account for about 80% of the instances), or an OTHER category for rare meanings like (8i), or a SUPPORT category for purely syntactic occurrences like (8g). This scheme implies a trifurcation of **preposition functions**: a group of freely combining **semantic senses**, the **selectional** (idiomatic) uses, and those that serve as **syntactic support**. Our hope is that the "grab-bag" categories OTHER and SUPPORT will streamline a first-pass annotation while leaving open the possibility of revisiting difficult cases in subsequent passes.

[27] Of course, it is possible to paraphrase the sentences in (9) without a preposition: *We bought a new TV **so** (that) we can watch the election coverage.* This suggests a certain amount of semantic overlap between prepositions and clausal conjunctions.

We will aim to annotate the prepositions in 50,000 word selections of the three datasets with MWE annotations (§2.2.6). Given these data, a straightforward modeling approach will be to train a supervised discriminative classifier in the manner of Hovy et al. (2010). As with MWE modeling, we will examine the effect of syntactic features, which Hovy et al. (2010) generally found to give slight gains over simply using lexical and POS features. If possible we will also train and evaluate a German preposition model on the corpus of Müller et al. (2010).

Cross-lingual 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; Zelinksy-Wibbelt, 1993) but is of practical interest as well, especially when it comes to machine translation (see §6). Here we propose to investigate whether features from parallel data can help bootstrap a **monolingual** preposition function classifier. The foreign word aligned to the English preposition would in many cases provide disambiguating context. For example, two of the French equivalents of for are the prepositions *pour* (GOAL, DESTINATION) and *pendant* (DURATION).

How can parallel data be exploited to improve a supervised model trained on non-parallel data? After training on a small annotated dataset in English, we might then self-train on the English side of parallel sentences to learn weights for the cross-lingual features in addition to the monolingual ones. These new features would provide "scaffolding" which could help learn a better classifier for prepositions in parallel context. Finally, after the scaffolded model makes new predictions on the parallel data, it could "wean" itself off of the scaffolding features by self-training with only the monolingual features. As far as we are aware, this variant of self-training has never been tried[28] and could result in a better monolingual preposition classifier without any additional annotator effort.

4 A Unified Approach to Token Lexical Semantics

Thus far, we have considered three avenues to analyzing the chunking and semantic categorization of lexical expressions. It is best to think of these approaches not as discordant, but in harmony. In fact, the sequence tagging-chunking representations advanced above can be integrated.

Figure 17 sketches how this can be done for two sentences. There are two levels of analysis: a chunking level (including single-word, contiguous multiword, and gappy multiword chunks) and a tagging level (where every chunk receives 0 or 1 tags). Only a few tokens (e.g. punctuation, determiners, pronouns) remain unanalyzed. Multiword units like a little that are not headed by a noun, verb, or preposition are chunked but not sense-tagged. Coarse noun and verb senses use WordNet supersense categories (§1); those that are also named entities are marked with the -NE flag. Note that because sense tagging is at the lexical expression level, the semantics of corner and advice (both of which could be analyzed with noun supersenses) are subsumed by their containing expressions. Preposition functions are tagged as described in §3.[29]

Combining the results of the annotation projects discussed above will yield a corpus of sentences fully annotated with the integrated representation. From these data we can learn and evaluate a unified lexical semantic analyzer in much the same way as the aforementioned models.

Computationally, a unified model will have a larger search space than any of the component models. However, the situation should not be as bad as it first appears because the POS tags (from preprocessing) can be used as a filter, limiting the number of possibilities for each token. If efficiency remains a challenge, alternate dynamic programming strategies that have been shown to produce speedups with large label sets (Kaji et al., 2010) can be tried.

A minute later they turned the corner into the side street where the
N:TIME V:MOTION _V:MOTION P:TRAJECTORY N:ARTIFACT
Hog 's Head 's sign creaked a little , though there was no breeze .
N:ARTIFACT-NE N:ARTIFACT V:PERCEPTION N:PHENOMENON

It even got a little worse during a business trip to the city , so
V:CHANGE _V:CHANGE P:SIMULEVT N:ACT P:DEST N:LOC
on the advice of a friend I set up an appointment with True Massage .
P:REASON _P:REASON P:SPEAKER N:PERSON V:ACT N:ACT P:COM N:GROUP-NE

Figure 17: Sentences (1) and (2) annotated for supersenses, named entities, multiword expressions, and prepositions. "_" indicates the continuation of a gappy multiword unit. The label N:LOC indicates the nominal LOCATION category. P:COM stands for the COMITATIVE preposition sense.

[28] Burkett et al. (2010) explored a similar setting, but assumed monolingual models were available for both languages of the parallel data.

[29] It remains to be determined whether function-tagged prepositions in MWEs (e.g. prepositional verbs) should be included in the integrated scheme, as they apply to only part of a chunk.

The central question in our *intrinsic* evaluation of this model will be: to what extent do the different pieces of the representation complement and reinforce each other? In previous work, semantic field categories similar to supersenses have been used for MWE extraction (Piao et al., 2003), and another study found that the best predictors of a transitive preposition's semantics are its head and object (Hovy et al., 2010)—no doubt due in part to the *meanings* of the head and object, which could be represented with supersenses. Though we are not proposing to model syntactic relations directly, we hope the model will enable fruitful information-sharing among nearby tokens. Experimental scenarios can include independent runs of the component models vs. a pipeline vs. a single joint model.

A final note is that nothing in the proposed representation or modeling approach is inherently specific to English. While for practical reasons the data annotated with this integrated representation will be limited to English, for each of the components we are aware of at least one comparable representation in another language. And though our unified model will exploit rich English language data sources in its features, we hope to show that even without these features a reasonably effective analyzer can be built—which would suggest our general approach to coarse lexical semantics through token-driven corpus annotation and sequence modeling is a viable one for any language where basic morphological processing is available.

We now turn briefly to applications and related topics.

5 Application to Frame-Semantic Parsing

FrameNet (Baker et al., 1998) is a linguistically rich semantic lexicon and corpus for predicate-argument structures. It organizes predicates into scenes, or **frames**, which are listed in the lexicon. Associated with each frame definition is a list of **lexical units** 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. Figure 18 gives an example sentence with a single frame annotation.

In previous work, we developed SEMAFOR, a system that uses probabilistic modeling to analyze the frame-semantic structure of an English input sentence (Das et al., 2010).[30] Originally a SemEval 2007 shared task (Baker et al., 2007), this combines a kind of word sense disambiguation (finding and disambiguating frame-evoking predicates) with semantic role labeling (finding arguments to each predicate and labeling them with roles).

[30] SEMAFOR has since seen a number of improvements (Das, 2012).

Another reader **takes** Christine Sutton **to task** on a semantic point .
JUDGMENT_DIRECT_ADDRESS: Communicator Addressee Topic

Figure 18: 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. Other lexical units include chide.v, compliment.{n,v}, harangue.v, tell off.v, telling off.n, tongue-lashing.n, and upbraid.v.

Here we propose to investigate whether SEMAFOR can exploit the output of a lexical semantic analyzer to better predict frame parses.

5.1 Target identification

The first phase of frame-semantic parsing is to detect frame-evoking expressions (called predicates or **targets**) in the sentence. SEMAFOR uses heuristic matching against a whitelist of targets culled from the FrameNet lexicon and annotated data. This list includes some multiword targets, but the current heuristics do not match gappy targets. In principle an accurate lexical analyzer should help improve recall (due to gappy targets) and precision (due to possible false positive multiwords), though these are rare enough[31] that performance gains are expected to be negligible. Another possibility to consider is to use the supersense tags N:ACT and N:EVENT to identify eventive nouns (e.g., malpractice) that may be missing from the lexicon. Finally, because prepositions are so ambiguous, the current heuristics do not identify any of the few that evoke frames; preposition function tags should enable more sensitive filtering heuristics.

5.2 Frame identification

The next step chooses one of 877 frames for each of the identified targets. This is accomplished with a feature-based conditional model learned from sentences with full frame annotations. Because the training data is relatively small (20,000 frame instances in FrameNet 1.5), adding new features that semantically categorize the target and its context—e.g. supersense and preposition function tags—may improve the model's generalization power.

[31] An analysis of the SemEval training data found just 4% of targets were multiword and 1% were gappy.

5.3 Argument identification

A second feature-based model brackets and classifies arguments, conditional on the inferred frame. Again, due to data sparseness, new features for the supersense of the semantic head of the candidate argument as well as the preposition function of a candidate PP argument would likely lead to better valency generalizations. Additionally, a new feature for argument candidates that violate multiword unit boundaries is expected to improve argument bracketing.

Because SEMAFOR currently leverages a large number of features, including syntactic information from a dependency parser, new features (even if they are predictive) may not substantially affect performance. Yet there is definitely room for improvement in multiple phases: the F_1 score for argument identification currently stands at 80% with oracle frames and 64% with predicted frames (Das, 2012, p. 73). Further analysis and experimentation is needed to understand and remedy the current system's shortcomings.

6 Application to 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 MT (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 (10) are mapped to the nonsensical French output (11a) instead of to (11b), suggesting that *mind* is being translated separately from *make up*:

(10) a. She was unable to make up the Count's mind.
b. She was unable to make up the mind of the Count.

(11) 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 20 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 unsupervised by a full-scale MT system. Better methods for building SMT systems with explicit information about lexical items may result from this research. Alternatively, the analysis might reveal new insights into current systems' ability to work around unanalyzed input, perhaps suggesting novel ways of recruiting parallel data (or even the systems themselves) to improve monolingual lexical semantic analysis.

6.1 Planned Experiments

Because the lexical semantic analyzer will expect well-formed English input, we will experiment with translation out of English. Specifically, we intend to build MT systems for two high-resource language pairs: English-French and English-German, using the 3 million word News Commentary corpus from the WMT translation task (Callison-Burch et al., 2012). This will allow us to examine the role of lexical semantics in two language families without the confound of morphology (in morphologically richer languages many of the functions of English prepositions will be assumed by case-marking affixes/clitics, which would require special handling). For evaluation we will measure BLEU score (Papineni et al., 2002) on the standard WMT test sets.[32]

[32] At present, METEOR (Banerjee and Lavie, 2005) is only available for translation into English. TER tends to behave similarly to BLEU.

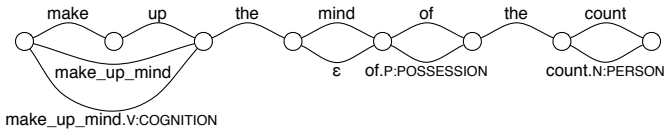

Figure 19: Partial lattice for (10b).

Prepositions are known to be especially challenging for machine translation (Gustavii, 2005), and are a high-value target due to their frequency. Following on our investigations from §3, we will investigate whether conjoining preposition tokens with automatic function tags produces more reliable word alignments, and ultimately better translations.

Then, we will consider MWEs and supersense tags from our analyzer. We will examine whether automatic word alignment and phrase extraction procedures tend to respect the unit status of MWEs. If MWEs are frequently broken up, then simply adding a phrase for the entire MWE may enable the decoder to form better hypotheses.

Finally, the multiple levels of structure in our lexical semantic representation suggest a model which has the flexibility to choose the best level of generalization that is supported by the data. We will therefore experiment with **lattice translation** (Dyer et al., 2008), in which each input sentence at test time is specified as a lattice. The lattice will be constructed with three levels of structure: the plain sentence; the MWE-chunked[33] sentence according to our lexical analyzer; and the full chunked and semantically tagged analysis (Figure 19). In choosing a path through the sentence lattice, the decoder will then be free to mix and match the different granularities of representation. For comparison we will also build a system with a Hiero-style grammar (Chiang, 2007), which can handle gappy chunks directly.

6.2 Related Work

Surprisingly, adpositions have received little attention in the SMT paradigm (Baldwin et al., 2009). An exception is the work of Toutanova and Suzuki (2007) in generating Japanese case-marking postpositions in English-Japanese SMT, which uses a target side reranker. Here we propose to focus instead on improving the

[33] Multiword chunks will be provided to the MT system as words-with-spaces. Gappy chunks cannot be represented directly in the lattice, so we will use a canonical member of the gappy MWE (such as the verb) to determine the chunk's position in the lattice.

representation on the source side.

Word sense disambiguation has been found to yield at best small gains in SMT systems (Carpuat and Wu, 2005; Cabezas and Resnik, 2005; Chan et al., 2007). 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 studies that have used *coarse-grained* senses, 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 to solve with our approach.

Several studies have modeled various kinds of MWEs within MT systems. Among these are studies by Carpuat and Diab (2010) and Ramisch (2012), both of which sought to improve phrase-based statistical MT out of English by identifying English MWEs. Carpuat and Diab (2010) used heuristic matching of the source side against English WordNet entries to improve an English-Arabic SMT system (trained on 2 million sentence pairs). They experimented with two methods: conjoining MWEs as words-with-spaces in preprocessing; and adding a translation model feature counting the number of MWEs in the source language side of the phrase pair, so as to penalize translation hypotheses that break MWEs. Each method produced a modest improvement in BLEU and TER scores. Ramisch (2012) built several smaller-scale English-Portuguese systems (trained on about 4,000 sentence pairs) with different methods of incorporating information about English phrasal verbs. Automatic (BLEU/NIST) and human evaluations were inconclusive, with little difference between the baseline system and five variants. The approach proposed here will similarly conjoin source side MWEs, but with three important differences: first, we aim to recognize many more kinds MWEs expressions than just phrasal verbs or WordNet entries, so we expect to have greater impact on the results; second, we will integrate semantic tags in our representation; and third, we will use lattice translation, which is able to back off to a less refined representation where called for by the data.

Other investigations have similarly manipulated the source side to improve source-target correspondences in SMT systems: Yeniterzi and Oflazer (2010), for instance, modified the English source string on the basis of syntax to build complex multiwords, improving factored phrase-based translation into Turkish.

For the word alignment subtask, Fraser and Marcu (2007) developed a model that is capable of inferring M-to-N alignments, where there are multiple, possibly nonconsecutive words on both the source and target sides.

7 Concluding Remarks

The lexical semantic analysis agenda presented here is to build new pathways between linguistic corpus annotation, statistical modeling, and natural language applications—bridged by a reasonably simple, yet general, representation of units and categories of lexical meaning. Multiword unit analysis, supersense analysis, and preposition function analysis will be the core components, new datasets and tools will be generated, and two external tasks will offer measures of practical impact. Related tasks (better integration of tokenization and POS tagging/morphological analysis, finer-grained semantic representations, new sense inventories, context beyond the sentence, and so on) lie farther down the road.

8 Timeline

<table>
<tr><th></th><th></th><th>Supersenses</th><th>Multiword Expressions</th><th>Prepositions</th></tr>
<tr><td rowspan="3">2012
2013</td><td>DEC
JAN
FEB</td><td rowspan="3">Wrap up Arabic SST research (NAACL?)</td><td rowspan="3">Refinement of MWE annotation task (ACL?)</td><td rowspan="3">—</td></tr>
<tr><td>MAR
APR
MAY</td><td rowspan="2">Frame-semantic parsing infrastructure</td><td rowspan="3">MWE annotation & annotation modeling</td><td rowspan="2">Prototype prep annotation</td></tr>
<tr><td>JUN
JUL
AUG</td><td rowspan="2">SST for frame parsing</td><td rowspan="3">Prep annotation</td></tr>
<tr><td>SEP
OCT
NOV
DEC</td><td rowspan="2">SST for MT; English SST annotation: nouns, then verbs</td><td>MWE identification experiments</td></tr>
<tr><td>2014</td><td>JAN
FEB</td><td>MWE for frame parsing</td><td>English prep tagging</td></tr>
<tr><td></td><td>MAR
APR
MAY</td><td colspan="3">Unified model; frame parsing and MT evaluations</td></tr>
<tr><td></td><td>JUN
JUL
AUG</td><td colspan="3">Writing and job search</td></tr>
</table>

A Supersense Tagset for Nouns

Here is the complete supersense tagset for nouns. Each tag is briefly described by its `symbol`, NAME, short description, and **examples**.

0 **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

^ **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.

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.

C Multiword Expression Annotation Instructions (Pilot Study 1, §2.2.3)

Multiword Expression Annotation

The Task

You are given a (pre-tokenized) English sentence. **Your mission is to partition the sentence into lexical expressions, each consisting of one or more tokens.**

Most tokens will remain as they are, but some belong to a multiword expression and should therefore be joined to other tokens. What is meant by *multiword expression*? Intuitively, **any expression that (a) is a proper name, or (b) ought to be listed in a dictionary because its structure and/or meaning and/or frequency are not predictable solely on the basis of its components**. (This definition includes, but is not limited to, idioms and noncompositional expressions.) Some examples:

- a lot of
- in spite of
- as well as
- for example
- of course
- in person
- meet with (somebody)
- meet up
- meet up with (somebody)
- kick the bucket
- fire and brimstone
- green thumb
- outer space
- computer science
- hurt like hell
- miles per gallon
- cut to the chase
- Mr. Barack Obama

Joining consecutive tokens with an underscore character (e.g., `miles_per_gallon`) marks them as a multiword expression.

Discontiguous Expressions

The components of the expression need not be contiguous in the sentence. For example:

- make a decision, made the best decision, etc.
- hold someone hostage
- drive someone crazy
- cut the vacation short
- the person with whom I met

There are two ways to mark discontiguous expressions. As long as an expression contains 2 contiguous parts separated by a gap, and no part of any other discontiguous expression falls within that gap, you can use a trailing underscore on the first part and a leading underscore on the second part. Alternatively, you can mark tokens or contiguous expressions with numeric indices like `|1`, `|2`, etc.

- `my idea was met_with_ much _scorn` (i.e., *met*, *with*, and *scorn* all belong to the same expression). Equivalently: `my idea was met|1 with|1 much scorn|1` or `my idea was met_with|4 much scorn|4`. (The number itself is not important, so long as all parts of an expression have the same index.)
- `You should hold|1 your head|1 quite high|1`

Example

Here is a full sentence annotated according to my (Nathan's) intuitions:

```
It even got_ a little _worse during a business_trip to the city ,
so on|1 the advice|1 of a friend I set_up an appointment with True_Massage .
```

Notes

1. Do not split any of the tokens.
2. If a word is obviously **misspelled**, do not correct the spelling but interpret it as the intended word. If any of the characters from the original sentence disappear, the interface will yell at you.
3. Multiword expressions should only include **punctuation tokens** that are in the middle of the expression or obviously does not belong to the rest of the sentence: *I , Robot*, *Yahoo!*.
4. Don't worry about the token's **morphological form**.
5. Keep in mind that an expression may behave differently in different contexts. For example, the bigram *met with*: *I met with my supervisor*, *my idea was met with scorn*, *the person I met with dark red hair*.

Language is vibrant and messy, and these guidelines are admittedly (and intentionally) somewhat vague. **You will have to make a lot of judgment calls.** This is just a pilot round, so you are encouraged to keep notes of difficulties for future discussion. Note that some sentences will not contain any multiword expressions.

Once you've started the annotation, please don't discuss the data with other annotators—we want to avoid biasing anyone at this stage. There will be opportunity for discussion later on.

References

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. Springer Netherlands, 2003.

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.

Collin Baker, Michael Ellsworth, and Katrin Erk. SemEval-2007 Task 19: frame semantic structure extraction. In *Proc. of SemEval*, pages 99–104, Prague, Czech Republic, June 2007.

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.

Tim Baldwin, Valia Kordoni, and Aline Villavicencio. Prepositions in applications: a survey and introduction to the special issue. *Computational Linguistics*, 35(2):119–149, 2009.

Timothy Baldwin and Su Nam Kim. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL, 2010. ISBN 978-1420085921.

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 Netherlands, 2006.

Satanjeev Banerjee and Alon Lavie. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proc. of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005.

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.

Kenneth R. Beesley and Lauri Karttunen. *Finite state morphology*. University of Chicago Press, Chicago, 2003.

Eduard Bejček and Pavel Straňák. Annotation of multiword expressions in the Prague dependency treebank. *Language Resources and Evaluation*, 44(1):7–21, 2010.

Ann Bies, Justin Mott, Colin Warner, and Seth Kulick. *English Web Treebank*. Linguistic Data Consortium, Philadelphia, PA, 2012. LDC2012T13.

Daniel M. Bikel, Richard Schwartz, and Ralph M. Weischedel. An algorithm that learns what's in a name. *Machine Learning*, 34(1):211–231, February 1999.

Julia Birke and Anoop Sarkar. A clustering approach for the nearly unsupervised recognition of nonliteral language. In *Proc. of EACL*, pages 329–336, Trento, Italy, April 2006.

Leonard Bloomfield. *Language*. Henry Holt, New York, 1933.

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, January 2001.

David Burkett, Slav Petrov, John Blitzer, and Dan Klein. Learning better monolingual models with unannotated bilingual text. In *Proc. of CoNLL*, pages 46–54, Uppsala, Sweden, July 2010.

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.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Findings of the 2012 Workshop on Statistical Machine Translation. In *Proc. of WMT*, pages 10–51, Montréal, Canada, June 2012.

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.

Bob Carpenter. Multilevel bayesian models of categorical data annotation. Technical report, Alias-i, Inc., 2008. URL http://lingpipe.files.wordpress.com/2008/11/carp-bayesian-multilevel-annotation.pdf.

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, 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.

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.

David Chiang. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2): 201–228, 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.

Noam Chomsky. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA, 1965.

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 Michael Collins and Mark Steedman, editors, *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.

William W. Cohen and Vitor R. Carvalho. Stacked sequential learning. In *Proc. of International Joint Conferences on Artificial Intelligence*, pages 671–676, Edinburgh, Scotland, 2005.

Michael Collins. Discriminative training methods for Hidden Markov Models: theory and experiments with perceptron algorithms. In *Proc. of EMNLP*, pages 1–8, Philadelphia, PA, 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.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. The VNC-Tokens dataset. In *Proc. of MWE*, pages 19–22, Marrakech, Morocco, 2008.

William Croft. *Radical Construction Grammar: syntactic theory in typological perspective*. Oxford University Press, Oxford, 2001.

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. *Semi-supervised and latent-variable models of natural language semantics*. PhD thesis, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, 2012. URL http://www.dipanjandas.com/files/thesis.pdf.

Dipanjan Das and Slav Petrov. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proc. of ACL-HLT*, pages 600–609, Portland, Oregon, USA, June 2011.

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, June 2010.

A. Philip Dawid and Allan M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28, January 1979.

Nicolas Denand and Monique Rolbert. Contextual processing of locative prepositional phrases. In *Proc. of Coling*, pages 1353–1359, Geneva, Switzerland, August 2004.

Mona Diab and Pravin Bhutada. Verb noun construction MWE token classification. In *Proc. of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, pages 17–22, 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 / Heidelberg, 2009.

Mona Diab and Philip Resnik. An unsupervised method for word sense tagging using parallel corpora. In *Proc. of ACL*, pages 255–262, Philadelphia, Pennsylvania, USA, July 2002.

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.

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.

Christopher Dyer, Smaranda Muresan, and Philip Resnik. Generalizing word lattice translation. In *Proc. of ACL-HLT*, pages 1012–1020, Columbus, Ohio, June 2008.

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, 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.

Afsaneh Fazly, Suzanne Stevenson, and Ryan North. Automatically learning semantic knowledge about multiword predicates. *Language Resources and Evaluation*, 41(1):61–89, 2007.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103, 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, MA, 1998.

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.

Richard Fothergill and Timothy Baldwin. Fleshing it out: a supervised approach to MWE-token and MWE-type classification. In *Proc. of 5th International Joint Conference on Natural Language Processing*, pages 911–919, Chiang Mai, Thailand, November 2011.

Richard Fothergill and Timothy Baldwin. Combining resources for MWE-token classification. In *Proc. of *SEM*, pages 100–104, Montréal, Canada, June 2012.

Alexander Fraser and Daniel Marcu. Getting the structure right for word alignment: LEAF. In *Proc. of EMNLP-CoNLL*, pages 51–60, Prague, Czech Republic, June 2007.

Kevin Gimpel and Noah A. Smith. Rich source-side context for statistical machine translation. In *Proc. of WMT*, pages 9–17, Columbus, Ohio, 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, July 2011.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. Part-of-speech tagging for Twitter: annotation, features, and experiments. In *Proc. of ACL-HLT*, pages 42–47, Portland, Oregon, USA, June 2011.

Adele E. Goldberg. *Constructions at work: the nature of generalization in language*. Oxford University Press, Oxford, 2006.

Laura Gonnerman and Mary-Jane Blais. L2 processing of English phrasal verbs. The 31st Second Language Research Forum, October 2012.

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), to appear.

Stefan Th. Gries. Phraseology and linguistic theory: a brief survey. In Sylviane Granger and Fanny Meunier, editors, *Phraseology: an interdisciplinary perspective*, pages 3–25. Amsterdam: John Benjamins, 2008.

Ebba Gustavii. Target language preposition selection – an experiment with transformation-based learning and aligned bilingual data. In *Proc. of the 10th European Association for Machine Translation Conference (EAMT)*, pages 112–118, Budapest, 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.

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, October 2008.

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.

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, June 2008.

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 / Heidelberg, 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.

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.

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.

Nobuhiro Kaji, Yasuhiro Fujiwara, Naoki Yoshinaga, and Masaru Kitsuregawa. Efficient staggered decoding for sequence labeling. In *Proc. of ACL*, pages 485–494, Uppsala, Sweden, July 2010.

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.

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.

Henry Kučera and W. Nelson Francis. *Computational analysis of present-day American English*. Brown University Press, Providence, RI, 1967.

Mirella Lapata and Alex Lascarides. Detecting novel compounds: the role of distributional evidence. In *Proc. of EACL*, pages 235–242, 2003.

Linlin Li and Caroline Sporleder. Linguistic cues for distinguishing literal and non-literal usages. In *Coling 2010: Posters*, pages 683–691, Beijing, China, August 2010.

Kenneth C. Litkowski. Proposed next steps for The Preposition Project. Technical Report 12-01, CL Research, Damascus, MD, 2012. URL http://www.clres.com/online-papers/NextTPPSteps.pdf.

Kenneth C. 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, University of Essex - Colchester, United Kingdom, 2005.

Kenneth C. 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, 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.

Ron Martinez and Norbert Schmitt. A phrasal expressions list. *Applied Linguistics*, 33(3): 299–320, July 2012.

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.

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, NJ, USA, March 1993.

Behrang Mohit, Nathan Schneider, Rishav Bhowmick, Kemal Oflazer, and Noah A. Smith. Recall-oriented learning of named entities in Arabic Wikipedia. In *Proc. of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 162–173, Avignon, France, April 2012.

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.

Robert Munro, Steven Bethard, Victor Kuperman, Vicky Tzuyin Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen, and Harry Tily. Crowdsourcing and language studies: the new generation of linguistic data. In *Proc. of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 122–130, Los Angeles, June 2010.

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 / Heidelberg, 2011.

Kishorjit Nongmeikapam, Dhiraj Laishram, Naorem Singh, Ngariyanbam Chanu, and Sivaji Bandyopadhyay. Identification of reduplicated multiword expressions using CRF. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing: Proceedings of the 12th International Conference (CICLing 2011)*, volume 6608 of *Lecture Notes in Computer Science*, pages 41–51. Springer, Berlin, 2011.

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. Idioms. *Language*, 70(3):491–538, September 1994.

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, 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, and Nathan Schneider. Part-of-speech tagging for Twitter: Word clusters and other advances. Technical Report CMU-ML-12-107, Carnegie Mellon University, Pittsburgh, Pennsylvania, September 2012. URL http://www.ark.cs.cmu.edu/TweetNLP/owoputi+etal.tr12.pdf.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proc. of ACL*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002.

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*, 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.

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.

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.

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.

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, MA, 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 / Heidelberg, 2011.

Terry Regier. *The human semantic potential: spatial language and constrained connectionism*. MIT Press, Cambridge, MA, September 1996.

J. K. Rowling. *Harry Potter and the Half-Blood Prince*. Arthur A. Levine Books, New York, NY, 2005.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. FrameNet II: extended theory and practice, September 2010. URL https://framenet2.icsi.berkeley.edu/docs/r1.5/book.pdf.

Patrick Saint-Dizier. PrepNet: a multilingual lexical description of prepositions. In *Proc. of LREC*, volume 6, Genoa, Italy, 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 / Heidelberg, 2012.

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.

Violeta Seretan. *Syntax-based collocation extraction*. Number 44 in Text, Speech and Language Technology. Springer-Verlag, New York, January 2011. DOI: 10.1007/978-94-007-0134-2.

Rita Simpson-Vlach and Nick C. Ellis. An academic formulas list: new methods in phraseology research. *Applied Linguistics*, 31(4):487–512, September 2010.

Noah A. Smith. *Linguistic Structure Prediction*. Number 13 in Synthesis Lectures on Human Language Technologies. Morgan & Claypool, San Rafael, CA, May 2011.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. Cheap and fast — but is it good? Evaluating non-expert annotations for natural language tasks. In *Proc. of EMNLP*, pages 254–263, Honolulu, Hawaii, October 2008.

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.

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.

Kristina Toutanova and Hisami Suzuki. Generating case markers in machine translation. In *Proc. of NAACL-HLT*, pages 49–56, Rochester, New York, April 2007.

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.

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.

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, Canada, June 2012.

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.

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.

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. *Semi-compositional noun + verb constructions: theoretical questions and computational linguistic analyses*. Ph.D. dissertation, University of Szeged, Szeged, Hungary, August 2011. URL http://www.inf.u-szeged.hu/~vinczev/PhD/PhD_thesis_Vincze_Veronika.pdf.

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.

Janyce M. Wiebe, Rebecca F. Bruce, and Thomas P. O'Hara. Development and use of a gold-standard data set for subjectivity classifications. In *Proc. of ACL*, pages 246–253, College Park, Maryland, USA, June 1999.

Stefanie Wulff. *Rethinking idiomaticity: a usage-based approach*. Research in Corpus and Discourse. Continuum International Publishing Group, 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*. Mouton, Berlin, 2010.

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.

David Yarowsky, Grace Ngai, and Richard Wicentowski. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proc. of HLT*, San Diego, CA, March 2001.

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.

Reyyan Yeniterzi and Kemal Oflazer. Syntax-to-morphology mapping in factored phrase-based statistical machine translation from English to Turkish. In *Proc. of ACL*, pages 454–464, Uppsala, Sweden, July 2010.

Cornelia Zelinksy-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*, pages 351–390. Mouton de Gruyter, New York, 1993.

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

Index of linguistic examples